Christian P. Robert

The Bayesian Choice

From Decision-Theoretic Foundations to Computational Implementation

Second Edition

Springer

Christian P. Robert
CEREMADE
Universite Paris Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris cedex 16
France
xian@ceremade.dauphine.fr

Library of Congress Control Number: 2007926596

ISBN 978-0-387-71598-8 e-ISBN 978-0-387-71599-5

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To my reference prior,
Brigitte,
and to two major updates,
Joachim and Rachel.

Preface to the Paperback Edition

What could not be changed must be endured.
Robert Jordan, *New Spring, Prequel to The Wheel of Time*

Thanks

While this paperback edition is almost identical to the second edition of *The Bayesian Choice*, published in 2001, and thus does not require a specific introduction, it offers me the opportunity to thank several groups of people for their contributions that made this edition possible.

First, the changes, when compared with the second edition, are only made of corrections of typographical and conceptual errors (whose updated list can be found on my Webpage[1]). Almost all errors have been pointed out to me by friends, colleagues, unknown lecturers or anonymous readers who (always kindly and sometimes apologetically) sent me emails asking me to clarify a specific paragraph, a formula or a problem that did not make sense to them. Needless to say, I am very grateful to those numerous contributors for making the book more accurate and I obviously encourage all contributors who think there could be an error in the current edition to contact me because they cannot be wrong! Either there indeed is a mistake that needs to be set right or there is no mistake but the context is ambiguous at best and the corresponding text needs to be rewritten. Thanks, then, to Guido Consonni, Estelle Dauchy, Arnaud Doucet, Pierre Druihlet, Ed Green, Feng Liang, Jean-Michel Marin, M.R.L.N. Panchanana, Fabrice Pautot, and Judith Rousseau.

Second, working with my colleague Jean-Michel Marin on the design of a course for teaching Bayesian Statistics from a practical and computational perspective (a venture now published as *Bayesian Core* by Springer in early 2007) was a very important moment in that I realized that the material in this very book, *The Bayesian Choice*, was essential in communicating the essential relevance and coherence of the Bayesian approach

[1] `http://www.ceremade.dauphine.fr/~xian/books.html`

through its decision-theoretic foundations, while the message contained in the other book and transmitted only through processing datasets is that the Bayesian methodology is a universal and multifaceted tool for data analysis. While introducing wider and less mathematical audiences to the elegance and simplicity of the Bayesian methodology in a shorter and therefore more focussed volume was also necessary, if only because some learn better from examples than from theory, I came to the conclusion that there was no paradox in insisting on those foundations in another book! I am therefore immensely thankful to Jean-Michel Marin for initiating this epiphany (if I may rightly borrow this expression from Joyce!), as well as for several years of intense collaboration. Similarly, the DeGroot Prize committee of the ISBA—International Society for Bayesian Analysis—World meeting of 2004 in Valparaiso, Chile, greatly honored me by attributing to *The Bayesian Choice* this prestigious prize. In doing so, this committee highlighted the relevance of both foundations and implementation for the present and future of Bayesian Statistics, when it stated that the *"book sets a new standard for modern textbooks dealing with Bayesian methods, especially those using MCMC techniques, and that it is a worthy successor to DeGroot's and Berger's earlier texts".* I am quite indebted to the members of the committee for this wonderful recognition.

Third, it has been more than 18 years since I started working with John Kimmel from Springer New York (on a basic Probability textbook with Arup Bose that never materialized), and I always appreciated the support he provided over the various editions of the books. So, when he presented me with the possibility to publish this paperback edition, I first got some mixed feelings, because he made me feel like a classics author! This caused my kids poking endless fun at me and, in the end, I am quite grateful to John for the opportunity to teach from this book to a wider audience and thus hopefully exposing them to the beauty of Bayesian theory and methodology. Short of embarking upon a translation of *The Bayesian Choice* into Chinese or Arabic, I do not think there is much more he could do to support the book!

In Memoriam

This is a sheer consequence of time moving on, unfortunately, but I lost another dear friend since the last publication of *The Bayesian Choice*. José Sam Lazaro passed away last Spring: a mathematician, a professor and a colleague at the Université de Rouen, a music addict and a movie aficionado that made me discover *Der Tod und das Mädchen* as well as *The Night of The Hunter*, an intense piano player, a memorable tale teller, he was above all a philosopher and a friend. Although he would have made a joke out of it, I would like to dedicate this edition to his memory and wish him well to play this final and endless sonata...

Valencia and Paris **Christian P. Robert**
February 2007

Preface to the Second Edition

> "You can never know everything," Lan said quietly, "and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway."
>
> **Robert Jordan**, *Winter's Heart, Book IX of the Wheel of Time.*

Overview of Changes

Why a second edition? When thinking about it, this is more like a third edition, since the previous edition of *The Bayesian Choice* was the translation of the French version, and already included updates and corrections. The reasons for a new edition of the book are severalfold. The Bayesian community has grown at an incredible pace since 1994. The previous version not only overlooks important areas in the field but misses the significant advances that have taken place in the last seven years.

Firstly, the MCMC[2] revolution has fueled considerable advances in Bayesian modeling, with applications ranging from medical Statistics, to signal processing, to finance. While present in the 1994 edition, these methods were not emphasized enough: for instance, MCMC methods were not presented until the penultimate chapter.

Another significant advance that needed attention is the development of new testing approaches and, more generally, of model choice tools in connection with, and as a result of, MCMC techniques such as reversible jump. Other important expansions include hierarchical and dynamic models, whose processing only began to emerge in the early 1990s.

This second edition is not revolutionary, compared with the 1994 edition. It includes, however, important advances that have taken place since then. The only new chapter deals with model choice (Chapter 7) and is isolated from general testing theory (Chapter 5), since model choice is indeed a different problem and also because it calls for new, mostly computational,

[2] MCMC stands for *Markov chain Monte Carlo*, a simulation methodology which was (re)discovered in the early 1990s by the Bayesian community.

tools. For this reason, and also to emphasize the increasing importance of computational techniques, Chapter 6—previously Chapter 9—has been placed earlier in the book, after the presentation of the fundamentals of Bayesian Statistics. Chapter 6 could almost be called a new chapter in that its presentation has been deeply renovated in the light of ten years of MCMC practice. In Chapter 3, the material on noninformative priors has been expanded and includes, in particular, matching priors, since the research activity has been quite intense in this area in the past few years. Chapter 4 still deals with general estimation problems, but I have incorporated a new section on dynamic models, since those are quite central to the development of Bayesian Statistics in applied fields such as signal processing, finance and econometrics. Despite Delampady's criticisms of Chapter 11 in *The Mathematical Reviews*, I have decided to leave this chapter in: it does not hurt, when one is finished reading a book, to take an overall and more philosophical view of the topic because the reader has very likely acquired enough perspective to understand such arguments. (In a strictly textbook implementation, this chapter can be suggested as an additional reading, comparable with the Notes.)

Another noteworthy change from the previous edition is the decreased emphasis on decision-theoretic principles. Although I still believe that statistical procedures must be grounded on such principles, the developments in the previous decade have mainly focused on methodology, including computational methodology, rather than attacking broader and more ambitious decision problems (once again, including computational methodology). The second part of the book (starting with Chapter 6) is therefore less decision-theoretic and, in contrast to others, chapters such as Chapters 8 and 9 have hardly been changed.

At a more typographical level, subsections and separations have been introduced in many sections to improve visibility and reading, and more advanced or more sketchy parts have been relegated to a *Notes* section at the end of each chapter, following the approach adopted in *Monte Carlo Statistical Methods*, written with George Casella. The end of an example is associated with the ‖ symbol, while the end of a proof is indicated by the □□ symbol.

Several books on Bayesian Statistics have appeared in the interim, among them Bernardo and Smith (1994), Carlin and Louis (1996, 2000a), Gelman et al. (1996), O'Hagan (1994), and Schervish (1995). However, these books either emphasize deeper theoretical aspects at a higher mathematical level (Bernardo and Smith (1994), O'Hagan (1996), or Schervish (1996)) and are thus aimed at a more mature audience than this book, or they highlight a different vision of the practice of Bayesian Statistics (Carlin and Louis (2000a) or Gelman et al. (1996)), missing for instance the connection with Decision Theory developed in this book.

Course Schedules

My advice about running a course based on this book has hardly changed. In a first course on Bayesian analysis, the basic chapters (Chapters 1–6) should be covered almost entirely, with the exception of the Notes and Sections 4.5 and 5.4, while a course focusing more on Decision Theory could skip parts of Chapters 1 and 3, and Chapter 4 altogether, to cover Chapters 7–9. For a more advanced curriculum for students already exposed to Bayesian Statistics, my suggestion is first to cover the impropriety issue in Section 1.5, the noninformative priors in Section 3.5, the dynamic models in Section 4.5 and Notes 4.7.3 and 4.7.4. I would also spend time on the testing issues of Chapter 5 (with the possible exception of Sections 5.3 and 5.4). Then, after a thorough coverage of simulation methods via Chapter 6, I would move to the more controversial topic of model choice in Chapter 7, to recent admissibility results as in Section 8.2.5 and Note 8.7.1, and to the hierarchical and empirical modelings of Chapter 10. In this later scenario, the Notes should be most helpful for setting out reading seminars.

Thanks

I have always been of two minds about including a thank-you section in a book: on the one hand, it does not mean anything to most readers, except maybe to bring to light some of the author's idiosyncrasies that might better remain hidden! It may also antagonize some of those concerned because they are not mentioned, or because they are not mentioned according to their expectations, or even because they *are* mentioned! On the other hand, the core of ethical requirements for intellectual works is that sources should be acknowledged. This extends to suggestions that contributed to making the work better, clearer or simply different. And it is a small token of gratitude to the following people for the time spent on the successive drafts of this edition that their efforts should be acknowledged in print for all to behold!

Although this is "only" a revision, the time spent on this edition was mostly stolen from evenings, early mornings and week-ends, that is from Brigitte, Joachim and Rachel's time! I am thus most grateful to them for reading and playing (almost) quietly while I was typing furiously and searching desperately through piles of material for this or that reference, and for listening to Bartoli and Gudjónsson, rather than to Manau or Diana Krall! I cannot swear this book-writing experience will never happen again but, in the meanwhile, I promise there will be more time available for reading *Mister Bear to the Rescue*, and for besieging the Playmobil castle in full scale, for playing chess and for biking on Sunday afternoons!

I am thankful to several people for the improvements in the current edition! First, I got a steady stream of feedback and suggestions from those who taught from the book. This group includes Ed Green, Tatsuya Kubokawa, and Marty Wells. In particular, Judith Rousseau, radical biker

and Jordanite as well as Bayesian, definitely was instrumental in the reorganization of Chapter 3. I also got many helpful comments from many people, including the two "Cambridge Frenchies" Christophe Andrieu and Arnaud Doucet (plus a memorable welcome for a retreat week in Cambridge to finish Chapter 6), Jim Berger (for his support in general, and for providing preprints on model choice in particular), Olivier Cappé (who also installed Linux on my laptop, and consequently brought immense freedom for working on the book anywhere, from the sand-box to the subway, and, lately, to CREST, where Unix is now banned!), Maria DeIorio, Jean-Louis Fouley, Malay Ghosh (through his very supportive review in JASA), Jim Hobert (who helped in clarifying Chapters 6 and 10), Ana Justel, Stephen Lauritzen (for pointing out mistakes with Wishart distributions), Anne Philippe, Walter Racugno (who gave me the opportunity to teach an advanced class in model choice in Ca'liari last fall, thus providing the core of Chapter 7), Adrian Raftery, Anne Sullivan Rosen (about the style of this preface), and Jean-Michel Zakoian (for his advice on the new parts on dynamic models). I also take the opportunity to thank other friends and colleagues such as George Casella, Jérôme Dupuis, Merrilee Hurn, Kerrie Mengersen, Eric Moulines, Alain Monfort, and Mike Titterington, since working with them gave me a broader vision of the field, which is hopefully incorporated in this version. In particular, the experience of writing *Monte Carlo Statistical Methods* with George Casella in the past years left its mark on this book, not only through the style file and the inclusion of Notes, but also as a sharper focus on essentials. Manuela Delbois helped very obligingly with the transformation from TEX to LATEX, and with the subsequent additions and indexings. And, last but not least, John Kimmel and Jenny Wolkowicki, from Springer-Verlag, have been very efficient and helpful in pushing me to write this new edition for the former, in keeping the whole schedule under control and in getting the book published on time for the latter. Needless to say, the usual *proviso* applies: all remaining typos, errors, confusions and obscure statements are mine and only mine!

In Memoriam

A most personal word about two people whose *absence* has marked this new edition: in the summer 1997, I lost my friend Costas Goutis in a diving accident in Seattle. By no means am I the only one to feel keenly his absence, but, beyond any doubt, this book would have benefited from his vision, had he been around. Two summers later, in 1999, Bernhard K. Flury died in a mountain accident in the Alps. While the criticisms of our respective books always focussed on the cover colors, even to the extent of sending one another pirated versions in the "right" color, the disappearance of his unique humor has taken a measure of fun out of the world.

Paris, France — Christian P. Robert
March 2001

Preface to the First Edition

> From where we stand, the rain seems random.
> If we could stand somewhere else, we would see the order in it.
>
> — **T. Hillerman** (1990) *Coyote Waits*.

This book stemmed from a translation of a French version that was written to supplement the gap in the French statistical literature about Bayesian Analysis and Decision Theory. As a result, its scope is wide enough to cover the two years of the French graduate Statistics curriculum and, more generally, most graduate programs. This book builds on very little prerequisites in Statistics and only requires basic skills in calculus, measure theory, and probability. Intended as a preparation of doctoral candidates, this book goes far enough to cover advanced topics and modern developments of Bayesian Statistics (complete class theorems, the Stein effect, hierarchical and empirical modelings, Gibbs sampling, etc.). As usual, what started as a translation eventually ended up as a deeper revision because of the comments of French readers, adjustments to the different needs of American programs, and because my perception of things has changed slightly in the meantime. As a result, this new version is quite adequate for a general graduate audience of an American university.

In terms of level and existing literature, this book starts at a level similar to those of the introductory books of Lee (1989) and Press (1989), but it also goes further and keeps up with most of the recent advances in Bayesian Statistics, while justifying the theoretical appeal of the Bayesian approach on decision-theoretic grounds. Nonetheless, this book differs from the reference book of Berger (1985a) by including the more recent developments of the Bayesian field (the Stein effect for spherically symmetric distributions, multiple shrinkage, loss estimation, decision theory for testing and confidence regions, hierarchical developments, Bayesian computation, mixture estimation, etc.). Moreover, the style is closer to that of a textbook in the sense that the progression is intended to be linear. In fact, the exposition of

the advantages of a Bayesian approach and of the existing links with other axiomatic systems (fiducial theory, maximum likelihood, frequentist theory, invariance, etc.) does not prevent an overall unity in the discourse. This should make the book easier to read by students; through the years and on both sides of the blackboard(!), I found most Statistics courses disturbing because a wide scope of methods was presented simultaneously with very little emphasis on ways of discriminating between competing approaches. In particular, students with a strong mathematical background are quite puzzled by this multiplicity of theories since they have not been exposed previously to conflicting systems of axioms. A unitarian presentation that includes other approaches as limiting cases is thus more likely to reassure the students, while giving a broad enough view of Decision Theory and even of parametric Statistics.

The plan[3] of the book is as follows: Chapter 1 is an introduction to statistical models, including the Bayesian model and some connections with the Likelihood Principle. The book then proceeds with Chapter 2 on Decision Theory, considered from a classical point of view, this approach being justified through the axioms of rationality and the need to compare decision rules in a coherent way. It also includes a presentation of usual losses and a discussion of the Stein effect. Chapter 3 gives the corresponding analysis for prior distributions and deals in detail with conjugate priors, mixtures of conjugate priors, and noninformative priors, including a concluding section on prior robustness. Classical statistical models are studied in Chapter 4, paying particular attention to normal models and their relations with linear regression. This chapter also contains a section on sampling models that allows us to include the pedagogical example of capture-recapture models. Tests and confidence regions are considered separately in Chapter 5, since we present the usual construction through $0-1$ losses, but also include recent advances in the alternative decision-theoretic evaluations of testing problems. The second part of the book dwells on more advanced topics and can be considered as providing a basis for a more advanced graduate course. Chapter 8 covers complete class results and sufficient/necessary admissibility conditions. Chapter 9 introduces the notion of invariance and its relations with Bayesian Statistics, including a heuristic section on the Hunt–Stein theorem. Hierarchical and empirical extensions of the Bayesian approach, including some developments on the Stein effect, are treated in Chapter 10. Chapter 6 is quite appealing, considering the available literature, as it incorporates in a graduate textbook an introduction to state-of-the-art computational methods (Laplace, Monte Carlo and, mainly, Gibbs sampling). In connection with this chapter, a short appendix provides the usual pseudo-random generators. Chapter 11 is a more personal conclusion on the advantages of Bayesian theory, also mentioning the most common criticisms of the Bayesian approach. French readers may appreciate that

[3] The chapter and section numbers have been adapted to the current edition.

a lot of effort has been put into the exercises of each chapter in terms of volume and difficulty. They now range from very easy to difficult, instead of being uniformly difficult! The most difficult exercises are indexed by asterisks and are usually derived from research papers (covering subjects such as *spherically symmetric distributions* (1.1), *the Pitman nearness criticism* (2.57–2.62), *marginalization paradoxes* (3.44–3.50), *multiple shrinkage* (10.38), etc.). They should benefit most readers by pointing out new directions of Bayesian research and providing additional perspectives.

A standard one-semester course should cover the first five chapters (with the possible omission of Note 2.8.2, §2.5.4, §2.6, §3.4, Note 4.7.1, §4.3.3, and §5.4). More advanced (or longer) courses can explore the material presented in Chapters 8, 9, and 10, bearing in mind that a detailed and rigorous treatment of these topics requires additional reading of the literature mentioned in those chapters. In any case, I would advise against entirely forgoing Chapter 6. Even a cursory reading of this chapter may be beneficial to most students, by illustrating the practical difficulties related to the computation of Bayesian procedures and the corresponding answers brought by simulation methods.

This book took many excruciatingly small steps and exacted a heavy toll on evenings, weekends, and vacations... It is thus only a small indication of my gratitude that this book be dedicated to Brigitte (although she might take this as a propitiatory attempt for future books!!!). Many persons are to be thanked for the present version of this book. First and foremost, Jim Berger's "responsibility" can be traced back to 1987 when he invited me to Purdue University for a year and, as a result, considerably broadened my vision of Statistics; he emphasized his case by insisting very vigorously that I translate the French version and urging me along the whole time. My gratitude to Jim goes very deep when I consider his strong influence in my "coming-of-age" as a statistician. Mary-Ellen Bock, Anirban Das Gupta, Edward George, Gene (formerly Jiunn) Hwang, and Marty Wells were also very instrumental in my progression towards the Bayesian choice, although they do not necessarily support this choice. In this regard, George Casella must be singled out for his strong influence through these years of intense collaboration and friendship, even during his most severe (and "unbearable") criticisms of the Bayesian paradigm! I am also quite grateful to Jean-François Angers, Dean Foster, and Giovanni Parmigiani for taking the risk of using a preliminary version of these notes in their courses, as well as for their subsequent comments. Thanks to Teena Seele for guiding my first steps in TEX, as well as some delicate points in this book—never use \def as an abbreviation of definition! I am also grateful to Elsevier North-Holland for granting me permission to use Diaconis and Ylvisaker's (1985) figures in §3.3. Last, and definitely not least, Kerrie Mengersen and Costas Goutis put a lot of time and effort reading through a preliminary version and provided many helpful comments on content, style, and clarity, while adding a

touch of Ausso-Greek accent to the tone. (In addition, Costas Goutis saved the subject index from utter destruction!) They are thus partly responsible for the improvements over previous versions (but obviously not for the remaining defects!), and I am most grateful to them for their essential help.

Paris, France
May 1994

Christian P. Robert

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

"Sometimes the Pattern has a randomness to it—to our eyes, at least—but what chance that you should meet a man who could guide you in this thing, and he one who could follow the guiding?"

Robert Jordan, *The Eye of the World, Book I of the Wheel of Time.*

1.1 Statistical problems and statistical models

The main purpose of statistical theory is to derive from observations of a random phenomenon an *inference* about the probability distribution underlying this phenomenon. That is, it provides either an analysis (description) of a past phenomenon, or some predictions about a future phenomenon of a similar nature. In this book, we insist on the *decision-oriented* aspects of statistical inference because, first, these analysis and predictions are usually motivated by an objective purpose (whether a company should launch a new product, a racing boat should modify its route, a new drug should be put on the market, an individual should sell shares, etc.) having measurable consequences (monetary results, position at the end of the race, recovery rate of patients, benefits, etc.). Second, to propose inferential procedures implies that one should stand by them, i.e., that the statistician thinks they are preferable to alternative procedures. Therefore, there is a need for an evaluative tool that allows for the comparison of different procedures; this is the purpose of *Decision Theory*. As with most formal definitions, this view of Statistics ignores some additional aspects of statistical practice such as those related to *data collection* (surveys, design of experiments, etc.). This book does, as well, although we do not want to diminish the importance of these omitted topics.

We also insist on the fact that Statistics should be considered an *interpretation* of natural phenomena, rather than an *explanation*. In fact, statistical inference is based on a *probabilistic modeling* of the observed phenomenon and implies a necessarily reductive formalization step since without this probabilistic support it cannot provide any useful conclusion (or decision).

Example 1.1.1 Consider the problem of forest fires. They usually appear at random, but some ecological and meteorological factors influence their eruption. Determining the probability p of fire as a *function* of these factors should help in the prevention of forest fires, even though such modeling is obviously unable to lead to the eradication of forest fires and cannot possibly encompass all the factors involved. A more reductive approach is to assume a parametrized shape for the function p, including physical constraints on the influential factors. For instance, denoting by h the humidity rate, t the average temperature, x the degree of management of the forest, a *logistic model*

$$p = \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)/\left[1 + \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)\right]$$

could be proposed, the statistical step dealing with the evaluation of the parameters $\alpha_1, \alpha_2, \alpha_3$. ‖

To impose a probability modeling on unexplained phenomena seems to be overly reductive in some cases because a given phenomenon can be entirely deterministic, although the regulating function of the process is unknown and cannot be recovered from the observations. This is, for instance, the case with *chaotic phenomena*, where a deterministic sequence of observations cannot be distinguished from a sequence of random variables, in the sense of a statistical test (see Bergé, Pommeau, and Vidal (1984) and Gleick (1987) for introductions to chaos theory). *Pseudo-random generators* are actually based on this indeterminacy. While they use iterative deterministic algorithms such as

$$a_{t+1} = f(a_t),$$

they imitate (or *simulate*) rather well the behavior of a sequence of random variables (see Devroye (1985), Gentle (1998), Robert and Casella (2004), and Appendix B for a list of the most common generators).

Although valid on philosophical grounds, this criticism does not hold if we consider the probabilistic modeling from the *interpretation* perspective mentioned above. This modeling simultaneously incorporates the available information about the phenomenon (influential factors, frequency, amplitude, etc.) *and* the uncertainty pertaining to this information. It thus authorizes a *quantitative* discourse on the problem by providing via probability theory a genuine *calculus of uncertainty* going beyond the mere description of deterministic modelings. This is why a probabilistic interpretation is necessary for statistical inference; it provides a framework replacing a singular phenomenon in the globality of a model and thus allows for analysis and generalizations. Far from being a misappropriation of the inferential purposes, the imposition of a probabilistic structure that is only an approximation of reality is essential for the subsequent statistical modeling to induce a deeper and more adequate understanding of the considered phenomenon.

Obviously, probabilistic modeling can only be defended if it provides an adequate representation of the observed phenomenon. A more down-to-earth criticism of probabilistic modeling is therefore that, even when a modeling is appropriate, it is difficult to know exactly the probability distribution underlying the generation of the observations, e.g., to know that it is normal, exponential, binomial, etc., except in special cases.

Example 1.1.2 Consider a radioactive material with unknown half-life H. For a given atom of this material, the time before disintegration follows exactly an exponential distribution[1] with parameter $\log(2)/H$. The observation of several of these particles can then lead to an inference about H. ‖

Example 1.1.3 In order to determine the number N of buses in a town, a possible inferential strategy goes as follows: observe buses during an entire day and keep track of their identifying number. Then repeat the experiment the next day by paying attention to the number of buses been already observed on the previous day, n. If 20 buses have been observed the first day and 30 the second day, n is distributed as a hypergeometric[1] random variable, $\mathcal{H}(30, N, 20/N)$, and the knowledge of this distribution leads, for instance, to the approximation of N by $20(30/n)$. This method, called *capture-recapture*, has induced numerous and less anecdotal developments in ecology and population dynamics (see Chapter 4). ‖

We could create many other examples where the distribution of the observations is exactly known, its derivation based upon physical, economical, etc., considerations. In the vast majority of cases, however, statistical modeling is reductive in the sense that it only approximates the reality, losing part of its richness but gaining in efficiency.

Example 1.1.4 Price and salary variations are closely related. One way to represent this dependence is to assume the linear relation

$$\Delta P = a + b\,\Delta S + \epsilon,$$

where ΔP and ΔS are the price and salary variations, a and b are unknown coefficients and ϵ is the error factor. A further, but drastic, reduction can be obtained by assuming that ϵ is normally distributed because, while ϵ is indeed a random variable, there are many factors playing a role in the determination of prices and salaries and it is usually impossible to determine the distribution of ϵ. Nonetheless, besides a justification through the *Central Limit Theorem* (i.e., the additional influence of many small factors of similar magnitude), this advanced modeling also allows for a more thorough statistical analysis, which remains valid even if the distribution of ϵ is not exactly normal. (See also Exercise 1.3.) ‖

[1] See Appendix A for a survey of the most common distributions.

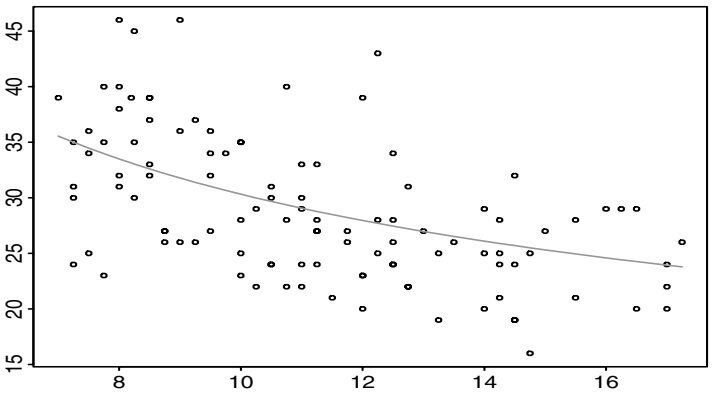

Figure 1.1.1. *Plot of monthly unemployment rate versus number of accidents (in thousands) in Michigan, from 1978 to 1987. (*Source: *Lenk (1999).)*

Example 1.1.5 Consider the dataset depicted by Figure 1.1, which plots the monthly unemployment rate against the monthly number of accidents (in thousands) in Michigan, from 1978 to 1987. Lenk (1999) argues in favor of a connection between these two variates, in that higher unemployment rates lead to less traffic on the roads and thus fewer accidents. A major step towards reduction is then to postulate a parametric structure in the dependence, such as the *Poisson regression* model

$$N|\varrho \sim \mathcal{P}(\exp\{\beta_0 + \beta_1 \log(\varrho)\}), \tag{1.1.1}$$

where N denotes the number of accidents and ϱ the corresponding unemployment rate. Figure 1.1 also depicts the estimated expectation $\mathbb{E}[N|\varrho]$, which tends to confirm the decreasing impact of unemployment upon accidents. But the validity of the modeling (1.1.1) first needs to be assessed, using goodness-of-fit and other model choice techniques (see Chapter 7). ‖

In some cases, the reductive effect is deliberately sought as a positive *smoothing effect* which partly removes unimportant perturbations of the phenomenon and often improves its analysis by highlighting the major factors, as in the following example.

Example 1.1.6 Radiographs can be represented as a 1000×1200 grid of elementary points, called *pixels*, which are grey levels represented by numbers between 0 and 256. For instance, Figure 1.1.6 provides the histogram of the grey levels for a typical chest radiograph. If we consider a pixel to be a discrete random variable taking values in $\{0, 1, \ldots, 256\}$, the histogram gives an approximation of the distribution of this random variable. As shown by the figure, this distribution is rather complex, but approximately bimodal. This particularity is observed on most radiographs and suggests

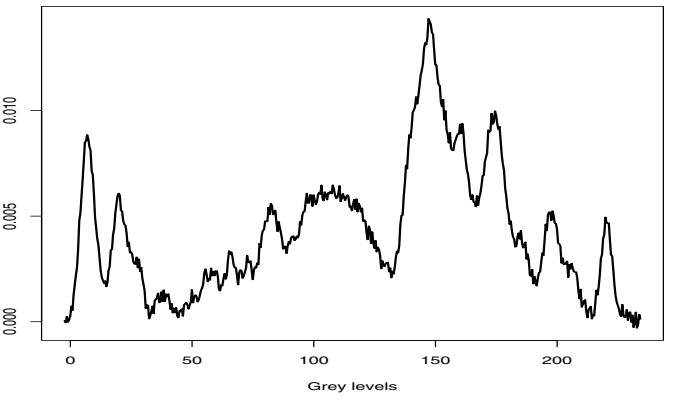

Figure 1.1.2. *Histogram of the grey levels of a chest radiograph and its modeling by a two-component mixture. (Source: Plessis (1989).)*

a modeling of the distribution through a continuous approximation by a *mixture of two normal distributions*, with density

$$(1.1.2)\quad f(x) = \frac{p}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] .$$

Obviously, this modeling considerably smoothes the histogram (see Figure 1.1.6), but also allows a description of the image through five parameters, with no substantial loss of information. The two important modes of the true distribution have actually been shown to correspond to two regions of the chest, the *lungs* and the *mediastinum*. This smoothing technique is used in an image-processing algorithm called *Parametric Histogram Specification* (see Plessis (1989)). We will consider the Bayesian estimation of mixture distributions in detail in Section 6.4. ‖

Given this imperative of reduction of the complexity of the observed phenomenon, two statistical approaches contend. A first approach assumes that statistical inference must incorporate as much as possible of this complexity, and thus aims at estimating the distribution underlying the phenomenon under minimal assumptions, generally using functional estimation (density, regression function, etc.). This approach is called *nonparametric*. Conversely, the *parametric* approach represents the distribution of the observations through a density function $f(x|\theta)$, where only the parameter θ (of finite dimension) is unknown.

We consider that this second approach is more pragmatic, since it takes into account that a finite number of observations can efficiently estimate only a finite number of parameters. Moreover, a parametric modeling authorizes an evaluation of the inferential tools for *finite sample sizes*, contrary to the more involved nonparametric methods, which are usually

justified only asymptotically, therefore strictly only apply when the sample size becomes *infinite* (see, however, Field and Ronchetti (1990), who study the applicability of asymptotic results for finite sample sizes). Of course, some nonparametric approaches, like rank tests (Hajek and Sidàk (1967)), completely evacuate the estimation aspect by devising distribution-free statistics, but their applicability is limited to testing settings.

Both approaches have their interest and we shall not try to justify any further the parametric choice. Quite naturally, there is also an extensive literature on model construction. See Cox (1990) and Lehmann (1990) for references, as well as reflections on the very notion of a statistical model. We will see in Chapter 7 some approaches to the comparison of models, which can be used in the modeling stage, that is, when a model is sought in order to fit the data and several potential models contend.

In this book, we only consider parametric modeling. We assume that the observations supporting the statistical analysis, $x_1, \ldots, x_n$, have been generated from a parametrized probability distribution, i.e., x_i $(1 \leq i \leq n)$ has a distribution with density $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1})$ on $\mathbb{R}^p$, such that the parameter θ_i is unknown and the function f_i is known (see Exercise 1.2 about the formal ambiguity of this definition and Note 1.8.2 for indications about the Bayesian approach to nonparametrics). This model can then be represented more simply by

$$x \sim f(x|\theta),$$

where x is the vector of the observations and θ is the set of the parameters, $\theta_1, \ldots, \theta_n$, which may all be equal. This approach is unifying, in the sense that it represents similarly an isolated observation, dependent observations, and repeated independent and identically distributed (*i.i.d.*) observations $x_1, \ldots, x_n$ from a common distribution, $f(x_1|\theta)$. In the latter case, $x = (x_1, \ldots, x_n)$ and

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Notice that densities of discrete and continuous random variables will be denoted identically in this book, the reference measure being generally provided by the setting. Moreover, we use the notation *"x is distributed according to f"* or "$x \sim f$" instead of *"x is an observation from the distribution with density f"* for the sake of conciseness[2]. Most of the time, the sample is reduced to a single observation, for simplification reasons, but

[2] This book does not follow the usual probabilistic convention that random variables are represented by capital letters, X say, and their *realization*, that is, their observed value, by the corresponding lower case letter, x, as in $P(X \leq x)$. This is because from a Bayesian point of view, we always condition on the realized value x and, besides, consider the parameter, θ say, as a random variable: the use of capital Greek letters then gets confusing in the extreme since Θ is rather, by convention, the parameter space. This also facilitates the use of conditional expressions, which abound in Bayesian computations. In cases potentially provoking confusion, we will revert to the capital-lower case convention.

also because we are usually dealing with distributions where the sample size does not matter, since they allow for sufficient statistics of constant dimension (see Section 1.3 and Chapter 3).

Definition 1.1.7 *A parametric statistical model consists of the observation of a random variable x, distributed according to $f(x|\theta)$, where only the parameter θ is unknown and belongs to a vector space Θ of finite dimension.*

Once the statistical model is defined, the main purpose of the statistical analysis is to lead to an *inference* on the parameter θ. This means that we use the observation x to improve our knowledge on the parameter θ, so that one can take a decision related with this parameter, i.e., either estimate a function of θ or a future event which distribution depends on θ. The inference can deal with some components of θ, precisely ("*What is the value of θ_1?*") or not ("*Is θ_2 larger than θ_3?*"). A distinction is often made between *estimation* and *testing* problems, depending on whether the exact value of the parameters (or of some functions of the parameters), or just a hypothesis about these parameters, is of interest. For instance, the two reference books of classical Statistics, Lehmann (1986) and Lehmann and Casella (1998), deal, respectively, with each of these themes. Other authors have proposed a more subtle distinction between *estimation* and *evaluation* of estimation procedures (see, for instance, Casella and Berger (1990)). More generally, inference covers the random phenomenon directed by θ and thus includes *prediction*, that is, the evaluation of the distribution of a future observation y depending on θ (and possibly the current observation x), $y \sim g(y|\theta, x)$. As shown later, these divisions are somehow artificial, since all inferential problems can be expressed as estimation problems when considered from a decision-theoretic perspective.

The choice of the parametric approach made in this book can be criticized, since we cannot always assume that the distribution of the observations is known up to a (finite dimensional) parameter, but we maintain that this reduction allows for deeper developments in the inferential process, even though this may seem a paradoxical statement. Criticisms on the reductive aspects of the statistical approach and, a fortiori, on the parametric choice, are actually seconded by other criticisms about the choice of the evaluation criteria and the whole purpose of Decision Theory, as we will see in Chapter 2. However, we stand by these choices on the ground that these increasingly reductive steps are minimal requirements for a statistical approach to be coherent (that is, self-consistent). Indeed, the ultimate goal of statistical analysis is, in the overwhelming majority of cases, to support a *decision* as being *optimal* (or at least reasonable). It is thus necessary to be able to compare the different inferential procedures at hand. The next section presents the foundations of Bayesian statistical analysis, which seems to us to be the most appropriate approach for this determination of optimal procedures, while also being the most coherent method[3],

[3] As reported in Robins and Wasserman (2000), there are several formal definitions of

since it builds up these procedures by starting from required properties, instead of the reverse, namely, verifying the good behavior of procedures selected in an ad-hoc manner. The Bayesian choice, as presented in this book, may appear as an unnecessary reduction of the inferential scope, and it has indeed been criticized by many as being so. But we will see in the following chapters that this reduction is both necessary and beneficial. Chapter 11 summarizes various points of a defense of the Bayesian choice, and can be read in connection with the previous arguments[4].

Notice that there also exists a Bayesian approach to nonparametric Statistics. It usually involves prior distributions on functional spaces, such as Dirichlet processes. See Ferguson (1973, 1974) and Escobar (1989), Escobar and West (1994), Dey et al. (1998), Müller et al. (1999) and Note 1.8.2 for references in this area. Example 1.4.3 provides an illustration of the interest of the Bayesian approach in this setting.

1.2 The Bayesian paradigm as a duality principle

Compared[5] with probabilistic modeling, the purpose of a statistical analysis is fundamentally an *inversion* purpose, since it aims at retrieving the causes—reduced to the parameters of the probabilistic generating mechanism—from the effects—summarized by the observations[6]. In other words, when observing a random phenomenon directed by a parameter θ, statistical methods allow to deduce from these observations an *inference* (that is, a summary, a characterization) about θ, while probabilistic modeling characterizes the behavior of the future observations *conditional* on θ. This inverting aspect of Statistics is obvious in the notion of the *likelihood* function, since, formally, it is just the sample density rewritten in the proper order,

$$\ell(\theta|x) = f(x|\theta), \tag{1.2.1}$$

i.e., as a function of θ, which is *unknown*, depending on the observed value x. Historically, the *fiducial approach* of Fisher (1956) also relies on this inversion (see Note 1.8.1).

A general description of the inversion of probabilities is given by *Bayes's Theorem*: If A and E are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}$$

coherence, from Savage (1954) to Heath and Sudderth (1989), which all lead to the conclusion that a procedure is coherent if, and only if, it is Bayesian.

[4] This chapter and Chapter 11 are worth re-reading once the more technical points of the inferential process and the issues at hand have been fully understood.

[5] The word *paradigm*, which is a grammatical term, is used here as an equivalent for *model* or *principles*.

[6] At the time of Bayes and Laplace, i.e., at the end of the eighteenth century, Statistics was often called *Inverse Probability* because of this perspective. See Stigler (1986, Chapter 3).

$$= \frac{P(E|A)P(A)}{P(E)} .$$

In particular,

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)} , \tag{1.2.2}$$

when $P(B) = P(A)$. To derive this result through the machinery of modern axiomatized probability theory is trivial. However, it appears as a major conceptual step in the history of Statistics, being the first *inversion* of probabilities. Equation (1.2.2) expresses the fundamental fact that, for two equiprobable causes, the ratio of their probabilities given a particular effect is the same as the ratio of the probabilities of this effect given the two causes. This theorem also is an actualization principle since it describes the updating of the likelihood of A from $P(A)$ to $P(A|E)$ once E has been observed. Thomas Bayes (1764) actually proved a continuous version of this result, namely, that given two random variables x and y, with conditional distribution[7] $f(x|y)$ and marginal distribution $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\,dy} .$$

While this inversion theorem is quite natural from a probabilistic point of view, Bayes and Laplace went further and considered that the *uncertainty* on the parameters θ of a model could be modeled through a *probability* distribution π on Θ, called *prior distribution*. The inference is then based on the distribution of θ conditional on x, $\pi(\theta|x)$, called *posterior distribution* and defined by

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} . \tag{1.2.3}$$

Notice that $\pi(\theta|x)$ is actually proportional to the distribution of x conditional upon θ, i.e., the likelihood, multiplied by the prior distribution of θ. (It seems that the full generality of (1.2.3) was not perceived by Bayes, but by Laplace, who developed it to a greater extent.) The main addition brought by a Bayesian statistical model is thus to consider a probability distribution on the parameters.

Definition 1.2.1 *A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$, and a prior distribution on the parameters, $\pi(\theta)$.*

In statistical terms, Bayes's Theorem thus actualizes the information on θ by extracting the information on θ contained in the observation x. Its impact is based on the daring move that puts causes (observations) and

[7] We will often replace *distribution* with *density*, assuming that the later is well defined with respect to a natural dominating measure, like the Lebesgue measure. It is only in advanced settings, such as the Haar measure in Chapter 9, that a finer level of measure theory will be needed.

effects (parameters) on the same conceptual level, since both of them have probability distributions. From a statistical modeling viewpoint, there is thus little difference between observations and parameters, since conditional manipulations allow for an interplay of their respective roles. Notice that, historically, this perspective that parameters directing random phenomena can also be perceived as random variables goes against the atheistic determinism of Laplace[8] as well as the clerical position of Bayes, who was a nonconformist minister. By imposing this fundamental modification to the perception of random phenomena, these two mathematicians created modern statistical analysis and, in particular, Bayesian analysis.

Indeed, the recourse to the prior distribution π on the parameters of a model is truly revolutionary. There is in fact a major step from the notion of an *unknown* parameter to the notion of a *random* parameter, and many statisticians place an absolute boundary between the two concepts, although they accept the probabilistic modeling on the observation(s). They defend this point of view on the ground that, even though in some particular settings the parameter is produced under the simultaneous action of many factors and can thus appear as (partly) random, as for instance, in quantum physics, the parameter to be estimated cannot be perceived as resulting from a random experiment in most cases. A typical setting occurs when estimating physical quantities like the speed of light, c. An answer in this particular setting is that the limited accuracy of the measurement instruments implies that the true value of c will never be known, and thus that it is justified to consider c as being uniformly distributed on $[c_0 - \epsilon, c_0 + \epsilon]$, if ϵ is the maximal precision of the measuring instruments and c_0 the obtained value.

We will consider in Chapter 3 some approaches to the delicate problem of prior distribution determination. However, more fundamentally, we want to stress here that the importance of the prior distribution in a Bayesian statistical analysis is not at all that the parameter of interest θ can (or cannot) be perceived as generated from π or even as a random variable, but rather that the use of a prior distribution is the best way to summarize the available information (or even the lack of information) about this parameter, as well as the residual uncertainty, thus allowing for incorporation of this imperfect information in the decision process. (Similar reasoning led Laplace to develop statistical models, despite his determinism.) A more technical point is that the only way to construct a mathematically justified approach operating conditional upon the observations is to introduce a corresponding distribution on the parameters. See also Lindley (1990, §3) for a detailed axiomatic justification of the use of prior distributions.

Let us conclude this section with the historical examples of Bayes and Laplace.

Example 1.2.2 (Bayes (1764)) A billiard ball W is rolled on a line of

[8] *"We must envision the present state of the Universe as the effect of its anterior state and as the cause of the following state"* – Laplace (1795).

length one, with a uniform probability of stopping anywhere. It stops at p. A second ball O is then rolled n times under the same assumptions and X denotes the number of times the ball O stopped on the left of W. *Given X, what inference can we make on p?*

In modern terminology, the problem is then to derive the posterior distribution of p given X, when the prior distribution on p is uniform on $[0,1]$ and $X \sim \mathcal{B}(n,p)$, the binomial distribution (see Appendix A). Since

$$\begin{aligned} P(X = x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \\ P(a < p < b \text{ and } \mathrm{X} = \mathrm{x}) &= \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp \end{aligned}$$

and

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp,$$

we derive that

$$\begin{aligned} P(a < p < b|X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)}\ , \end{aligned}$$

i.e., that the distribution of p conditional upon $X = x$ is a beta distribution, $\mathcal{B}e(x+1, n-x+1)$ (see Appendix A). ‖

In the same spirit, Laplace introduced a probabilistic modeling of the parameter space. But his examples are more advanced than Bayes's, in the sense that the prior distributions Laplace considers are based on abstract reasoning, instead of the physical basis of Bayes's prior distribution.[9]

Example 1.2.3 (Laplace (1773)) An urn contains a number n of black and white cards. If the first card drawn out of the urn is white, what is the probability that the proportion p of white cards is p_0? In his resolution of the problem, Laplace assumes that all numbers from 2 to $n-1$ are equally likely values for pn, i.e., that p is uniformly distributed on $\{2/n, \ldots, (n-1)/n\}$. The posterior distribution of p can then be derived using Bayes's Theorem and

$$\begin{aligned} P(p = p_0|\,\text{data}\,) &= \frac{p_0 \times 1/(n-2)}{\sum_{p=2/n}^{(n-1)/n} p \times 1/(n-2)} \\ &= \frac{n\, p_0}{n(n-1)/2 - 1}. \end{aligned}$$ ‖

[9] It is also possible to picture a more Machiavellian Bayes who picked up this particular example in order to circumvent potential criticisms of the choice of the prior. But it seems that this was not the case, i.e., that Bayes was actually studying this example for its own sake. See Stigler (1986) for more details.

The above choice of the prior distribution can obviously be attacked as being partly arbitrary. However, in Laplace's view of probability theory, most events can be decomposed into elementary *equiprobable* events and, therefore, in this particular case, it seems reasonable to consider the events $\{p = i/n\}$ $(2 \leq i \leq n-1)$ as elementary events. A similar reasoning justifies the following example.

Example 1.2.4 (Laplace (1786)) Considering male and female births in Paris, Laplace wants to test whether the probability x of a male birth is above $1/2$. For $251,527$ male and $241,945$ female births, assuming that x has a uniform prior distribution on $[0,1]$, Laplace obtains

$$P(x \leq 1/2|(251,527; 241,945)) = 1.15 \times 10^{-42}.$$

(see Stigler (1986, p. 134) and Exercise 1.6). He then deduces that this probability x is more than likely to be above 50%. Still assuming a uniform prior distribution on this probability, he also compares the male births in London and Paris and deduces that the probability of a male birth is significantly higher in England. ‖

The following example worked out by Laplace is even more interesting because, from a practical point of view, it provides a method of deriving optimal procedures and, from a theoretical point of view, it is the first formal derivation of a Bayes estimator.

Example 1.2.5 In astronomy, one frequently gets several observations of a quantity ξ. These measurements are independently distributed according to a distribution that is supposed to be unimodal and symmetric around ξ. If we put a uniform distribution on the parameter ξ, it should be a "uniform distribution on $(-\infty, +\infty)$", which is not defined as a probability distribution. However, if we agree on this formal extension (see Section 1.5 for a justification), we can work with the Lebesgue measure on $(-\infty, +\infty)$ instead.

Using this *generalized distribution*, Laplace (1773) establishes that the *posterior median* of ξ, i.e. the median for the distribution of ξ conditional on the observations, is an optimal estimator in the sense that it minimizes the average absolute error

$$\mathbb{E}^{\xi}[\,|\xi - \delta|\,] \tag{1.2.4}$$

in δ, where $\mathbb{E}^{\xi}[\cdot]$ denotes the expectation under the distribution of ξ (see Appendix C for a list of usual notations). This result justifies the use of the posterior median as an estimator of ξ, whatever the distribution of the observation. Although established more than two centuries ago, it is strikingly modern (generality of the distribution, choice of a loss function to evaluate the estimators) and Laplace extended it in 1810 by establishing a similar result for squared error.

Surprisingly, though, Laplace was rather unsatisfied with this result because he still needed the distribution of the observation error to be able to calculate the resulting estimator. He first considered, in 1774, the double

exponential distribution

$$\varphi_\xi(x) = \frac{\xi}{2} e^{-\xi|x|}, \qquad x \in \mathbb{R},\ \xi > 0, \tag{1.2.5}$$

also called the *Laplace distribution*, which supposedly involved the resolution of a fifteenth degree equation for three observations. (Actually, Laplace made a mistake and the correct equation is cubic, as shown by Stigler (1986).) Then, in 1777, he looked at the even less tractable alternative

$$\varphi_\xi(x) = \frac{1}{2\xi} \log\left(\xi/|x|\right) \mathbb{I}_{|x|\le\xi}, \qquad \xi > 0,$$

where $\mathbb{I}$ denotes the indicator function. It was only in 1810 when Legendre and Gauss independently exposed the importance of the *normal distribution*, that Laplace was able to compute his (Bayes) estimators explicitly, since he then thought this was the ideal error distribution. ‖

We will consider again this example, along with other optimality results, in Chapter 2, when we study different loss functions to evaluate estimation procedures and the associated Bayes estimators. Let us stress here that the main consequence of the Bayes and Laplace works has been to introduce the *conditional perspective* in Statistics, i.e., to realize that parameters and observations are fundamentally identical objects, albeit differently perceived.[10] To construct in parallel a probability distribution on the parameter space completes this equivalence and, through Bayes's Theorem, allows a quantitative discourse on the causes, i.e., in our parametric framework an inference on the parameters. As mentioned above, the choice of the prior distribution is delicate, but its determination should be incorporated into the statistical process in parallel with the determination of the distribution of the observation. Indeed, a prior distribution is the best way to include residual information into the model. In addition, Bayesian statistical analysis provides natural tools to incorporate the uncertainty associated with this information in the prior distribution (possibly through a hierarchical modeling, see Chapter 10). Lastly, as pointed out in Lindley (1971), the Bayesian paradigm is intrinsically logical: given a set of required properties represented by the loss function and the prior distribution, the Bayesian approach provides estimators satisfying these requirements, while other approaches evaluate the properties of estimators derived for reasons external to the inferential framework.

1.3 Likelihood Principle and Sufficiency Principle

1.3.1 Sufficiency

Classical Statistics can be envisaged as being directed by principles often justified by "common sense" or additional axioms. On the contrary, the

[10] Again, this is why this book indistinctly writes random variables, observations and parameters in lower case.

Bayesian approach naturally incorporates most of these principles with no restraint on the procedures to be considered, and also definitely rejects other principles, such as *unbiasedness*. This notion once was a cornerstone of Classical Statistics and restricted the choice of estimators to those who are on average correct (see Lehmann and Casella (1998)). While intuitively acceptable, it imposes too stringent conditions on the choice of the procedures and often leads to inefficiency in their performances. (See, e.g., the Stein effect in Note 2.8.2.) More importantly, the number of problems which allow for unbiased solutions is a negligible percentage of all estimation problems (Exercise 1.12). Despite these drawbacks, a recent statistical technique like the bootstrap (Efron (1982), Hall (1992)) was introduced as a way to (asymptotically) reduce the bias.

Two fundamental principles are followed by the Bayesian paradigm, namely the Likelihood Principle and the Sufficiency Principle.

Definition 1.3.1 *When $x \sim f(x|\theta)$, a function T of x (also called a statistic) is said to be sufficient if the distribution of x conditional upon $T(x)$ does not depend on θ.*

A sufficient statistic $T(x)$ contains the whole information brought by x about θ. According to the *factorization theorem*, under some measure theoretic regularity conditions (see Lehmann and Casella (1998)), the density of x can then be written as

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x)),$$

if g is the density of $T(x)$. We will see in Chapter 2 that, when an estimator is evaluated under a convex loss, the optimal procedures only depend on sufficient statistics (this is the *Rao–Blackwell Theorem*). In particular, when the model allows for a *minimal sufficient* statistic, i.e., for a sufficient statistic that is a function of all the other sufficient statistics, we only have to consider the procedures depending on this statistic or equivalently the restricted statistical model associated with this statistic. The concept of sufficiency has been developed by Fisher and is associated with the following principle.

Sufficiency Principle *Two observations x and y factorizing through the same value of a sufficient statistic T, that is, such that $T(x) = T(y)$, must lead to the same inference on θ.*

Example 1.3.2 Consider $x_1, \ldots, x_n$ independent observations from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ (see Appendix A). The factorization theorem then implies that the pair $T(x) = (\bar{x}, s^2)$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

is a sufficient statistic for the parameter (μ, σ), with density

$$g(T(x)|\theta) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-(\bar{x}-\theta)^2 n/2\sigma^2} \frac{(s^2)^{(n-3)/2} e^{-s^2/2\sigma^2}}{\sigma^n \Gamma(n-1/2) 2^{n-1/2}}.$$

Therefore, according to the Sufficiency Principle, inference on μ should only depend on this two-dimensional vector, whatever the sample size n is. We will see in Chapter 3 that the existence of a sufficient statistic of constant dimension is in a sense characteristic of *exponential families*[11]. ‖

Example 1.3.3 Consider $x_1 \sim \mathcal{B}(n_1, p)$, $x_2 \sim \mathcal{B}(n_2, p)$, and $x_3 \sim \mathcal{B}(n_3, p)$, three binomial independent observations when the sample sizes n_1, n_2, and n_3 are known. The likelihood function is then

$$f(x_1, x_2, x_3|p) = \binom{n_1}{x_1}\binom{n_2}{x_2}\binom{n_3}{x_3} p^{x_1+x_2+x_3}(1-p)^{n_1+n_2+n_3-x_1-x_2-x_3}$$

and the statistics

$$T_1(x_1, x_2, x_3) = x_1 + x_2 + x_3 \qquad \text{or} \qquad T_2(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3}$$

are sufficient, on the contrary of $x_1/n_1 + x_2/n_2 + x_3/n_3$. ‖

The Sufficiency Principle is generally accepted by most statisticians, in particular because of the Rao–Blackwell Theorem, which rules out estimators which do not depend only on sufficient statistics. In *model choice* settings, it is sometimes criticized as being too drastically reductive, but note that the Sufficiency Principle is only legitimate when the statistical model is actually the one underlying the generation of the observations. Any uncertainty about the distribution of the observations should be incorporated into the model, a modification which almost certainly leads to a change of sufficient statistics. A similar cautionary remark applies to the Likelihood Principle.

1.3.2 The Likelihood Principle

This second principle is partly a consequence of the Sufficiency Principle. It can be attributed to Fisher (1959) or even to Barnard (1949), but was formalized by Birnbaum (1962). It is strongly defended by Berger and Wolpert (1988) who provide an extended study of the topic. In the following definition, the notion of *information* is to be considered in the general sense of the collection of all possible inferences on θ, and not in the mathematical sense of Fisher information, defined in Chapter 3.

Likelihood Principle *The information brought by an observation x about θ is entirely contained in the likelihood function $\ell(\theta|x)$. Moreover, if x_1 and*

[11] For other distributions, sufficiency is not a very interesting concept, since the dimension of a sufficient statistic is then of the order of the dimension of the observation x (or of the corresponding sample), as detailed in Chapter 3.

x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$\ell_1(\theta|x_1) = c\ell_2(\theta|x_2)$$

for every θ, they then bring the same information about θ and must lead to identical inferences.

Notice that the Likelihood Principle is only valid when

(i) inference is about the *same* parameter θ; and

(ii) θ includes *every* unknown factor of the model.

The following example provides a (now classical) illustration of this principle.

Example 1.3.4 While working on the audience share of a TV series, $0 \le \theta \le 1$ representing the part of the TV audience, an investigator found nine viewers and three nonviewers. If no additional information is available on the experiment, two probability models at least can be proposed:

(1) the investigator questioned 12 persons, thus observed $x \sim \mathcal{B}(12, \theta)$ with $x = 9$;

(2) the investigator questioned N persons until she obtained 3 nonviewers, with $N \sim \mathcal{N}eg(3, 1-\theta)$ and $N = 12$.

In other words, the random quantity in the experiment can be either 9 or 12. (Notice that it could also be both.) The important point is that, for both models, the likelihood is proportional to

$$\theta^3(1-\theta)^9.$$

Therefore, the Likelihood Principle implies that the inference on θ should be identical for both models. As shown in Exercise 1.23, this is not the case for the classical approach. ‖

Since the Bayesian approach is entirely based on the posterior distribution

$$\pi(\theta|x) = \frac{\ell(\theta|x)\pi(\theta)}{\int \ell(\theta|x)\pi(\theta)d\theta}$$

(see (1.2.3) and Section 1.4), which depends on x only through $\ell(\theta|x)$, the Likelihood Principle is automatically satisfied in a Bayesian setting.

On the contrary, the classical or *frequentist* approach[12] focuses on the *average* behavior properties of procedures and thus justifies the use of an estimator for reasons that can contradict the Likelihood Principle. This perspective is particularly striking in testing theory. For instance, if $x \sim$

[12] The theory built up by Wald, Neyman and Pearson in the 1950s is called *frequentist*, because it evaluates statistical procedures according to their long-run performances, that is, on the average (or in *frequency*) rather than focusing on the performance of a procedure for the obtained observation, as a conditional approach would do. The frequentist approach will be considered in detail in Chapters 2 and 5.

$\mathcal{N}(\theta, 1)$ and if the hypothesis to be tested is $H_0 : \theta = 0$, the classical Neyman–Pearson test procedure at level 5% is to reject the hypothesis if $x = 1.96$, on the basis that $P(|x - \theta| \geq 1.96) = 0.05$, thus conditioning on the event $|x| > 1.96$ rather than $x = 1.96$ (which is impossible for the frequentist theory). The frequency argument associated with this procedure is then that, in 5% of the cases when H_0 is true, it rejects wrongly the null hypothesis. Such arguments come to contradict the Likelihood Principle because tail behaviors may vary for similar likelihoods (see Exercises 1.17 and 1.23). The opposition between the frequentist and Bayesian paradigms is stronger in testing theory than in point estimation, where the frequentist approach usually appears as a limiting case of the Bayesian approach (see Chapter 5).

Example 1.3.5 Consider x_1, x_2 i.i.d. $\mathcal{N}(\theta, 1)$. The likelihood function is then

$$\ell(\theta|x_1, x_2) \propto \exp\{-(\bar{x} - \theta)^2\}$$

with $\bar{x} = (x_1 + x_2)/2$. Now, consider the alternative distribution

$$g(x_1, x_2|\theta) = \pi^{-3/2} \frac{e^{-(x_1+x_2-2\theta)^2/4}}{1 + (x_1 - x_2)^2}.$$

This distribution gives a likelihood function proportional to $\ell(\theta|x_1, x_2)$ and therefore should lead to the same inference about θ. However, the distribution g is quite different from $f(x_1, x_2|\theta)$; for instance, the expectation of $(x_1 - x_2)$ is not defined. Therefore, the estimators of θ will have different frequentist properties if they do not depend only on $\bar{x}$. In particular, the confidence regions on θ may differ significantly because of the heavier tails of g. ‖

Example 1.3.6 Another implication of the Likelihood Principle is the *Stopping Rule Principle* in sequential analysis. A *stopping rule* τ can be defined as follows. If the experiments $\mathcal{E}_i$ lead to observations $x_i \in \mathcal{X}_i$, with $x_i \sim f(x_i|\theta)$, consider a corresponding sequence $\mathcal{A}_i \subset \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ such that the criterion τ takes the value n if $(x_1, \ldots, x_n) \in \mathcal{A}_n$, i.e., the experiment stops after the nth observation only if the first n observations are in $\mathcal{A}_n$. The likelihood of $(x_1, \ldots, x_n)$ is then

$$\begin{aligned} \ell(\theta|x_1, \ldots, x_n) &= f(x_1|\theta) f(x_2|x_1, \theta) \\ &\quad \ldots f(x_n|x_1, \ldots, x_{n-1}, \theta) \mathbb{I}_{\mathcal{A}_n}(x_1, \ldots, x_n), \end{aligned}$$

thus depends only on τ through the sample $x_1, \ldots, x_n$. This implies the following principle.

Stopping Rule Principle *If a sequence of experiments, $\mathcal{E}_1, \mathcal{E}_2, \ldots$, is directed by a stopping rule, τ, which indicates when the experiments should stop, inference about θ must depend on τ only through the resulting sample.*

Example 1.3.4 illustrates the case where two different stopping rules lead to the same sample: either the sample size is fixed to be 12, or the experiment is stopped when 9 viewers have been interviewed. Another striking (although artificial) example of a stopping rule is to observe $x_i \sim \mathcal{N}(\theta, 1)$ and to take τ as the first integer n such that

$$|\bar{x}_n| = \left| \sum_{i=1}^{n} x_i/n \right| > 1.96/\sqrt{n}.$$

In this case, the stopping rule is obviously incompatible with frequentist modeling since the resulting sample *always* leads to the rejection of the null hypothesis $H_0 : \theta = 0$ at the level 5% (see Chapter 5). On the contrary, a Bayesian approach avoids this difficulty (see Raiffa and Schlaifer (1961) and Berger and Wolpert (1988, p. 81)). ‖

1.3.3 Derivation of the Likelihood Principle

A justification of the Likelihood Principle has been provided by Birnbaum (1962) who established that it is implied by the Sufficiency Principle, conditional upon the acceptance of a second principle.

Conditionality Principle *If two experiments on the parameter θ, $\mathcal{E}_1$ and $\mathcal{E}_2$, are available and if one of these two experiments is selected with probability p, the resulting inference on θ should only depend on the selected experiment.*

This principle seems difficult to reject when the selected experiment is known, as shown by the following example.

Example 1.3.7 (Cox (1958)) In a research laboratory, a physical quantity θ can be measured by a precise but often busy machine, which provides a measurement, $x_1 \sim \mathcal{N}(\theta, 0.1)$, with probability $p = 0.5$, or through a less precise but always available machine, which gives $x_2 \sim \mathcal{N}(\theta, 10)$. The machine being selected at random, depending on the availability of the more precise machine, the inference on θ when it has been selected should not depend on the fact that the alternative machine *could have been selected*. In fact, a classical confidence interval at level 5% taking into account this selection, i.e., averaging over all the possible experiments, is of half-length 5.19, while the interval associated with $\mathcal{E}_1$ is of half-length 0.62 (Exercise 1.20). ‖

The equivalence result of Birnbaum (1962) is then as follows.

Theorem 1.3.8 *The Likelihood Principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.*

Proof. We define first the *evidence* associated with an experiment $\mathcal{E}$, $Ev(\mathcal{E}, x)$, as the collection of the possible inferences on the parameter θ directing this experiment. Let $\mathcal{E}^*$ denote the *mixed* experiment starting

with the choice of $\mathcal{E}_i$ with probability 0.5 ($i = 1, 2$), thus with result (i, x_i). Under these notations, the Conditionality Principle can be written as

$$Ev(\mathcal{E}^*, (j, x_j)) = Ev(\mathcal{E}_j, x_j). \tag{1.3.1}$$

Consider x_1^0 and x_2^0 such that

$$\ell(\cdot|x_1^0) = c\ell(\cdot|x_2^0). \tag{1.3.2}$$

The Likelihood Principle then implies

$$Ev(\mathcal{E}_1, x_1^0) = Ev(\mathcal{E}_2, x_2^0). \tag{1.3.3}$$

Let us assume that (1.3.2) is satisfied. For the mixed experiment $\mathcal{E}^*$ derived from the two initial experiments, consider the statistic

$$T(j, x_j) = \begin{cases} (1, x_1^0) & \text{if } j = 2,\ x_2 = x_2^0, \\ (j, x_j) & \text{otherwise,} \end{cases}$$

which takes the same value for $(1, x_1^0)$ and for $(2, x_2^0)$. Then, this statistic is sufficient, since, if $t \neq (1, x_1^0)$,

$$P_\theta(X^* = (j, x_j)|T = t) = \mathbb{I}_t(j, x_j)$$

and

$$P_\theta(X^* = (1, x_1^0)|T = (1, x_1^0)) = \frac{c}{1 + c},$$

due to the proportionality of the likelihood functions. The Sufficiency Principle then implies that

$$Ev(\mathcal{E}^*, (1, x_1)) = Ev(\mathcal{E}^*, (2, x_2)) \tag{1.3.4}$$

and, combined with (1.3.1), leads to (1.3.3).

The reciprocal of this theorem can be derived for the Conditionality Principle from the fact that the likelihood functions of (j, x_j) and x_j are proportional, and for the Sufficiency Principle from the factorization theorem. □□

Evans, Fraser and Monette (1986) have shown that the Likelihood Principle can also be derived as a consequence of a stronger version of the Conditionality Principle.

1.3.4 Implementation of the Likelihood Principle

It thus seems quite justified to follow the Likelihood Principle because this principle can be derived from the unassailable Sufficiency and Conditionality Principles. However, this principle is altogether too vague, since it does not lead to the selection of a particular procedure when faced with a given inferential problem. It has been argued that the role of the statistician should stop with the determination of the likelihood function (Box and Tiao (1973)) since it is sufficient for clients to draw their inference, but this extreme view only stands in the most simple cases (or from a Bayesian decisional point of view, if the decision-maker also provides a

prior distribution and a loss function). In large (parameter) dimensions, the likelihood function is also difficult to manipulate because of the lack of proper representations tools.

The vagueness of the Likelihood Principle calls for a reinforcement of the axiomatic bases of the inferential process, i.e., for additional structures in the construction of statistical procedures. For instance, an effective implementation of the Likelihood Principle is the *maximum likelihood estimation* method, as briefly described in Section 1.3.5. Similarly, the Bayesian paradigm allows for implementation of the Likelihood Principle in practice, with the additional advantage of including the decision-related requirements of the inferential problem, and even getting optimal procedures from a frequentist point of view (see below).

If we keep in mind the inversion aspect of Statistics presented in Section 1.2, it is tempting to consider the likelihood as a generalized density in θ, whose mode would then be the maximum likelihood estimator, and to work with this density as with a regular distribution. This approach seems to have been advocated by Laplace when he suggested using the uniform prior distribution when no information was available on θ (see Examples 1.2.3–1.2.5). Similarly, Fisher introduced the fiducial approach (see Note 1.8.1) to try to circumvent the determination of a prior distribution while putting into practice the Likelihood Principle, the choice of his distribution being objective (since depending only on the distribution of the observations). However, this approach is at its most defensible when θ is a location parameter (see also Example 1.5.1), since it leads in general to paradoxes and contradictions, the most immediate being that $\ell(\theta|x)$ is not necessarily integrable as a function of θ (Exercise 1.26). The derivation of objective posterior distributions actually calls for a more advanced theory of *noninformative* distributions (see Chapter 3), which shows that the likelihood function cannot always be considered the most natural posterior distribution.

Many approaches have been suggested to implement the Likelihood Principle like, for instance, *penalized likelihood theory* (Akaike (1978, 1983)) or *stochastic complexity theory* (Rissanen (1983, 1990)). See also Bjørnstad (1990) for a survey of non-Bayesian methods derived from the Likelihood Principle in the prediction area. The overall conclusion of this section is nonetheless that, apart from the fact that many of these theories have a Bayesian flavor, a truly Bayesian approach is the most appropriate to take advantage of the Likelihood Principle (see Berger and Wolpert (1988, Chapter 5) for an extensive discussion of this point).

1.3.5 Maximum likelihood estimation

The Likelihood Principle is altogether distinct from the *maximum likelihood estimation* approach, which is only one of several ways to implement the Likelihood Principle. Because we encounter this technique quite often in the next chapters, and also because it can be situated at the fringe of

the Bayesian paradigm, we recall briefly some basic facts about the maximum likelihood approach. Extended coverage can be found in Lehmann and Casella (1998).

When $x \sim f(x|\theta)$ is observed, the maximum likelihood approach considers the following estimator of θ

$$\hat{\theta} = \arg\sup_{\theta} \ell(\theta|\mathrm{x}), \tag{1.3.5}$$

i.e., the value of θ that maximizes the density at x, $f(x|\theta)$, or, informally, the probability of observing the given value of x. The maximization (1.3.5) is not always possible (see, e.g., the case of a mixture of two normal distributions, which is detailed in Chapter 6) or can lead to several (equivalent) global maxima (see, e.g., the case of a Cauchy distribution, $\mathcal{C}(0,1)$, with two well-separated observations). Nevertheless, the maximum likelihood estimator method is widely used, partly because of this intuitive motivation of maximizing the probability of occurrence and partly because of strong asymptotic properties (*consistency* and *efficiency*). An interesting feature of maximum likelihood estimators is also that they are *parameterization-invariant*. That is to say, for any function $h(\theta)$, the maximum likelihood estimator of $h(\theta)$ is $h(\hat{\theta})$ (even when h is not one-to-one). This property is not enjoyed by any other statistical approach (except by Bayes estimators in the special case of *intrinsic losses*. See Section 2.5.4).

The maximum likelihood method also has drawbacks. First, the practical maximization of $\ell(\theta|x)$ can be quite complex, especially in multidimensional and constrained settings. Consider, for instance, the examples of a mixture of normal distributions, of a truncated Weibull distribution

$$\ell(\theta_1,\theta_2|x_1,\ldots,x_n) = (\theta_1\theta_2)^n (x_1\ldots x_n)^{\theta_1} \exp\left\{-\theta_2 \sum_{i=1}^{n} x_i^{\theta_1}\right\}$$

(see Exercise 1.29), or of a 10×10 table where $x_{ij} \sim \mathcal{N}(\theta_{ij},1)$ when θ_{ij} is increasing in i and j (see Robert and Hwang (1996) and Exercises 1.30 and 1.31). Some numerical procedures, such as the EM algorithm of Dempster et al. (1977) for missing data models or the algorithm of Robertson et al. (1988) for order-restricted parameter spaces, have been tailored to this approach, but unsolved difficulties remain.

Second, a maximization technique is bound to give estimators that lack smoothness, as opposed to integration for instance. This is particularly true when the parameter space is restricted. For example, Saxena and Alam (1982) show that, if $x \sim \chi^2_p(\lambda)$, that is, a noncentral chi-squared distribution with p degrees of freedom[13], the maximum likelihood estimator of λ is equal to 0 for $x < p$. Similarly, maximum likelihood estimators can be quite unstable, i.e., vary widely for small variations of the observations, at least for reduced sample sizes (see Exercise 1.32).

[13] This example also exhibits a limitation of the invariance mentioned above: when $y \sim \mathcal{N}_p(\theta, I_p)$, the maximum likelihood estimator of $\lambda = ||\theta||^2$ is $||y||^2 = x \sim \chi^2_p(\lambda)$, which differs from the maximum likelihood estimator based on x (see Exercise 3.55).

A last but important defect of the maximum likelihood approach is that it lacks decision-theoretic and probabilistic supports. In fact, it does not incorporate the requirements of a decision-theoretic analysis and also fails to provide evaluation tools for the estimators it proposes. For instance, tests are not possible in a purely maximum likelihood context: it is necessary to call for frequentist justifications, even if they are based upon a likelihood ratio (see Section 5.3). Similarly, confidence regions of the form $C = \{\theta;\ \ell(\theta)/\ell(\hat{\theta}) \geq c\}$, which are asymptotically shortest, will not depend solely on the likelihood function if the bound c is to be chosen to achieve coverage at a given level α.

1.4 Prior and posterior distributions

Let us assume at this point that, in addition to the sample distribution, $f(x|\theta)$, a prior distribution on θ, $\pi(\theta)$, is available, that is, that we deal with a complete Bayesian model. Chapter 3 considers the preliminary problem of deriving this distribution from the prior information. Given these two distributions, we can construct several distributions, namely:

(a) the *joint distribution* of (θ, x),

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta)\,;$$

(b) the *marginal distribution* of x,

$$\begin{aligned} m(x) &= \int \varphi(\theta, x)\, d\theta \\ &= \int f(x|\theta)\pi(\theta)\, d\theta\,; \end{aligned}$$

(c) the *posterior distribution* of θ, obtained by Bayes's formula,

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, d\theta} \\ &= \frac{f(x|\theta)\pi(\theta)}{m(x)}\,; \end{aligned}$$

(d) the *predictive distribution* of y, when $y \sim g(y|\theta, x)$, obtained by

$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x) d\theta\,.$$

Example 1.4.1 (Example 1.2.2 continued) If $x \sim \mathcal{B}(n, p)$ and $p \sim \mathcal{B}e(\alpha, \beta)$ (with $\alpha = \beta = 1$ in the particular case of Bayes),

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n, \\ \pi(p) &= \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \leq p \leq 1. \end{aligned}$$

The joint distribution of (x, p) is then

$$\varphi(x,p) = \frac{\binom{n}{x}}{B(\alpha,\beta)} p^{\alpha+x-1}(1-p)^{n-x+\beta-1}$$

and the marginal distribution of x is

$$\begin{aligned} m(x) &= \frac{\binom{n}{x}}{B(\alpha,\beta)} B(\alpha+x, n-x+\beta) \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}, \end{aligned}$$

since the posterior distribution of p is

$$\pi(p|x) = \frac{p^{\alpha+x-1}(1-p)^{\beta+n-x-1}}{B(\alpha+x, \beta+n-x)},$$

i.e., a beta distribution $\mathcal{B}e(\alpha+x, \beta+n-x)$. ‖

Among these distributions, the central concept of the Bayesian paradigm is the *posterior distribution*. In fact, this distribution operates conditional upon the observations, thus operates automatically the *inversion* of probabilities defined in Section 1.2, while incorporating the requirement of the Likelihood Principle. It thus avoids averaging over the unobserved values of x, which is the essence of the frequentist approach. Indeed, the posterior distribution is the updating of the information available on θ, owing to the information contained in $\ell(\theta|x)$, while $\pi(\theta)$ represents the information available a priori, that is, before observing x.

Notice that the Bayesian approach enjoys a specific kind of coherence (we will meet others in the following chapters) in that the order in which i.i.d. observations are collected does not matter (this is a consequence of the Likelihood Principle), but also that updating the prior one observation at a time, or all observations together, does not matter. In other words,

$$\begin{aligned} \pi(\theta|x_1,\ldots,x_n) &= \frac{f(x_n|\theta)\pi(\theta|x_1,\ldots,x_{n-1})}{\int f(x_n|\theta)\pi(\theta|x_1,\ldots,x_{n-1})d\theta} \\ &= \frac{f(x_n|\theta)f(x_{n-1}|\theta)\pi(\theta|x_1,\ldots,x_{n-2})}{\int f(x_n|\theta)f(x_{n-1}|\theta)\pi(\theta|x_1,\ldots,x_{n-2})d\theta} \\ &= \ldots \qquad (1.4.1) \\ &= \frac{f(x_n|\theta)f(x_{n-1}|\theta)\ldots f(x_1|\theta)\pi(\theta)}{\int f(x_n|\theta)f(x_{n-1}|\theta)\ldots f(x_1|\theta)\pi(\theta)d\theta}\,. \end{aligned}$$

It may happen that the observations do not modify the distribution of some parameters. This is obviously the case when the distribution of x does not depend on these parameters, as in some nonindentifiable settings.

Example 1.4.2 Consider one observation x from a normal

$$\mathcal{N}\left(\frac{\theta_1+\theta_2}{2}, 1\right)$$

distribution, with a prior π on (θ_1, θ_2) such that $\pi(\theta_1, \theta_2) = \pi_1(\theta_1 + \theta_2)\pi_2(\theta_1 - \theta_2)$. If we operate the change of variables

$$\xi_1 = \frac{\theta_1 + \theta_2}{2}, \qquad \xi_2 = \frac{\theta_1 - \theta_2}{2},$$

the posterior distribution of ξ_2 is then

$$\begin{aligned} \pi(\xi_2) &\propto \int_{\mathbb{R}} \exp\left\{-(x-\xi_1)^2/2\right\} 2\pi_1(2\xi_1)2\pi_2(2\xi_2)d\xi_1 \\ &\propto \pi_2(2\xi_2) \int_{\mathbb{R}} \exp\left\{-(x-\xi_1)^2/2\right\} \pi_1(2\xi_1)d\xi_1 \\ &\propto \pi_2(2\xi_2) \end{aligned}$$

for every observation x. The observation thus brings no information on ξ_2. ‖

We must warn the reader that *not* every nonidentifiable setting leads to this simple conclusion: depending on the choice of the prior distribution and on the reparameterization of the parameter θ in (θ_1, θ_2), where the distribution of x only depends on θ_1, the marginal posterior distribution of θ_2 may or may not depend on x (Exercise 1.45). An important aspect of the Bayesian paradigm in nonidentifiable settings is, however, that the prior distribution can be used as a tool to *identify* the parts of the parameter that are not covered by the likelihood, even though the choice of prior may have a bearing on the identifiable part.

This invariance from prior distribution to posterior distribution may also occur for some parameters when the number of parameters becomes too large compared with the sample size (Exercise 1.39).

Example 1.4.3 A general setting where this situation occurs is found when the number of parameters is infinite, for instance, when the inference encompasses a whole distribution. Studden (1990) considers n observations $x_1, \ldots, x_n$ from a *mixture of geometric distributions*,

$$x \sim \int_0^1 \theta^x (1-\theta)\, dG(\theta),$$

x taking its values in $\mathbb{N}$ and the probability distribution G being unknown. In this setting, G can be represented by the sequence of its noncentral moments $c_1, c_2, \ldots$. The likelihood function is then derived from $P(X = k) = c_k - c_{k+1}$. Studden (1990) shows that, although the c_i are constrained by an infinite number of inequalities (starting with $c_1 > c_2 > c_1^2$), it is possible to derive (algebraically) independent functions of the c_i's, $p_1, p_2, \ldots$, taking values in $[0, 1]$ and such that c_i only depends on $p_1, \ldots, p_i$ (see Exercise 1.46 for details). Therefore, if the prior distribution of $p_1, p_2, \ldots$ is

$$\pi(p_1, p_2, \ldots) = \prod_{i=1}^{+\infty} \pi_i(p_i),$$

and if the largest observation in the sample is k, the posterior distribution of $p_{k+2}, p_{k+3}, \ldots$ does not depend on the observations:

$$\pi(p_{k+2}, \ldots | x_1, \ldots, x_n) = \pi(p_{k+2}, \ldots) = \prod_{i=k+2}^{+\infty} \pi_i(p_i). \quad \|$$

Conversely, the marginal distribution does not involve the parameter of interest θ. It is therefore rarely of direct use, except in the *empirical Bayesian approach* (see Chapter 10), since the posterior distribution is much more adapted to inferential purposes. The marginal distribution can, however, be used in the derivation of the prior distribution if the available information has been gathered from different experiments, that is, dealing with different θ's as in *meta-analysis* (see Mosteller and Chalmers (1992), Mengersen and Tweedie (1993), and Givens et al. (1997)).

Given a probability distribution π on θ, the Bayesian inferential scope is much larger than the classical perspective. For instance, not only the mean, mode, or median of $\pi(\theta|x)$ can be computed, but also evaluations of the performances of these estimators (through their variance and higher-order moments) are available. Moreover, the knowledge of the posterior distribution also allows for the derivation of *confidence regions* through highest posterior density (HPD) regions, that is, regions of the form

$$\{\theta; \pi(\theta|x) \geq k\},$$

in both unidimensional and multidimensional cases. Similarly, it is possible to derive quite naturally the probability of a hypothesis H_0, by conditioning on the observations, i.e., $P^\pi(\theta \in H_0|x)$. Let us stress that the Bayesian approach is the only one justifying such an expression because the expression $P(\theta = \theta_0) = 0.95$ is meaningless unless θ is a random variable. From a Bayesian point of view, this expression signifies that we are ready to bet that θ is equal to θ_0 with a 95/5 odds ratio, or, in other words, that the uncertainty about the value of θ is reduced to a 5% zone. Chapters 4 and 5 are devoted to the study of estimation techniques that incorporate the decisional requirements. We just illustrate the simplicity of this derivation by constructing a confidence interval in the following example.

Example 1.4.4 Consider $x \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$. Therefore, for a given[14] x,

$$\pi(\theta|x) \quad \propto \quad f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{20}\right)$$

[14] The proportionality symbol $\propto$ is to be taken for functions of θ (not of x). While being entirely rigorous, computations using proportionality signs lead to greater efficiency in the derivation of posterior distributions. In fact, probability densities are uniquely determined by their functional form, and the normalizing constant can be recovered, when necessary, at the end of the computation. This technique will therefore be used extensively in this book. Obviously, it is not always appropriate, for instance when the proportionality constant is 0 or infinity, as seen in Section 1.5.

$$\begin{aligned} &\propto \exp\left(-\frac{11\theta^2}{20} + \theta x\right) \\ &\propto \exp\left(-\frac{11}{20}\left\{\theta - (10x/11)\right\}^2\right) \end{aligned}$$

and $\theta|x \sim \mathcal{N}(\frac{10}{11}x, \frac{10}{11})$. A natural confidence region is then

$$\begin{aligned} C &= \{\theta; \pi(\theta|x) > k\} \\ &= \left\{\theta; \left|\theta - \frac{10}{11}x\right| > k'\right\} . \end{aligned}$$

We can also associate a *confidence level* α with this region in the sense that, if $z_{\alpha/2}$ is the $\alpha/2$ quantile of $\mathcal{N}(0,1)$,

$$C_\alpha = \left[\frac{10}{11}x - z_{\alpha/2}\sqrt{\frac{10}{11}},\ \frac{10}{11}x + z_{\alpha/2}\sqrt{\frac{10}{11}}\right]$$

has a posterior probability $(1-\alpha)$ of containing θ. ‖

We will see in Chapter 10 that a posterior distribution can sometimes be decomposed into several levels according to a hierarchical structure, the parameters of the first levels being treated as random variables with additional prior distributions. But this decomposition is instrumental and does not modify the fundamental structure of the Bayesian model.

A problem we did not mention above is that, although all posterior quantities are automatically defined from a conceptual point of view as integrals with respect to the posterior distribution, it may be quite difficult to provide a numerical value in practice and, in particular, an explicit form of the posterior distribution cannot always be derived. In fact, the complexity of the posterior distributions increases when the parameters are continuous and when the dimension of Θ is large.

These computational difficulties are studied in Chapter 6, where we provide some general solutions. Still, they should not be considered a major drawback of the Bayesian approach. Indeed, Computational Statistics is currently undergoing such a rapid development that we can clearly reject the notion of a prior distribution chosen for its computational tractability, even though we may still rely on these particular distributions to present simpler and clearer examples in this book. On the contrary, it is stimulating to see that we are getting closer to the goal of providing more powerful and efficient statistical tools because of these new computational techniques, as they allow for the use of more complex prior distributions, which are in turn more representative of the available prior information.

1.5 Improper prior distributions

When the parameter θ can be treated as a random variable with known probability distribution π, we saw in the previous section that Bayes's

Theorem is the basis of Bayesian inference, since it leads to the posterior distribution. In many cases, however, the prior distribution is determined on a subjective or theoretical basis that provides a σ-finite measure on the parameter space Θ instead of a probability measure, that is, a measure π such that

$$\int_\Theta \pi(\theta)\, d\theta = +\infty.$$

In such cases, the prior distribution is said to be *improper* (or *generalized*). (An alternative definition of generalized Bayes estimators is considered in Chapter 2.)

When this distribution stems from subjective reasons, that is, when the decision-maker is evaluating the relative likelihoods of different parts of the parameter space Θ (see Chapter 3), it really makes sense that, for large parameter spaces, for instance when Θ is noncountable, the sum of these weights, that is, the measure of Θ, should be infinite.

Example 1.5.1 Consider a distribution $f(x-\theta)$ where the *location parameter* θ is in $\mathbb{R}$ with no restriction. If no prior information is available on the parameter θ, it is quite acceptable to consider that the likelihood of an interval $[a,b]$ is proportional to its length $b-a$, therefore that the prior is proportional to the *Lebesgue measure* on $\mathbb{R}$. This was also the distribution selected by Laplace (see Example 1.2.5). ‖

When such improper prior distributions are derived by automatic methods from the density $f(x|\theta)$ (see Chapter 3), they seem more open to criticism, but let us point out the following points.

(1) These automatic approaches are usually the only way to derive prior distributions in noninformative settings, that is, in cases where the only available (or retained) information is the knowledge of the sample distribution, $f(x|\theta)$. This generalization of the usual Bayesian paradigm thus makes possible a further extension of the scope of Bayesian techniques.

(2) The performances of the estimators derived from these generalized distributions are usually good enough to justify these distributions. Moreover, they often permit recovery of usual estimators like maximum likelihood estimators, thus guaranteeing a closure of the inferential field by presenting alternative approaches at the boundary of the Bayesian paradigm.

(3) The generalized prior distributions often occur as limits of proper distributions (according to various topologies). They can thus be interpreted as extreme cases where the reliability of the prior information has completely disappeared and seem to provide a more *robust* (or more *objective*) answer in terms of a possible *misspecification* of the prior distribution (i.e., a wrong interpretation of the sparse prior information).

(4) Such distributions are generally more acceptable to non-Bayesians, partly for reasons (2) and (3), but also because they may have frequentist justifications, such as:

(i) *minimaxity*, which is related to the usually improper "*least favorable* distributions", defined in Chapter 2);

(ii) *admissibility*, as proper and some improper distributions lead to admissible estimators, while admissible estimators sometimes only correspond to Bayes estimators (see Chapter 8); and

(iii) *invariance*, as the best equivariant estimator is a Bayes estimator for the generally improper *Haar measure* associated with the transformation group (see Chapter 9).

(5) A recent perspective (see, e.g., Berger (2000)) is that improper priors should be preferred to vague proper priors such as, a $\mathcal{N}(0, 100^2)$ distribution say, because the later gives a false sense of safety owing to properness, while lacking robustness in terms of influence on the resulting inference.

These reasons do not convince all Bayesians (see, e.g., Lindley (1965)), but the inclusion of improper distributions in the Bayesian paradigm allows for a closure of the inferential scope (figuratively as well as topologically).

From a more practical perspective, the fact that the prior distribution is improper weakens the above symmetry between the observations and the parameters, but *as long as the posterior distribution is defined*, Bayesian methods apply as well. In fact, the notion of conditional measures is not clearly defined in measure theory, although Hartigan (1983) advocates such an extension, but the convention is to take the posterior distribution $\pi(\theta|x)$ associated with an improper prior π as given by Bayes's formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

when the pseudo marginal distribution $\int_\Theta f(x|\theta)\pi(\theta)\,d\theta$ is well defined. This is an imperative condition for using improper priors, which (almost) always hold for proper priors (Exercise 1.47).

Example 1.5.2 (Example 1.5.1 continued) If $f(x-\theta)$ is the density of the normal distribution $\mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, an arbitrary constant, the pseudo marginal distribution is the measure

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and, by Bayes's formula, the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

i.e., corresponds to $\mathcal{N}(x, 1)$. Notice that the constant ϖ does not play a role in the posterior distribution, and that the posterior distribution is actually the likelihood function. Therefore, even though improper priors cannot be normalized, it does not matter because the constant is of no interest for the statistical inference (but see Chapter 5 for an important exception). ‖

According to the Bayesian version of the Likelihood Principle, only posterior distributions are of importance. Therefore, the generalization from proper to improper prior distributions should not cause problems, in the sense that the posterior distribution corresponding to an improper prior can be used similarly to a regular posterior distributions, *when it is defined*. (Obviously, the interpretation of the prior distribution is more delicate.) For instance, in Example 1.5.1, the relative prior weight of any interval is null, but this does not mean that this interval is unlikely a priori. Actually, a misinterpretation of improper priors as regular prior distributions may lead to difficulties like *marginalization paradoxes* (see Chapter 3) because the usual calculus of conditional probability does not apply in this setting. As expressed by Lindley (1990), *the mistake is to think of them [non-informative priors] as representing ignorance*.

It may happen that, for some observations x, the posterior distribution is not defined (Exercises 1.49–1.52). The usual solution is to determine the improper answer as a limit for a sequence of proper distributions (while also checking the justifications of the improper distribution).

Example 1.5.3 Consider a binomial observation, $x \sim \mathcal{B}(n,p)$, as in the original example of Bayes. Some authors (see Novick and Hall (1965) and Villegas (1977)) reject Laplace's choice of the uniform distribution on $[0,1]$ as automatic prior distribution because it seems to be biased against the extreme values, 0 and 1. They propose to consider instead Haldane's (1931) prior

$$\pi^*(p) \propto [p(1-p)]^{-1}.$$

In this case, the marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$. Therefore, $\pi(p|x)$ does not exist for these two extreme values of x, since the product $\pi^*(p)p^x(1-p)^{n-x}$ cannot be normalized for these two values. For the other values, the posterior distribution is $\mathcal{B}e(x, n-x)$, with posterior mean x/n, which is also the maximum likelihood estimator.

The difficulty in 0 and n can be overcome as follows. The prior measure π^* appears as a limit of unnormalized beta distributions,

$$\pi_{\alpha,\beta}(p) = p^{\alpha-1}(1-p)^{\beta-1},$$

when α and β go to 0. These distributions $\pi_{\alpha,\beta}$ lead to beta posterior distributions, $\mathcal{B}e(\alpha+x, \beta+n-x)$, notwithstanding the lack of the normalizing factor, since the choice of the constant in the prior distribution is irrelevant. The posterior distribution $\pi_{\alpha,\beta}(p|x)$ has the expectation

$$\delta^{\pi}_{\alpha,\beta}(x) = \frac{x+\alpha}{\alpha+\beta+n},$$

which goes to x/n when α and β go to 0. If the posterior mean is the quantity of interest, we can then extend the inferential procedure to the cases $x = 0$ and $x = n$ by taking also x/n as a formal Bayes estimator. ‖

Example 1.5.4 Consider $x \sim \mathcal{N}(0, \sigma^2)$. It follows from invariance considerations that an interesting prior distribution on σ is the measure $\pi(\sigma) = 1/\sigma$ (see Chapter 9). It gives the posterior distribution

$$\pi(\sigma^2|x) \propto \frac{e^{-x^2/2\sigma^2}}{\sigma^2},$$

which is not defined for $x = 0$. However, owing to the continuity of the random variable x, this difficulty is of little importance compared with Example 1.5.3. ‖

Obviously, these limiting arguments are ad-hoc expedients which are not always justified, in particular because the resulting estimator may depend on the choice of the converging sequence. An example of this phenomenon is provided by Richard (1973) (see also Bauwens (1991)) in the case of a normal distribution $\mathcal{N}(\theta, \sigma^2)$, when $\pi(\theta)$ is the Lebesgue measure and σ^{-2} is distributed according to a gamma distribution $\mathcal{G}(\alpha, s_0^2)$, i.e., when

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^{2(\alpha+1)}} e^{-s_0^2/2\sigma^2};$$

the estimator of θ then depends on the behavior of the ratio $s_0^2/(\alpha - 1)$ when both numerator and denominator go to 0.

Moreover, when estimating a *discontinuous* function of θ, the estimator for the limiting distribution may differ from the limit of the estimators. This is, for instance, the case in testing theory with the *Jeffreys–Lindley paradox* (see Chapter 5). Finally, there may be settings such that improper prior distributions cannot be used easily, like in mixture estimation (see Exercise 1.57 and Chapter 6) or in testing theory when testing two-sided hypotheses (see Exercises 1.61–1.63 and Chapter 5).

It is thus important to exercise additional caution when dealing with improper distributions in order to avoid ill-defined distributions. In this book, improper distributions will always be used under the implicit assumption that the corresponding posterior distributions are defined, even though there are settings where this condition could be relaxed (see Note 1.8.3).

The practical difficulty is in checking the propriety (or *properness*) condition

$$\int f(x|\theta)\pi(\theta)\, d\theta < \infty$$

in complex settings like hierarchical models (see Exercise 1.67 and Chapter 10), where the use of improper priors on the upper level of the hierarchy is quite common. The problem is even more crucial because new computational tools like MCMC algorithms (Chapter 6) do not require in practice

this checking of properness (see Note 1.8.3, and Hobert and Casella (1996, 1998)).

Let us stress again that the main justification for using improper prior distributions is to provide a completion of the Bayesian inferential field for subjective, axiomatic (in relation with *complete class results*, see Chapter 8), and practical reasons. This extension does not modify the complexity of the inference, however, because the posterior distribution is truly a probability distribution.

1.6 The Bayesian choice

To close this introduction, let us impress upon the reader that there is such a thing as a Bayesian choice. Thus, it is always possible to adhere to this choice, or to opt for other options. While we are resolutely advocating for this choice, there is no reason to become overly strident. Most statistical theories, such as those presented in Lehmann and Casella (1998), have a reasonable level of coherence and most often agree when the number of observations gets large compared with the number of parameters (see Note 1.8.4).

If we do not present these options here, it is both for philosophical and practical reasons (exposed in Chapter 11), and also for the purpose of presenting an unified discourse on Statistics, where all procedures logically follow from a given set of axioms. This is indeed for us *the* compelling reason for adhering to the Bayesian choice, namely, the ultimate coherence of the axioms of Bayesian statistical inference. By modeling the unknown parameters of the sampling distribution through a probability structure, i.e., by probabilizing uncertainty, the Bayesian approach authorizes a quantitative discourse on these parameters. It also allows incorporation in the inferential procedure of the prior information *and* of the imprecision of this information. Besides, apart from subjective and axiomatic arguments in favor of the Bayesian approach, which is the only system allowing for conditioning on the observations (and thus for an effective implementation of the Likelihood Principle), Bayes estimators are also quintessential for the frequentist optimality notions of Decision Theory. In fact, they can provide essential tools even to those statisticians who reject prior elicitation and the Bayesian interpretation of reality.

1.7 Exercises

Section[15] 1.1

1.1 *(Kelker (1970)) A vector $x \in \mathbb{R}^p$ is distributed according to a *spherically symmetric distribution* if $e.x$ has the same distribution than x for every

[15] The exercises with stars are more advanced, but offer a broader view of the topics treated in each chapter. They can be be treated as useful complements, or as a guided lecture of relevant papers by most readers.

orthogonal transform e.

a. Show that, when a spherically symmetric distribution has a density, it is a function of $x^t x$ only.

b. Show that, if the density of x is $\varphi(x^t x)$, the density of $r = ||x||$ is proportional to
$$r^{p-1}\,\varphi(r^2),$$
and give the proportionality coefficient.

c. Show that, if $x = (x_1', x_2')'$ with $x_1 \in \mathbb{R}^q$ and $x_2 \in \mathbb{R}^{p-q}$, and $||x||^2 = ||x_1||^2 + ||x_2||^2$, the density of $(r_1, r_2) = (||x_1||, ||x_2||)$ is proportional to
$$r_1^{q-1}\, r_2^{p-q-1}\, \varphi\left(r_1^2 + r_2^2\right).$$

d. Deduce that
$$U = \frac{||x_1||^2}{||x_1||^2 + ||x_2||^2}$$
is distributed according to a beta distribution $\mathcal{B}e(q/2,\ (p-q)/2)$.

e. Conclude that
$$\frac{p-q}{q}\,\frac{||x_1||^2}{||x_2||^2}$$
is distributed according to the F-distribution $\mathcal{F}_{p-q,q}$ independently of the spherically symmetric distribution of x. Deduce that the F-ratio is a *robust* quantity in the sense that its distribution is constant on a range of spherically symmetric distributions.

1.2 *(Gouriéroux and Monfort (1996)) This exercise points out that the boundary between parametric and nonparametric models is quite difficult to determine. However, in the second case, the parameter cannot be identified.

a. Show that a c.d.f. is characterized by the values it takes at the rational numbers.

b. Deduce that the collection of the c.d.f.'s on $\mathbb{R}$ has the power of continuum (i.e., the cardinal of the set of the parts of $\mathbb{N}$, the set of natural integers) and thus that all probability distributions on $\mathbb{R}$ can be indexed by a real parameter.

1.3 Show that, if $x_1, \ldots, x_n$ are known explanatory variables and $y_1, \ldots, y_n$ are distributed as $\mathbb{E}[y_i] = bx_i$, the *least-squares estimator* of b, solution of
$$\min_b \sum_{i=1}^n (y_i - bx_i)^2,$$
is also a maximum likelihood estimator under a normality assumption.

1.4 In Example 1.1.3, give the expectation of n. Does that mean that $20 \times 30/n$ is an unbiased estimator of N?

1.5 In Example 1.1.6, show that the moments of $x \sim f(x)$ can be written as $\mathbb{E}[x^k] = p\mathbb{E}[x_1^k] + (1-p)\mathbb{E}[x_2^k]$. Deduce a moment estimator of $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. [*Note:* Historically, this is the estimate of Pearson (1894).]

Section 1.2

1.6 Derive the probabilities of Example 1.2.4 from the approximation

$$\Phi(-x) \simeq \frac{1}{\sqrt{2\pi}x} e^{-x^2/2},$$

which is valid when x is large.

1.7 An examination has 15 questions, each with 3 possible answers. Assume that 70% of the students taking the examination are prepared and answer correctly each question with probability 0.8; the remaining 30% answer at random.

a. Characterize the distribution of S, score of a student if one point is attributed to each correct answer.

b. Eight correct answers are necessary to pass the examination. Given that a student has passed the examination, what is the probability that she was prepared?

1.8 Prove the discrete and continuous versions of Bayes's Theorem.

1.9 *(Romano and Siegel (1986)) The *Simpson paradox* provides an illustration of the need for a conditional approach in Statistics. Consider two medical treatments, T_1 and T_2, T_1 being applied to 50 patients and T_2 to 50 others. The result of the experiment gives the following survival percentages: 40% for treatment T_1, 32% for treatment T_2. Therefore, treatment T_1 seems better because it leads to an higher survival rate. However, if age is taken into account, dividing the subjects between juniors (50) and seniors (50), the success rates are described in the following table:

	T_1	T_2
junior	40	50
senior	10	35

and T_1 is worse than T_2 in both cases. Explain the paradox in terms of Bayes's Theorem.

1.10 Show that the quantity δ that minimizes (1.2.4) is the median of the distribution of ξ. Give the quantity δ that minimizes the average squared error $\mathbb{E}^\xi[(\xi - \delta)^2]$.

1.11 Find the median of the posterior distribution associated with the sampling distribution (1.2.5) and a flat prior $\pi(\xi) = 1$ on ξ. [*Note:* See Stigler (1986) for a resolution.]

Section 1.3

1.12 Show that, for a sample from a normal $\mathcal{N}(\theta, \sigma^2)$ distribution, there does not exist an unbiased estimator of σ but only of powers of σ^2.

1.13 Consider $x \sim P(\lambda)$. Show that $\delta(x) = \mathbb{I}_0(x)$ is an unbiased estimator of $e^{-\lambda}$ which is null with probability $1 - e^{-\lambda}$.

1.14 *A statistic S is said to be *ancillary* if its distribution does not depend on the parameter θ and it is said to be *complete* if $\mathbb{E}_\theta[g(S)] = 0$ for every θ implies $g(s) \equiv 0$. Show that, if S is complete and minimal sufficient, it is independent of every ancillary statistic. [*Note:* This result is called *Basu's Theorem*. The reverse is false.]

1.15 Consider a sample $x_1, \ldots, x_n$ of i.i.d. variables with c.d.f. F.

a. Give the density of the order statistic.

b. Show that $O = (X_{(1)}, ..., X_{(n)})$ is sufficient. What is the conditional distribution of $(X_1, ..., X_n)$ given O?

c. Consider $X_1, ..., X_n$ i.i.d. with totally unknown density. Show that O is then complete.

1.16 Show that a statistic T is sufficient if and only if

$$\ell(\theta|x) \propto \ell(\theta|T(x)).$$

1.17 (Berger and Wolpert (1988, p. 21)) Consider x with support $\{1, 2, 3\}$ and distribution $f(\cdot \mid 0)$ or $f(\cdot \mid 1)$, where

	x		
	1	2	3
$f(x\|0)$	0.9	0.05	0.05
$f(x\|1)$	0.1	0.05	0.85

Show that the procedure that rejects the hypothesis $H_0 : \theta = 0$ (to accept $H_1 : \theta = 1$) when $x = 2, 3$ has a probability 0.9 to be correct (under H_0 as well as under the alternative). What is the implication of the Likelihood Principle when $x = 2$?

1.18 Show that the Stopping Rule Principle given in Example 1.3.6 is a consequence of the Likelihood Principle for the discrete case. [*Note:* See Berger and Wolpert (1988) for the extension to the continuous case.]

1.19 For Example 1.3.6, show that the stopping rule τ is finite with probability 1. (*Hint:* Use the *law of the iterated logarithm*. See Billingsley (1986).)

1.20 Show that the confidence intervals of Example 1.3.7 are correct: under the mixed experiment, $x \sim 0.5\mathcal{N}(\theta, 0.1) + 0.5\mathcal{N}(\theta, 10)$ and $P(\theta \in [x - 5.19, x + 5.19]) = 0.95$, while, under experiment $\mathcal{E}_1$, $x \sim \mathcal{N}(\theta, 0.1)$ and $P(\theta \in [x - 0.62, x + 0.62) = 0.95$.

1.21 (Raiffa and Schlaifer (1961)) Show that, if $z \sim f(z|\theta)$ and if $x = t(z)$, x is a sufficient statistic if and only if for every prior π on θ, $\pi(\theta|x) = \pi(\theta|z)$.

1.22 Consider $x_1, \ldots, x_n$ distributed according to $\mathcal{E}xp(\lambda)$. The data is *censored* in the sense that there exist n random variables $y_1, \ldots, y_n$ distributed according to $f(y)$, independent of λ, and $z_1 = x_1 \wedge y_1, \ldots, z_n = x_n \wedge y_n$ are the actual observations.

a. Show that, according to the Likelihood Principle, the inference on λ should not depend on f.

b. Extend this independence to other types of censoring.

1.23 Compare the lengths of the confidence intervals at level 10% in the setting of Example 1.3.7.

1.24 (Berger (1985a)) In the setting of Example 1.3.4, show that, for the UMPU test of $H_0 : p = 1/2$, the null hypothesis will be accepted or rejected at level 5%, depending on the distribution considered. Deduce that the frequentist theory of tests is not compatible with the Likelihood Principle. (*Hint:* See Chapter 5 for definitions.)

1.25 This exercise aims at generalizing Examples 1.3.4 and 1.3.5 in the continuous case by showing that there can also be incompatibility between the frequentist approach and the Likelihood Principle in continuous settings.

a. If $f(x|\theta)$ is a density such that x is a complete statistic, show that there is no other density $g(x|\theta)$ such that the two likelihood functions $\ell_f(\theta|x) = f(x|\theta)$ and $\ell_g(\theta|x) = g(x|\theta)$ are proportional (in θ) for every x.

b. Consider now a sample $x_1, \ldots, x_n$ from $f(x|\theta)$. We assume that there exists a complete sufficient statistic $T(x_1, \ldots, x_n)$ of dimension 1 and an ancillary statistic $S(x_1, \ldots, x_n)$ such that the couple (T, S) is a one-to-one function of $(x_1, \ldots, x_n)$. Show that, if there exists another density $g(x_1, \ldots, x_n|\theta)$ such that the two likelihood functions are proportional,

$$\ell_g(\theta|x_1, \ldots, x_n) = \omega(x_1, \ldots, x_n)\ell_f(\theta|x_1, \ldots, x_n),$$

the proportionality factor ω only depends on $S(x_1, \ldots, x_n)$.

c. In the particular case when $f(x|\theta)$ is the exponential density, $f(x|\theta) = \theta e^{-\theta x}$, give an example of a density $g(x_1, \ldots, x_n|\theta)$ such that the two likelihood functions are proportional. (*Hint:* Find an ancillary statistic S and derive a function $h(x_1, \ldots, x_n)$ depending only on $S(x_1, \ldots, x_n)$ such that $\mathbb{E}_\theta[h(x_1, \ldots, x_n)] = 1$.)

The following exercises (1.27 to 1.36) present some additional aspects of maximum likelihood estimation.

1.26 Show that, if the likelihood function $\ell(\theta|x)$ is used as a density on θ, the resulting inference does not obey the Likelihood Principle (*Hint:* Show that the posterior distribution of $h(\theta)$, when h is a one-to-one transform, is not the transform of $\ell(\theta|x)$ by the Jacobian rule.)

1.27 Consider a Bernoulli random variable $y \sim \mathcal{B}([1 + e^\theta]^{-1})$.

a. If $y = 1$, show that there is no maximum likelihood estimator of θ.

b. Show that the same problem occurs when $y_1, y_2 \sim \mathcal{B}([1 + e^\theta]^{-1})$ and $y_1 = y_2 = 0$ or $y_1 = y_2 = 1$. Give the maximum likelihood estimator in the other cases.

1.28 Consider x_1, x_2 two independent observations from $\mathcal{C}(\theta, 1)$. Show that, when $|x_1 - x_2| > 2$, the likelihood function is bimodal. Find examples of x_1, x_2, x_3 i.i.d. $\mathcal{C}(\theta, 1)$ for which the likelihood function has three modes.

1.29 The *Weibull distribution* $\mathcal{We}(\alpha, c)$ is widely used in engineering and reliability. Its density is given by

$$f(x|\alpha, c) = c\alpha^{-1}(x/\alpha)^{c-1}e^{-(x/\alpha)^c}.$$

a. Show that, when c is known, this model is equivalent to a gamma model.

b. Give the likelihood equations in α and c and show that they do not allow for explicit solutions.

c. Consider an i.i.d. $x_1, \ldots, x_n$ sample from $\mathcal{We}(\alpha, c)$ censored from the right in y_0. Give the corresponding likelihood function when α and c are unknown and show that there is no explicit maximum likelihood estimators in this case either.

1.30 *(Robertson et al. (1988)) For a sample $x_1, \ldots, x_n$, and a function f on $\mathcal{X}$, the isotonic regression of f with weights ω_i is the solution of the minimization in g of

$$\sum_{i=1}^{n} \omega_i(g(x_i) - f(x_i))^2,$$

under the constraint $g(x_1) \leq \ldots \leq g(x_n)$.

a. Show that a solution to this problem is obtained by the *pool-adjacent-violators* algorithm: if f is not isotonic, find i such that $f(x_{i-1}) > f(x_i)$, replace $f(x_{i-1})$ and $f(x_i)$ by

$$f^*(x_i) = f^*(x_{i-1}) = \frac{\omega_i f(x_i) + \omega_{i-1} f(x_{i-1})}{\omega_i + \omega_{i-1}},$$

and repeat until the constraint is satisfied. Take $g = f^*$.

b. Apply to the case $n = 4$, $f(x_1) = 23$, $f(x_2) = 27$, $f(x_3) = 25$, $f(x_4) = 28$, when the weights are all equal.

1.31 ***(Exercise 1.30 cont.)** The simple *tree-ordering* is obtained when one compares some treatment effects with a control state. The isotonic regression is then obtained under the constraint $g(x_i) \geq g(x_1)$ for $i = 2, \ldots, n$.

a. Show that the following algorithm provides the isotonic regression g^*: if f is not isotonic, assume w.l.o.g. that the $f(x_i)$ are in increasing order ($i \geq 2$). Find the smallest j such that

$$A_j = \frac{\omega_1 f(x_1) + \ldots + \omega_j f(x_j)}{\omega_1 + \ldots \omega_j} < f(x_{j+1})$$

and take $g^*(x_1) = A_j = g^*(x_2) = \ldots = g^*(x_j)$, $g^*(x_{j+1}) = f(x_{j+1})$, ...

b. Apply to the case where $n = 5$, $f(x_1) = 18$, $f(x_2) = 17$, $f(x_3) = 12$, $f(x_4) = 21$ and $f(x_5) = 16$, with $\omega_1 = \omega_2 = \omega_5 = 1$ and $\omega_3 = \omega_4 = 3$.

1.32 (Olkin et al. (1981)) Consider n observations $x_1, \ldots, x_n$ from $\mathcal{B}(k, p)$ where both k and p are unknown.

a. Show that the maximum likelihood estimator of k, $\hat{k}$, is such that

$$(\hat{k}(1-\hat{p}))^n \geq \prod_{i=1}^{n} (\hat{k} - x_i) \quad \text{and} \quad ((\hat{k}+1)(1-\hat{p}))^n < \prod_{i=1}^{n} (\hat{k} + 1 - x_i),$$

where $\hat{p}$ is the maximum likelihood estimator of p.

b. If the sample is $16, 18, 22, 25, 27$, show that $\hat{k} = 99$.

c. If the sample is $16, 18, 22, 25, 28$, show that $\hat{k} = 190$ and conclude on the stability of the maximum likelihood estimator.

1.33 Give the maximum likelihood estimator of p for Example 1.1.6 if the other parameters are known and if there are two observations. Compare with the mean of the posterior distribution if $p \sim \mathcal{U}_{[0,1]}$.

1.34 (Basu (1988)) An urn contains 1000 tickets; 20 are tagged θ and 980 are tagged 10θ. A ticket is drawn at random with tag x.

a. Give the maximum likelihood estimator of θ, $\delta(x)$, and show that $P(\delta(x) = \theta) = 0.98$.

b. Suppose now there are 20 tickets tagged θ and 980 tagged $a_i\theta$ ($i \leq 980$), such that $a_i \in [10, 10.1]$ and $a_i \neq a_j$ ($i \neq j$). Give the new maximum likelihood estimator, δ', and show that $P(\delta'(x) < 10\theta) = 0.02$. Conclude about the appeal of maximum likelihood estimation in this case.

1.35 (Romano and Siegel (1986)) Given

$$f(x) = \frac{1}{x} \exp\left[-50\left(\frac{1}{x} - 1\right)^2\right] \qquad (x > 0),$$

show that f is integrable and that there exist $a, b > 0$ such that

$$\int_0^b af(x)dx = 1 \quad \text{and} \quad \int_1^b af(x)dx = 0.99.$$

For the distribution with density

$$p(y|\theta) = a\theta^{-1} f(y\theta^{-1}) \mathbb{I}_{[0,b\theta]}(y),$$

give the maximum likelihood estimator, $\delta(y)$, and show that $P(\delta(y) > 10\theta) = 0.99$.

1.36 (Romano and Siegel (1986)) Consider x_1, x_2, x_3 i.i.d. $\mathcal{N}(\theta, \sigma^2)$.

a. Give the maximum likelihood estimator of σ^2 if $(x_1, x_2, x_3) = (9, 10, 11)$ or if $(x_1, x_2, x_3) = (29, 30, 31)$.

b. Given three additional observations x_4, x_5, x_6, give the maximum likelihood estimator if $(x_1, \ldots, x_6) = (9, 10, 11, 29, 30, 31)$. Does this result contradict the Likelihood Principle?

Section 1.4

1.37 If $x \sim \mathcal{N}(\theta, \sigma^2)$, $y \sim \mathcal{N}(\varrho x, \sigma^2)$, as in an autoregressive model, with ϱ known, and $\pi(\theta, \sigma^2) = 1/\sigma^2$, give the predictive distribution of y given x.

1.38 If $y \sim \mathcal{B}(n, \theta)$, $x \sim \mathcal{B}(m, \theta)$, and $\theta \sim \mathcal{B}e(\alpha, \beta)$, give the predictive distribution of y given x.

1.39 Given a proper distribution $\pi(\theta)$ and a sampling distribution $f(x|\theta)$, show that the only case such that $\pi(\theta|x)$ and $\pi(\theta)$ are identical occurs when $f(x|\theta)$ does not depend on θ.

1.40 Consider a prior distribution π positive on Θ and $x \sim f(x|\theta)$. Assume that the likelihood $\ell(\theta|x)$ is bounded, continuous, and has a unique maximum $\hat{\theta}(x)$.

a. Show that, when considering a virtual sample $x_n = (x, \ldots, x)$ made of n replications of the original observation x, the posterior distribution $\pi(\theta|x_n)$ converges to a Dirac mass in $\hat{\theta}(x)$.

b. Derive a Bayesian algorithm for computing maximum likelihood estimators.

1.41 *Given a couple (x, y) of random variables, the marginal distributions $f(x)$ and $f(y)$ are not sufficient to characterize the joint distribution of (x, y).

a. Give an example of two different bivariate distributions with the same marginals. (*Hint:* Take these marginals to be uniform $\mathcal{U}([0, 1])$ and find a function from $[0, 1]^2$ to $[0, 1]^2$ which is increasing in both its coefficients).

b. Show that, on the contrary, if the two conditional distributions $f(x|y)$ and $f(y|x)$ are known, the distribution of the couple (x, y) is also uniquely defined.

c. Extend b. to a vector $(x_1, \ldots, x_n)$ such that the full conditionals $f_i(x_i|x_j, j \neq i)$ are known. [*Note:* This result is called the *Hammersley–Clifford* Theorem, see Robert and Casella (2004).]

d. Show that property b. does not necessarily hold if $f(x|y)$ and $f(x)$ are known, i.e., that several distributions $f(y)$ can relate $f(x)$ and $f(x|y)$. (*Hint:* Exhibit a counter-example.)

e. Give some sufficient conditions on $f(x|y)$ for the above property to be true. (*Hint:* Relate this problem to the theory of complete statistics.)

1.42 Consider $x_1, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$. Show that $\sum_{i=1}^n x_i$ is a sufficient statistic and give a confidence region as in Example 1.4.4 when $\pi(\lambda)$ is a $\mathcal{G}(\alpha, \beta)$ distribution. For a given α level, compare its length with an equal tail confidence region.

1.43 Give the posterior and the marginal distributions in the following cases:

(i) $x|\sigma \sim \mathcal{N}(0, \sigma^2)$, $1/\sigma^2 \sim \mathcal{G}(1, 2)$;

(ii) $x|\lambda \sim \mathcal{P}(\lambda)$, $\lambda \sim \mathcal{G}(2, 1)$;

(iii) $x|p \sim \mathcal{N}eg(10, p)$, $p \sim \mathcal{B}e(1/2, 1/2)$.

1.44 Show that, for a sample $x_1, \ldots, x_n$ from a distribution with conditional density $f(x_i|\theta, x_{i-1})$, the actualizing decomposition (1.4.1) also applies. [*Note:* The sequence x_i is then a Markov chain.]

1.45 Show that, in the setting of Example 1.4.2, the marginal posterior distribution on ξ_2 is different from the marginal prior distribution if $\pi(\xi_1, \xi_2)$ does not factorize in $\pi_1(\xi_1)\pi_2(\xi_2)$.

1.46 *(Studden (1990)) In the setting of Example 1.4.3, we define the *canonical moments* of a distribution and show that they can be used as a representation of this distribution.

a. Show that the first two moments c_1 and c_2 are related by the two following inequalities:
$$c_1^2 \leq c_2 \leq c_1$$
and that the sequence (c_k) is monotonically decreasing to 0.

b. Consider a kth degree polynomial
$$P_k(x) = \sum_{i=0}^{k} a_i x^i.$$
Deduce from
$$\int_0^1 P_k^2(x) g(x)\, dx \geq 0 \tag{1.7.1}$$
that
$$a^t C_k a \geq 0, \qquad \forall a \in \mathbb{R}^{k+1}, \tag{1.7.2}$$
where
$$C_k = \begin{pmatrix} 1 & c_1 & c_2 & \ldots & c_k \\ c_1 & c_2 & c_3 & \ldots & c_{k+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_k & c_{k+1} & & \ldots & c_{2k} \end{pmatrix}$$
and $a^t = (a_0, a_1, \ldots, a_k)$.

c. Show that for every distribution g, the moments c_k satisfy
$$\begin{vmatrix} 1 & c_1 & c_2 & \ldots & c_k \\ c_1 & c_2 & c_3 & \ldots & c_{k+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_k & c_{k+1} & & \ldots & c_{2k} \end{vmatrix} > 0. \tag{1.7.3}$$
(*Hint:* Interpret (1.7.2) as a property of C_k.)

d. Using inequalities similar to (1.7.1) for the polynomials $t(1-t)P_k^2(t)$, $tP_k^2(t)$, and $(1-t)P_k^2(t)$, derive the following inequalities on the moments of g:

$$\begin{vmatrix} c_1 - c_2 & c_2 - c_3 & \dots & c_{k-1} - c_k \\ c_2 - c_3 & c_3 - c_4 & \dots & c_k - c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_{k-1} - c_k & \dots & \dots & c_{2k-1} - c_{2k} \end{vmatrix} > 0, \tag{1.7.4}$$

$$\begin{vmatrix} c_1 & c_2 & \dots & c_k \\ c_2 & c_3 & \dots & c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_k & c_{k+1} & \dots & c_{2k-1} \end{vmatrix} > 0, \tag{1.7.5}$$

$$\begin{vmatrix} 1 - c_1 & c_1 - c_2 & \dots & c_{k-1} - c_k \\ c_1 - c_2 & c_2 - c_3 & \dots & c_k - c_{k+1} \\ \dots & \dots & \dots & \dots \\ c_{k-1} - c_k & \dots & \dots & c_{2k-2} - c_{2k-1} \end{vmatrix} > 0. \tag{1.7.6}$$

e. Show that (1.7.3) (resp. (1.7.4)) induces a lower (resp. upper) bound $\underline{c}_{2k}$ (resp. $\bar{c}_{2k}$) on c_{2k} and that (1.7.5) (resp. (1.7.6)) induces a lower (resp. upper) bound $\underline{c}_{2k-1}$ (resp. $\bar{c}_{2k-1}$) on c_{2k-1}.

f. Defining p_k as

$$p_k = \frac{c_k - \underline{c}_k}{\bar{c}_k - \underline{c}_k},$$

show that the relation between $(p_1, ..., p_n)$ and $(c_1, ..., c_n)$ is one-to-one for every n and that the p_i are independent.

g. Show that the inverse transform is given by the following recursive formulas. Let us define

$$q_i = 1 - p_i, \quad \zeta_1 = p_1, \quad \zeta_i = p_i q_{i-1} \qquad (i \geq 2).$$

Then

$$\begin{cases} S_{1,k} = \zeta_1 + \ldots + \zeta_k \quad (k \geq 1), \\ S_{j,k} = \sum_{i=1}^{k-j+1} \zeta_i S_{j-1,i+j-1} \quad (j \geq 2), \\ c_n = S_{n,n}. \end{cases}$$

Section 1.5

1.47 The problem with improper priors that the integral $\int_\Theta f(x|\theta)\pi(\theta)\,d\theta$ may not exist does not appear with proper priors.

a. Recall Fubini's theorem and apply to the couple of functions $(f(x|\theta), \pi(\theta))$.

b. Deduce that, if π is a finite positive measure,

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty \tag{1.7.7}$$

almost everywhere.

c. Show that, if π is improper and $f(x|\theta)$ has a finite support, then $\pi(\theta|x)$ is defined if, and only if, (1.7.7) is finite for every x in the support of $f(x|\theta)$.

1.48 Show that, if π is a positive measure on Θ, the integral (1.7.7) is positive almost everywhere.

1.49 (Fernandez and Steel (1999)) Consider n i.i.d. observations $x_1, \ldots, x_n$ from the mixture

$$p\mathcal{N}(\mu_0, \sigma_0^2) + (1-p)\mathcal{N}(\mu_0, \sigma_1^2)\,,$$

where p, μ_0 and σ_0 are known. The prior on σ_1 is a beta $\mathcal{B}e(\alpha, \beta)$ distribution. Show that, if $r \geq 1$ observations are equal to μ_0, the posterior distribution is only defined when $\alpha > r$. [*Note:* From a measure theoretical point of view, the set of x_i's equal to μ_0 is of measure zero. If one (or more) observation is exactly equal to μ_0, it means that the continuous mixture model is not appropriate.]

1.50 (Exercise 1.49 cont.) Consider an observation x from a normal distribution $\mathcal{N}(0, \sigma^2)$.

a. If the prior distribution on σ is an exponential distribution $\mathcal{E}xp(\lambda)$, show that the posterior distribution is not defined for $x = 0$.

b. If the prior distribution on σ is the improper prior $\pi(\sigma) = \sigma^{-1} \exp(-\alpha\sigma^{-2})$, with $\alpha > 0$, show that the posterior distribution is always defined.

1.51 (Exercise 1.50 cont.) Consider an observation y with $y = x - \lambda$, where x is distributed from Laplace's distribution,

$$f(x|\theta) = \theta^{-1} \exp(-|x|/\theta)\,,$$

and λ is distributed from

$$\pi(\lambda) = |\lambda|^{-1/2} \mathbb{I}_{[-1/2,1/2]}(\lambda)\,.$$

If θ is distributed from a gamma $\mathcal{G}(1/2, a)$ $(a > 0)$, show that, if $y = 0$, the posterior distribution is not defined.

1.52 (Musio and Racugno (1999)) Consider the Poisson $\mathcal{P}(\theta)$ model

$$P_\theta(X = x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, \ldots, \quad \theta > 0,$$

and the prior distribution $\pi(\theta) = 1/\theta$. Show that for $x = 0$, the posterior distribution is not defined.

1.53 (Raiffa and Schlaifer (1961)) Consider a $\mathcal{B}e(\alpha m, (1-m)\alpha)$ prior on $p \in [0, 1]$. Show that, if m is held fixed and α approaches 0, the prior distribution converges to a two-point mass distribution with weight m on $p = 1$ and $(1-m)$ on $p = 0$. Discuss the drawbacks of such a setting.

1.54 (Bauwens (1991)) Consider $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$ and

$$\pi(\theta, \sigma^2) = \sigma^{-2(\alpha+1)} \exp(-s_0^2/2\sigma^2).$$

a. Compute the posterior distribution $\pi(\theta, \sigma^2|x_1, \ldots, x_n)$ and show that it only depends on $\bar{x}$ and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

b. Derive the posterior expectation $\mathbb{E}^\pi[\theta|x_1, \ldots, x_n]$ and show that its behavior when α and s_0 both converge to 0 depends on the limit of the ratio $s_0^2/\alpha - 1$.

1.55 Show that if the prior $\pi(\theta)$ is improper and the sample space $\mathcal{X}$ is finite, the posterior distribution $\pi(\theta|x)$ is not defined for some values of x.

1.56 Consider $x_1, \ldots, x_n$ distributed according to $\mathcal{N}(\theta_j, 1)$, with $\theta_j \sim \mathcal{N}(\mu, \sigma^2)$ $(1 \leq j \leq n)$ and $\pi(\mu, \sigma^2) = \sigma^{-2}$. Show that the posterior distribution $\pi(\mu, \sigma^2|x_1, \ldots, x_n)$ is not defined.

1.57 In the setting of Example 1.1.6, that is, for a mixture of two normal distributions,

a. Show that the maximum likelihood estimator is not defined when all the parameters are unknown.

b. Similarly, show that it is not possible to use an improper prior of the form

$$\pi_1(\mu_1,\sigma_1)\pi_2(\mu_2,\sigma_2)\pi_3(p)$$

to estimate these parameters. (*Hint:* Write the likelihood as a sum of $n+1$ terms, depending on the number of observations allocated to the first component.)

[*Note:* Mengersen and Robert (1996) show that it is possible to use some improper priors by introducing prior dependence between the components.]

1.58 *(**Exercise 1.57 cont.**) For a mixture of two normal distributions (1.1.2), if the prior distribution on the parameters is of the form

$$\pi_1(\mu_1,\sigma_1)\pi_1(\mu_2,\sigma_2)\pi_3(p)$$

and $\pi_3(p)=\pi_3(1-p)$, show that the marginal posterior distribution of (μ_1,σ_1) is the same as the marginal posterior distribution of (μ_2,σ_2), whatever the sample. Deduce that the posterior mean of (μ_1,σ_1) is equal to the posterior mean of (μ_2,σ_2) and therefore that it is not a pertinent estimator. [*Note:* This problem is a consequence of the nonidentifiability of the component labels in a mixture. Solutions involve imposing identifying constraints such as the ordering $\mu_1\le\mu_2$, or using loss functions that are invariant under permutation of the component labels.]

1.59 Construct a limiting argument as in Example 1.5.3 to solve the indeterminacy of Example 1.5.4. Derive the posterior mean.

1.60 Show that, if the prior distribution is improper, the pseudo marginal distribution is also improper.

1.61 *(Hobert and Casella (1998)) Consider a random-effect model,

$$y_{ij}=\beta+u_i+\varepsilon_{ij},\qquad i=1,\dots,I,\ j=1,\dots,J,$$

where $u_i\sim\mathcal{N}(0,\sigma^2)$ and $\varepsilon_{ij}\sim\mathcal{N}(0,\tau^2)$. Under the prior

$$\pi(\beta,\sigma^2,\tau^2)=\frac{1}{\sigma^2\tau^2}\ ,$$

the posterior does not exist.

a. By integrating out the (unobservable) random-effects u_i, show that the full posterior distribution of (β,σ^2,τ^2) is

$$\begin{aligned}\pi(\beta,\sigma^2,\tau^2|y)\propto\ & \sigma^{-2-I}\,\tau^{-2-IJ}\,\exp\left\{-\tfrac{1}{2\tau^2}\,\textstyle\sum_{i,j}(y_{ij}-\bar y_i)^2\right\}\\ &\times\exp\left\{-\frac{J\sum_i(\bar y_i-\beta)^2}{2(\tau^2+J\sigma^2)}\right\}\,(J\tau^{-2}+\sigma^{-2})^{-I/2}\ .\end{aligned}$$

b. Integrate out β to get the marginal posterior density

$$\begin{aligned}\pi(\sigma^2,\tau^2|y)\ &\propto\ \frac{\sigma^{-2-I}\,\tau^{-2-IJ}}{(J\tau^{-2}+\sigma^{-2})^{I/2}}\,(\tau^2+J\sigma^2)^{1/2}\\ &\times\ \exp\left\{-\frac{1}{2\tau^2}\sum_{i,j}(y_{ij}-\bar y_i)^2-\frac{J}{2(\tau^2+J\sigma^2)}\sum_i(\bar y_i-\bar y)^2\right\}.\end{aligned}$$

c. Show that the full posterior is not integrable. (*Hint:* For $\tau\neq0$, $\pi(\sigma^2,\tau^2|y)$ behaves like σ^{-2} in a neighborhood of 0.)

d. Show that the conditional distributions

$$\begin{aligned} U_i|y,\beta,\sigma^2,\tau^2 &\sim \mathcal{N}\left(\frac{J(\bar{y}_i-\beta)}{J+\tau^2\sigma^{-2}}, (J\tau^{-2}+\sigma^{-2})^{-1}\right) , \\ \beta|u,y,\sigma^2,\tau^2 &\sim \mathcal{N}(\bar{y}-\bar{u},\tau^2/JI) , \\ \sigma^2|u,\beta,y,\tau^2 &\sim \mathcal{IG}\left(I/2,(1/2)\sum_i u_i^2\right) , \\ \tau^2|u,\beta,y,\sigma^2 &\sim \mathcal{IG}\left(IJ/2,(1/2)\sum_{i,j}(y_{ij}-u_i-\beta)^2\right) , \end{aligned}$$

are well defined. [*Note:* The consequences of this definition of the full posterior will be clarified in Chapter 6.]

1.62 *Consider a dichotomous probit model, where $(1 \le i \le n)$

$$P(d_i = 1) = 1 - P(d_i = 0) = P(z_i \ge 0) , \tag{1.7.8}$$

with $z_i \sim \mathcal{N}(r_i\beta, \sigma^2)$, $\beta \in \mathbb{R}$, r_i being a covariate. (Note that the z_i's are *not* observed.)

a. Show that the parameter (β, σ) is not identifiable.

b. For the prior distribution $\pi(\beta, \sigma) = 1/\sigma$, show that the posterior distribution is not defined.

c. For the prior distribution

$$\sigma^{-2} \sim \mathcal{G}a(1.5, 1.5) , \qquad \beta|\sigma \sim \mathcal{N}(0, 10^2) ,$$

show that the posterior distribution is well defined.

d. An identifying constraint is to impose $\sigma = 1$ on the model. Give sufficient conditions on the observations (d_i, r_i) for the posterior distribution on β to be defined if $\pi(\beta) = 1$.

e. Same question as d. when the normal distribution on the z_i's is replaced with the logistic function, that is,

$$P(d_i = 1) = 1 - P(d_i = 0) = \frac{\exp(r_i\beta)}{1+\exp(r_i\beta)} ,$$

which gives the dichotomous logit model.

1.63 *(Kubokawa and Robert (1994)) In *linear calibration models*, the interest is in determining values of the regressor x from observed responses y, as opposed to standard linear regression. A simplified version of this problem can be put into the framework of observing the independent random variables

$$y \sim \mathcal{N}_p(\beta, \sigma^2 I_p),\ z \sim \mathcal{N}_p(x_0\beta, \sigma^2 I_p),\ s \sim \sigma^2\chi^2_q , \tag{1.7.9}$$

with $x_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$. The parameter of interest is x_0.

a. A reference prior on (x_0, β, σ) yields the joint posterior distribution

$$\begin{aligned} \pi(x_0,\beta,\sigma^2|y,z,s) \propto\ & \sigma^{-(3p+q)-\frac{1}{2}} \exp\{-(s+\|y-\beta\|^2 \\ & +\|z-x_0\beta\|^2)/2\sigma^2\} \left(1+x_0^2\right)^{-1/2} . \end{aligned}$$

Show that this posterior is compatible with the sampling distribution (1.7.9).

b. Show that the marginal posterior distribution of x_0 is

$$\pi(x_0|y,z,s) \propto \frac{(1+x_0^2)^{(p+q-1)/2}}{\left\{\left(x_0 - \frac{y^t z}{s+\|y\|^2}\right)^2 + \frac{\|z\|^2+s}{\|y\|^2+s} - \frac{(y^t z)^2}{(s+\|y\|^2)^2}\right\}^{(2p+q)/2}} .$$

c. Deduce that the posterior distribution of x_0 is well defined.

[*Note:* See Osborne (1991) for an introduction and review of calibration problems. The model (1.7.9) is also equivalent to Fieller's (1954) problem. See, e.g., Lehmann and Casella (1998).]

Note 1.8.2

1.64 *(Diaconis and Kemperman (1996)) Show that the definition of the Dirichlet process $\mathcal{D}(F_0, \alpha)$ given in 1.8.2 is compatible with the following one: given an i.i.d. sequence x_i from F_0 and a sequence of weights ω_i such that

$$\omega_1 \sim \mathcal{B}e(1,\alpha), \quad \omega_1 + \omega_2 \sim \mathcal{B}e(1,\alpha)\mathbb{I}_{[\omega_1,1]}, \ldots$$

the random distribution

$$F = \sum_{i=1}^{\infty} \omega_i \delta_{x_i}$$

is distributed from $\mathcal{D}(F_0, \alpha)$.

1.65 ***(Exercise 1.64 cont.)** If $F \sim \mathcal{D}(F_0, \alpha)$, the quantity $X = \int xF(dx)$ is a random variable.

a. If $\alpha = 1$ and F_0 is a Cauchy distribution, show that X is also distributed as a Cauchy random variable. [*Note:* This relates to the characterizing property of Cauchy distributions that the average of Cauchy random variables is a Cauchy variable with the same parameters.]

b. If $\alpha = 1$ and $F_0 = \varrho\delta_0 + (1-\varrho)\delta_1$, show that X is distributed as a beta $\mathcal{B}e(\varrho, 1-\varrho)$ random variable.

c. Show that, if $\alpha = 1$ and F_0 is $\mathcal{U}_{[0,1]}$, X has the density

$$\frac{e}{\pi} \frac{\sin(\pi y)}{(1-y)^{(1-y)} y^y} .$$

[*Note:* See Diaconis and Kemperman (1996) for the general formula relating F_0 and the density of X.]

1.66 *(Diaconis and Kemperman (1996)) The Dirichlet process prior $\mathcal{D}(F_0, \alpha)$ can also be described via the so-called *Chinese restaurant process*. Consider a restaurant with many large tables and label each table j with a realization y_j from F_0. Then seat arrivals as follows: the first person to arrive sits at the first table. The $(n+1)$ th person sits at an empty table with probability $\alpha/(\alpha+n)$ and to the right of a seated person with probability $n/(\alpha+n)$.

a. If x_i denotes the label z_j of the table where the i th person sits, show that the sequence $x_1, x_2, \ldots$ is exchangeable (that is, that the distribution is invariant under any permutation of the indices).

b. Show that this sequence can be seen as i.i.d. replications from F, with F distributed from $\mathcal{D}(F_0, \alpha)$, using the conditional distribution given in Note 1.8.2.

c. Show that this definition is also compatible with the definition of Exercise 1.64.

Note 1.8.3

1.67 *(Hadjicostas and Berry (1999)) Consider independent observations x_i $(i = 1, \ldots, n)$ from Poisson distributions $\mathcal{P}oi(\lambda_i t_i)$, where the durations t_i are known. The prior on the λ_i's is a gamma distribution $\mathcal{G}(\alpha, \beta)$ with an independence assumption. The model is *hierarchical* because the parameters (α, β) are supposed to be distributed from a prior distribution $\pi(\alpha, \beta)$ such that

$$\pi(\alpha, \beta) \propto \alpha^{k_1}(\alpha + s_1)^{k_2}\beta^{k_3}(\beta + s_2)^{k_4}\,, \tag{1.7.10}$$

where the values k_i and $s_j > 0$ are known $(i = 1, \ldots, 4,\ j = 1, 2)$.

a. Show that the prior distribution (1.7.10) is proper if, and only if,

$$k_1 + k_2 + 1 < 0, \quad k_1 + 1 > 0, \quad k_3 + k_4 + 1 < 0, \quad k_3 + 1 > 0.$$

b. By integrating out the λ_i's from the joint distribution of the λ_i's and of (α, β), derive the posterior (marginal) distribution of (α, β).

c. Show that the posterior (marginal) distribution of (α, β) is defined (proper) if, and only if,

$$k_1 + y + 1 > 0, \quad k_3 + r + 1 > 0, \quad k_3 > k_1 + k_2$$

and either $k_3 + k_4 + 1 < 0$ or $k_3 + k_4 + 1 = 0$ and $k_1 + y > 0$, where

$$y = \sum_{i=1}^{n} \mathbb{I}_0(x_i), \qquad r = \sum_{i=1}^{n} x_i\,.$$

d. Verify that the conditions of a. imply the conditions of b. (as they should).

e. Show that the conditions of b. are satisfied when $(k_1, \ldots, k_4) = (-8, 0, -5, 0)$ and $(y, r) = (10, 337)$, while the conditions of a. are not satisfied.

f. Show that the conditions of b. are not satisfied when $(k_1, \ldots, k_4) = (-12, 0, 1, 1)$ and $(y, r) = (10, 337)$.

Note 1.8.4

1.68 *(Robins and Ritov (1997)) Consider i.i.d. observations (x_i, y_i) in $(0, 1)^k \times \mathbb{R}$ from the following model: $x \sim f(x)$, $y|x \sim \mathcal{N}(\theta(x), 1)$, with the mean function θ uniformly bounded on $(0, 1)^k$ and f in the set of densities such that $c < f(x) < 1/c$ uniformly on $(0, 1)^k$, where $c < 1$ is a fixed constant. Assume that the quantity of interest is

$$\varphi = \int_{(0,1)^k} \theta(x)dx\,.$$

a. Show that the space Θ of mean functions θ is infinite dimensional.

b. Give the likelihood $\ell(\theta, f)$ and show that it factorizes in a function of f times a function of θ.

c. When f is known, show that $(x_1, \ldots, x_n)$ is ancillary.

d. When f is unknown, show that $(x_1, \ldots, x_n)$ is θ-ancillary, in the sense that the conditional likelihood given $(x_1, \ldots, x_n)$ is a function of θ only, the marginal distribution of $(x_1, \ldots, x_n)$ is a function of f only, and the parameter space is a product space. (See Cox and Hinkley (1974) and Robins and Wasserman (2000) for details about this notion.)

e. When f is known, show that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{f(x_i)}$$

is a consistent estimator of φ. (In fact, it is $\sqrt{n}$ uniformly consistent.)

f. When f is unknown, Robins and Ritov (1997) have shown that there is no uniformly consistent estimator of φ. Deduce that, if the prior distribution on (θ, f) factorizes as $\pi_1(\theta)\pi_2(f)$, the Bayesian inference on θ (and thus φ) is the same whatever the value of f.

g. On the contrary, if the prior distribution on (θ, f) makes θ and f dependent, and if f is known to be equal to f_0, the posterior distribution will depend on f_0. Deduce that this dependence violates the Likelihood Principle.

[*Note:* The previous simplified description of Robins and Ritov (1997) follows from Robins and Wasserman (2000).]

1.8 Notes

1.8.1 A brief history of Bayesian Statistics

Books have been written on the history of Bayesian Statistics, including Stigler (1986), Dale (1991), Lad (1996) and Hald (1998), and we only point out here a few highlights in the development of Bayesian Statistics in the last two centuries.

As detailed in this chapter, the first occurrence of Bayes's formula took place in 1761, in the setting of the binomial example of Section 1.2, exposed by the Reverent Thomas Bayes before the Royal Society, and published posthumously by his friend R. Price in 1763. Pierre Simon Laplace then rediscovered this formula in a greater generality in 1773, apparently ignoring Bayes's previous work. The use of the Bayesian principle then became common in the 19th century, as reported in Stigler (1986), but criticisms started to arise by the end of the 19th century, as for instance in Venn (1886) or Bertrand (1889), focusing on the choice of the uniform prior and the resulting reparameterization paradoxes, as reported by Zabell (1989).

Then, despite further formalizations of the Bayesian paradigm by Edgeworth and Karl Pearson at the turn of the century and Keynes (1921) later, came first Kolmogorov, whose axiomatization of the theory of probabilities in the 1920s seemed to go against the Bayesian paradigm and the notion of subjective probabilities, and second Fisher, who moved away from the Bayesian approach (Fisher (1912)) to the definition of the likelihood function (Fisher (1922)), then to fiducial Statistics (Fisher (1930)), but never revised his opinion on Bayesian Statistics. This is slightly paradoxical, since fiducial Statistics was, in a sense, an attempt to overcome the difficulty of selecting the prior distribution by deriving it from the likelihood function (Seidenfeld (1992)), in the spirit of the *noninformative approaches* of Jeffreys (1939) and Bernardo (1979).

For instance, considering the relation $O = P + \epsilon$ where ϵ is an error term, fiducial Statistics argues that, if P (the cause) is known, O (the effect) is distributed according to the above relation. Conversely, if O is known, $P = O - \epsilon$ is distributed according to the symmetric distribution. In this perspective, observations and parameters play a *symmetric* role, depending on the way the

model is analyzed, i.e., depending on what is known and what is unknown. More generally, the fiducial approach consists of renormalizing the likelihood (1.2.1) so that it becomes a density *in* θ when

$$\int_{\Theta} \ell(\theta|x)\, d\theta < +\infty,$$

thus truly inverting the roles of x and θ. As can be seen in the above example, the argument underlying the causal inversion is totally conditional: conditional upon P, $O = P + \epsilon$ while, conditional upon O, $P = O - \epsilon$. Obviously, this argument does not hold from a probabilistic point of view: if O is a random variable and P is a (constant) parameter, to write $P = O - \epsilon$ does not imply that P becomes a random variable. Moreover, the transformation of $\ell(\theta|x)$ into a density is not always possible. The fiducial approach was progressively abandoned after the exposure of fundamental paradoxes (see Stein (1959), Wilkinson (1977) and the references in Zabell (1992)).

Jeffreys's (1939) book appears as the first modern treatise on Bayesian Statistics: it contains, besides the idea of a noninformative prior, those of predictive distribution, Bayes factors and improper priors. But it came out at the time of Fisher's development of likelihood Statistics and Neyman's (1934) confidence intervals and did not meet with the same success. Alternatives to Bayesian Statistics then became the standard in the 1930s with the introduction of maximum likelihood estimators and the development of a formalized theory of Mathematical Statistics, where prior distributions only appeared as a way of constructing formally optimal estimators as in Wald (1950) or Ibragimov and Has'minskii (1981) (see Chapter 8). Attempts to formalize further the Bayesian approach to Statistics by Gini or de Finetti from the 1930s to the 1970s did not bring it more popularity against the then-dominant Neyman–Pearson paradigm, even though the Bayesian community was growing and produced treatises such as those of Savage (1954) and Lindley (1965, 1971).

It can be argued that it is only recently that Bayesian Statistics got a new impetus, thanks to the development of new computational tools—which have always been central to the Bayesian paradigm—and the rapidly growing interest of practitioners in this approach to statistical modeling, as stressed in Berger's (2000) view of the present and future states of Bayesian Statistics. The vitality of current Bayesian Statistics can be seen through the percentage of Bayesian papers in Statistics journals as well as in other fields. It thus looks as though practitioners in this century will be taking better heed of Bayesian Statistics than their twentieth-century counterparts.

1.8.2 Bayesian nonparametric Statistics

While this book sticks to the parametric approach to Statistics, there is a large literature on Bayesian nonparametric Statistics. First, optimality notions such as minimaxity are central to functional estimation; similar to the parametric case (see Chapter 3), Bayes estimators can be used to determine minimaxity bounds and minimax estimators.

A second and much less formal aspect is to envision Bayesian prior modeling in a infinite dimensional space. This is obviously harder, for mathematical as well as prior construction, reasons. But a first solution is to stand in the grey area between parametric and non-parametric Statistics as in Example 1.4.3: the

number of parameters is finite but grows to infinity with the number of observations. This is, for instance, the case with kernel estimation, where a density is approximated by a mixture

$$\frac{1}{n\sigma}\sum_{i=1}^{n} K\left(\frac{x-x_i}{\sigma}\right),$$

K being a density function, and where σ can be estimated in a Bayesian way, with Hermite expansions (Hjort (1996)), or with wavelets bases (Müller and Vidakovic (1999, Chapter 1)), where a function f is decomposed in a functional basis,

$$f(x) = \sum_i \sum_j \omega_{ij} \Psi\left(\frac{x-\mu_i}{\sigma_j}\right),$$

where Ψ denotes a special function called the *mother wavelet*, like the Haar wavelet

$$\Psi(x) = \mathbb{I}_{[0,1/2)} - \mathbb{I}_{[1/2,1)},$$

where the scale and location parameters μ_i and σ_j are fixed and known, and where the coefficients ω_{ij} can be associated with a prior distribution like (Abramovich et al. (1998))

$$\omega_{ij} \sim \varrho_i \mathcal{N}(0, \tau_i^2) + (1-\varrho_i)\delta_0,$$

where δ_0 denotes the Dirac mass at 0.

A second solution, when estimating a c.d.f. F, is to put a prior distribution on F. The most common choice is to use a Dirichlet distribution $\mathcal{D}(F_0, \alpha)$ on F, F_0 being the prior mean and α the precision, as introduced in Ferguson (1974). This prior distribution enjoys the coherency property that, if $F \sim \mathcal{D}(F_0, \alpha)$, the vector $(F(A_1), \ldots, F(A_p))$ is distributed as a Dirichlet variable in the usual sense $\mathcal{D}_p(\alpha F_0(A_1), \ldots, \alpha F_0(A_p))$ for every partition $(A_1, \ldots, A_p)$. But it also leads to posterior distributions which are partly discrete: if $x_1, \ldots, x_n$ are distributed from F and $F \sim \mathcal{D}(F_0, \alpha)$, the marginal conditional distribution of x_1 given $(x_2, \ldots, x_n)$ is

$$\frac{\alpha}{\alpha+n-1}F_0 + \frac{1}{\alpha+n-1}\sum_{i=2}^{n}\delta_{x_i}.$$

(See also Exercises 1.64 and 1.66 for other characterizations.) The approximation of the posterior distribution requires advanced computational tools that will be developed in Chapter 6. (See Note 6.6.7 for more details.) Other proposals have thus appeared in the literature such as the *generalized Dirichlet distribution* (Hjort (1996)), *Pólya tree priors* (Fabius (1964), Lavine (1992)), *beta process prior* (Hjort (1990)), *Lévy process priors* (Phillips and Smith (1996)).

As a concluding note, let us mention that a recent trend in Bayesian statistics has been to study models with varying dimensions, such as mixtures, hidden Markov models and other dynamic models, as well as neural networks, thanks to new computational tools developed by Grenander and Miller (1994), Green (1995), Phillips and Smith (1996), or Stephens (1997) (see Chapter 6). This is, for instance, the case with mixtures

$$\sum_{i=1}^{k} p_{ik}\varphi(x|\theta_{ik})$$

where $\varphi(\cdot|\theta)$ is a parametrized density, the sum of the weights p_{ik} sum up to 1, and the number of components k is unknown. While this is a well-defined parametric problem, it is closer to nonparametric imperatives than to standard parametric estimation (see Richardson and Green (1997) or Stephens (2000)).

1.8.3 Proper posteriors

It must by now be clear from Section 1.5 that an improper prior π can only be used for inference purposes if (1.7.7) holds for the observation x at hand. If this condition is not satisfied, posterior quantities like the posterior mean or posterior median have no meaning, since, for instance, the ratio

$$\frac{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}$$

is not defined. To verify that (1.7.7) is satisfied in complex models can be quite difficult (see Exercises 1.61 and 1.62) or even simply impossible. Unfortunately, because of computational innovations such as the Gibbs sampler (see Chapter 6), it is possible to work directly from the relation $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ to simulate values from the posterior $\pi(\theta|x)$, but the simulation output does not always signal that the posterior does not exist (see Hobert and Casella (1998)). There are therefore examples in the literature where data have been analyzed with such undefined posteriors, this lack of definition of the posterior been only discovered years later.

We will see in Note 6.6.4, however, that there are good reasons to work with improper posteriors on extended spaces, that is, through a completion of θ in (α, θ), as long as the properness of the posterior $\pi(\theta|x)$ is satisfied.

1.8.4 Asymptotic properties of Bayes estimators

We do not develop the asymptotic point of view in this book for two main reasons, the first of which being that the Bayesian point of view is intrinsically conditional. When conditioning on the observation x, which may be a sample $(x_1, \ldots, x_n)$, there is no reason to wonder what might happen if n goes to infinity since n is fixed by the sample size. Theorizing on future values of the observations thus leads to a frequentist analysis, opposite to the imperatives of the Bayesian perspective. The second point is that, even though it does not integrate asymptotic requirements, Bayesian procedures perform well in a vast majority of cases under asymptotic criteria. It is not so paradoxical that, most often, the Bayesian perspective, and in particular the choice of the prior distribution, cease to be relevant when the number of observations gets infinitely large compared with the number of parameters. (There are well-known exceptions to this ideal setting, as in the *Neyman–Scott problem* of Example 3.5.10, in Diaconis and Freedman (1986), where the number of parameters increases with the number of observations and leads to inconsistent Bayes estimators, or in Robins and Ritov (1997), as detailed in Exercise 1.68.)

In a general context, Ibragimov and Has'minskii (1981, Chapter 1) show that Bayes estimators are *consistent*, that is, that they almost surely converge to the true value of the parameter when the number of observations goes to infinity. This is, for instance, the case with estimators δ_α ($\alpha \geq 1$) that minimize the posterior loss (see Chapter 2) associated with the loss function

$\mathrm{L}(\delta,\theta) = |\theta-\delta|^\alpha$, under fairly weak constraints on the prior distribution π and the sampling density $f(x|\theta)$. Ibragimov and Has'minskii (1981, Chapter 3) also establish (under more stringent conditions) the asymptotic efficiency of some Bayes estimates, that is, that the posterior distribution converges towards the true value at the rate $n^{-1/2}$. (See Schervish (1995) for more details.)

Barron et al. (1999) also give general conditions for consistency of a posterior distribution in the following sense: the posterior probability of every Hellinger neighborhood of the true distribution tends to 1 almost surely when the sample size goes to infinity. (The *Hellinger distance* between two densities f_1 and f_2 (or the corresponding distributions) is defined as

$$d(f_1, f_2) = \int \left(f_1(x)^{1/2} - f_2(x)^{1/2}\right)^2 dx\,.$$

We will use it for decision-theoretic purposes in Chapter 2.) The basic assumption on the prior distribution π is that it gives positive mass to every Kullback–Leibler neighborhood of the true distribution. (We will also use the Kullback–Leibler pseudo-distance in Chapter 2.)

We will come back to asymptotics, nonetheless, in Chapter 3 with the definition of noninformative priors via the asymptotic approximation of tail behaviors, and in Chapter 6 with the *Laplace approximation* to posterior integrals.