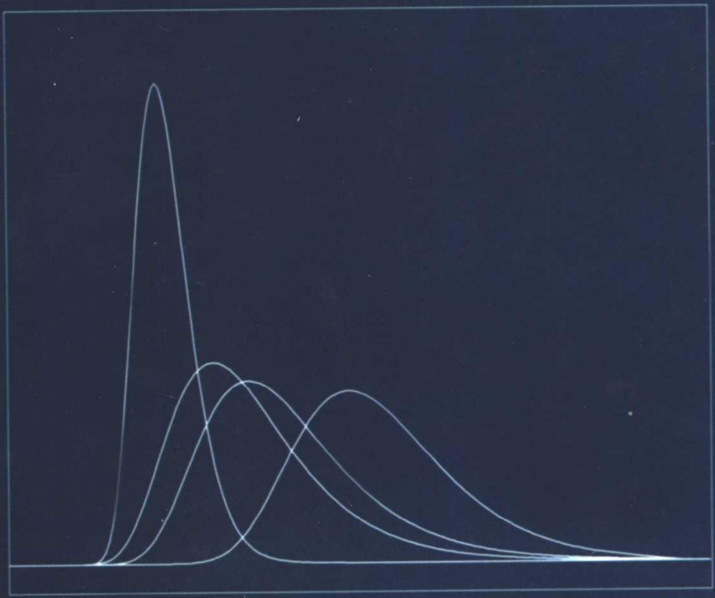
经典原版书库

贝叶斯方法

（英文版）

Cambridge Series in Statistical and Probabilistic Mathematics



Bayesian Methods

An Analysis for Statisticians and Interdisciplinary Researchers

（美） Thomas Leonard 爱丁堡大学 John S. J. Hsu 加州大学圣芭芭拉分校 著

机械工业出版社
China Machine Press

贝叶斯方法 （英文版）

Bayesian Methods
An Analysis for Statisticians and Interdisciplinary Researchers

“本书提供了有关最新现代贝叶斯统计方法的重要题材，文笔流畅，语言优美，其突出的特点是包括大量实际应用，涉及若干领域中AIC和BIC模型选择标准的运用和对比，通过效用理论以独特方式处理贝叶斯决策论，并论述了贝叶斯过程的频度特性，配备了可以扩展与加深书中内容的有趣和适中的自学练习。”

——Michael J. Evans, *Mathematical Review*

“以严密、纯熟的文笔介绍贝叶斯建模的基本原则，选材深思熟虑，按照研究生层次引入贝叶斯方法。”

——*Journal of the American Statistical Association*

贝叶斯“后验分布”或“预测分布”是对有关未知参或未来观测所需了解的每项事物的概括。本书以一种强有力和贴切的方式说明了如何运用贝叶斯统计技术，引导读者从具体数据中推测有关科学、医疗与社会问题的结论。本书解释了贝叶斯方法论所需的一些细微假设，并展示了如何运用这些假设去获取准确结论。本书所介绍的各种方法对计算机模拟的频度特性方面也非常适用。

本书生动地概述了有关费希尔方法（频度方法），同时全面强调了似然性，适合作为主流统计学的教程。本书讲述了效用理论的进展以及时间序列和预测，从而也适合计量经济学的学生阅读。另外，本书还包括线性模型、范畴数据分析、生存竞争分析、随机效应模型和非线性平滑等内容。

本书提供了许多运行实例、自学练习和实际应用，可作为高年级本科生和研究生的教材，同时也可供其他交叉学科的研究人员阅读。

作者简介

Thomas Leonard 于1973年在伦敦大学获得统计学博士学位。他曾在沃里克大学工作过，于1995年担任爱丁堡大学统计学系主席，还曾做过威斯康星-麦迪逊大学统计学教授。20世纪80年代，他最早将拉普拉斯算子引入到贝叶斯方法中。他发表了多篇有关统计学应用方面的论文，并作为统计学专家参与过多个美国法律诉讼案件。

John S. J. Hsu 加州大学圣芭芭拉分校统计学与应用概率论副教授、统计实验室主任，擅长研究应用问题，还建立了贝叶斯理论研究计划。由于在log-线性模型分析方面的贡献，他获得了爱丁堡大学的名誉职位。在Thomas Leonard和Kam-Wah Tsui的指导下，他于1990年在威斯康星-麦迪逊大学获得统计学博士学位。

封面设计 吴刚

ISBN 7-111-15832-6
9 787111 158325

HZ BOOKS
华章图书

华章网站 http://www.hzbook.com
网上购书：www.china-pub.com
北京市西城区百万庄南街1号 100037
读者服务热线：(010) 68995259 68995264
读者服务信箱：hzedu@hzbook.com

ISBN 7-111-15832-6/O · 397
定价：45.00元

经典原版书库

贝叶斯方法

(英文版)

Bayesian Methods

An Analysis for Statisticians and Interdisciplinary Researchers

(美) Thomas Leonard 爱丁堡大学
John S. J. Hsu 加州大学圣芭芭拉分校 著

机械工业出版社
China Machine Press

Thomas Leonard and John S. J. Hsu: Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers (ISBN 0-521-00414-4).

Originally published by Cambridge University Press in 1999.

This reprint edition is published with the permission of the Syndicate of the Press of the University of Cambridge, Cambridge, England.

本书版权登记号：图字：01-2005-0675

图书在版编目（CIP）数据

贝叶斯方法（英文版）/（美）伦纳德（Leonard, T.）等著. －北京：机械工业出版社，2005.1

（经典原版书库）

书名原文：Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers

ISBN 7-111-15832-6

Ⅰ. 贝…　Ⅱ. 伦…　Ⅲ. 贝叶斯推断－英文　Ⅳ. O212

中国版本图书馆CIP数据核字（2004）第135141号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）
责任编辑：迟振春
北京中兴印刷有限公司印刷·新华书店北京发行所发行
2005年1月第1版第1次印刷
787mm × 1092mm 1/16 · 21.5印张
印数：0 001-3 000 册
定价：45.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换
本社购书热线：（010）68326294

Preface

Statistics uses theoretical models and techniques to help applied researchers to extract, and infer, real-life, scientific, medical, and social conclusions from numerical data, which are subject to random uncertainty. For any particular study, it is important to combine theoretical and computational resources, together with applied skills, and an ability to interact with experts with knowledge relating to the background and usefulness of the data.

Many studies and data sets are nonstandard, and it is not always possible to provide a completely convincing analysis based upon preexisting techniques. Therefore, statisticians frequently need to develop new techniques, on line, for a particular practical study. Furthermore, the statistical state of the art is continuously evolving, and it is therefore important for researchers to continue to develop the available statistical methodology. Finally, when existing methodology is available, it is important that this should be applied with specific knowledge of the subtleties of the assumptions involved, together with their consequences.

There are nowadays two main streams of statistical thought. We will refer to these as the "Fisherian" and the "Bayesian" philosophies. The Fisherian philosophy is named after Sir Ronald Fisher and combines the "frequency approach" (unbiased estimators, hypothesis tests, and confidence intervals) with likelihood methods. The Fisherian philosophy also includes the "fiducial approach," an incomplete method, suggested by Fisher, which attempts to achieve some of the advantages of the Bayesian approach (e.g., good conditional inference, given the observed values of the data, combined with appealing frequency properties when repeating the experiment a number of times under identical conditions), but without the assumption of a "prior distribution."

The Bayesian philosophy is named after the Reverend Thomas Bayes and refers to such concepts as "prior and posterior knowledge," "prior, posterior, and predictive distributions," and "Bayes decision rules and estimators." The Bayesian approach possesses many advantages, even when viewed from a Fisherian viewpoint, in particular its inherent long-run frequency properties. In practical terms, this means that if computer simulations are used to compare the mean squared error, prediction error, coverage probability, or power of different procedures, then Bayesian methods can perform remarkably well. This validation is an essential ingredient, when combined with the construction of statistical techniques, and provides just one substantial justification of the Bayesian paradigm. Other advantages are summarized by Berger (1985) and Bernardo and Smith (1994), and in our introductions to Chapters 2, 3, 5, and 6 of the current text.

Chapter 1 describes a number of Fisherian procedures, which comprise important background to the Bayesian approach. It is, for example, essential for the reader to be able to construct and understand likelihood functions before attempting Bayesian techniques. The reader should also understand basic data analysis.

Chapter 2 provides an easy introduction to Bayesian ideas and utilizes easy forms of Bayes' theorem when the parameter space is discrete. These are of particular importance in medical and legal applications.

Chapter 3 develops the Bayesian paradigm when there is a single unknown parameter. In such cases, a univariate probability distribution readily summarizes the posterior information. Frequency properties of related estimators and decision rules are developed.

Chapter 4 provides a break to some of the technicalities and considers the "expected utility hypothesis" and its role in financial decision making. Some extensions to the expected utility hypothesis are considered.

Chapter 5 extends the ideas of Chapter 3 to statistical models with several parameters. Approaches to the linear statistical model, categorical data analysis, and time-series analysis are included.

Chapter 6 provides advanced studies of prior structures, posterior smoothing, and Bayes–Stein estimation. Many of the techniques again achieve appealing frequency properties. Computational techniques, already mentioned in Chapter 5, for approximating or simulating high-dimensional numerical integrations, for example, for providing adequate finite sample size analyses of nonlinear models, are developed further. These include Laplacian methods, importance sampling, and Markov Chain Monte Carlo Methods (MCMC).

The text contains 49 worked examples and 148 self-study exercises, which relate to special cases of methodology more broadly explained in the main body of the text. The reader is thereby provided with layers of knowledge, which can be studied at different levels. The volume progressively develops a number of special themes in a possibly unique manner. A large number of further practical examples are described throughout the text.

The bibliography integrates Bayesian statistics with other statistical methodologies and with interdisciplinary research. While the Bayesian references represent the last four decades of research, they do not provide an exhaustive reference list for the Bayesian literature.

Much of the material in this text has been previously taught to graduate students in statistics, economics, and business attending a Bayesian Decisions course at the University of Wisconsin–Madison, and to graduate students attending a Bayesian Inference course at the University of California at Santa Barbara. The text is also appropriate for the following readerships:

- Students attending a statistics course with a mixture of Fisherian and Bayesian philosophies, at final-year undergraduate or at Master's-degree level. In this case, the instructors should concentrate on the easier parts of Chapter 1, together with Chapters 2 and 3, and the easier parts of Chapter 5. If the course is taught within an economics graduate program, then Chapter 4 and Sections 5.3–5.7 will also be of interest, together with the simulation procedures of Chapter 6.

- Interdisciplinary research specialists wishing to develop statistical models and analyses relating to their own area. We have previously used techniques described in this text for interdisciplinary research in many areas, including geology, psychometrics, medicine, animal science, genetics, biology, archaeology, forensic science, civil engineering, plant science, pathology, and physics. We have been involved in many practical collaborations, as directors of statistical laboratories at the University of Edinburgh and The University of California, with these objectives in mind.
- Doctoral students, and other researchers, in statistics. For example, Chapters 5 and 6 will help you to achieve the research frontiers in Bayesian statistics. Chapters 3, 5, and 6 would provide useful material for an advanced graduate course in statistics.

The first co-author wishes to acknowledge his mentors Anne F. S. Mitchell, Dennis V. Lindley, and A. Philip Dawid for teaching him Bayesian statistics at Imperial College and University College, London. His early Bayesian ideas, also frequently employed in this volume, were further influenced by James M. Dickey, Irving Jack Good, Adrian F. M. Smith, Tony O'Hagan, Jim Q. Smith, Patricia M. E. Altham, and P. Jeffrey Harrison. The second co-author wishes to acknowledge David V. Hinkley and Raisa Feldman for their encouragement. Both co-authors are indebted to Arnold Zellner, George Tiao, and Kam-Wah Tsui for their outstanding help and encouragement. They would also like to thank George E. P. Box, Jeff C. F. Wu, Irwin Guttman, Colin G. Aitken, Grace Wahba, Nan Laird, Michael Newton, Greg Reinsel, Bob Miller, Douglas Bates, and Richard A. Johnson for their previous advice on Bayesian and other related methods contained in this volume. Peter Lee has provided very helpful information in relation to his own writings. Suggestions by Bob Barmisch, Robert McCullough, Peter Wakker, Derek Arthur, and John Searle are indicated in the text. Jerome Klotz has advised us on gambling with roulette. Orestis Papasouliotis collaborated on some of the recent methodological developments and prepared the mathematics and computer program for the graphs in the cover design (these are the posterior densities of the group means in an analysis of covariance model). Geoff McLachlan kindly provided us with a copy of his computer package for multivariate mixtures. Rod Leonard described valuable insights regarding the problem of spurious correlation in the context of the chemical industry.

We should also acknowledge the many graduate students attending our Bayesian courses who have helped or advised us over the years. These include, but are not limited to, Jean Deichtmann, Josep Ginebra-Molins, Robert Tempelman, Taskin Atilgan, Christian Ritter, Tom Chiu, and Jen-Ting Wang.

We would like to thank the following publishers and associations for granting us permission to reproduce previously published material: John Wiley & Sons for Figure 5.2.1 from T. Leonard and J. S. J. Hsu (1994), The Bayesian analysis of categorical data – a selective review, *in* P. R. Freeman and A. F. M. Smith (eds.), *Aspects of Uncertainty: A Tribute to D. V. Lindley*, copyright John Wiley & Sons Limited; the American Educational Research Association for Tables 5.2.4 and 5.2.5 from T. Leonard and M. R. Novick (1986), Bayesian full rank marginalization for two-way contingency tables, *Journal of Educational Statistics* **11**: 33–56; the Royal Statistical Society for Tables

5.2.7 and 5.2.8 from T. Leonard (1975), Bayesian estimation methods for two-way contingency tables, *Journal of Royal Statistical Society*, Ser. B, **37**: 23–37; the American Statistical Association for Table 6.5.1 from L. Sun, J. S. J. Hsu, I. Guttman, and T. Leonard (1996), Bayesian methods for variance component models, *Journal of the American Statistical Association* **91**: 743–52; and the Institute of Statistical Mathematics for Table 6.5.2 from T. Leonard (1984), Some data-analytic modifications to Bayes-Stein estimation, *Annals of the Institute of Statistical Mathematics* **36**: 11–21.

We would also like to thank Gloria Scallissi for her excellent typing of the first draft of the manuscript, Lauren Cowles for her continued encouragement on behalf of the publishers, David Tranah for some helpful suggestions, and Rena S. Wells for her excellent copyediting.

Contents

1

Introductory Statistical Concepts

1.0 Preliminaries and Overview

When analyzing numerical data, subject to random uncertainty, which have been collected in some scientific or real-life context, the first "golden rule" is to study the data, for example, using dotplots, bivariate scatterplots, relative frequency histograms, and contingency tables, before applying any formal statistical technique. Complicated data sets deserve several hours, days, or even weeks of study. When studying a data set, you should realize that data are not simply numbers but rather measurements or counts of real entities (e.g., birthweights of babies, numbers of students passing a college test, a measurement of a real chemical). Therefore, any tentative conclusions should be made in the contexts of their meaning in relation to these entities, the real background of the data, and how the data were collected. The same set of numerical data might mean something entirely different in different scientific or real-life contexts.

Sometimes, upon viewing the data, you may discover a particularly distinctive feature that yields a decisive conclusion. In this case, it may not be necessary, or indeed technically feasible, to proceed to a more formal analysis. For example, when investigating the years of service of French generals during the late eighteenth century (see Wetzler, 1983, Appendix), the conclusion was reached, upon viewing a distinctive spike in the scatterplot, that a number of the generals had been rather abruptly dismissed during the French Revolution. As another example, the State of Wisconsin was advised during a court action in 1986, and based upon data for a carefully collected random sample of $n = 120$ nursing homes, that the state was not adequately reimbursing the actual costs (in dollars per patient per day) for nursing homes with costs in excess of \$45. This conclusion was validated by a distinctive blip in an otherwise linear bivariate scatterplot. The State of Wisconsin conceded the case, largely because the state recognized that data based upon a representative sample (a random sample of 120 nursing homes from 600 nursing homes in Wisconsin) had been collected. George and Wecker (1985) also emphasize the importance of using good statistics in legal cases.

For the first of these analyses, a computer package was not used, since formal summary statistics such as the sample mean and variance would not be particularly relevant. For the second analysis, any attempt to fit a regression model without first carefully considering the data could have led to many hours of fruitless analysis. Similarly, you should try to avoid any "black box" data analytic technique that cannot be combined with an interaction between your thought processes and the data. It is particularly

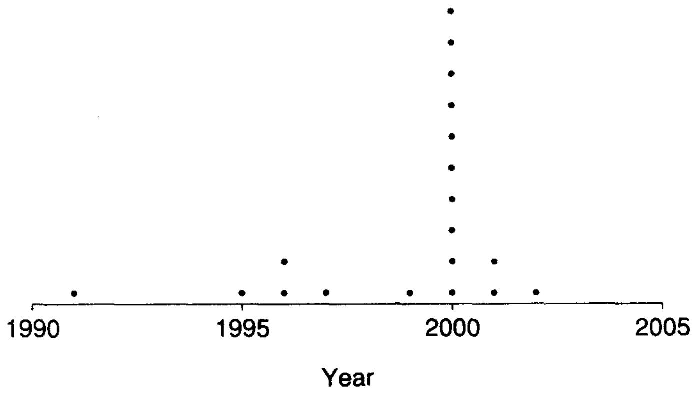


Figure 1.0.1. A dotplot with a spike.

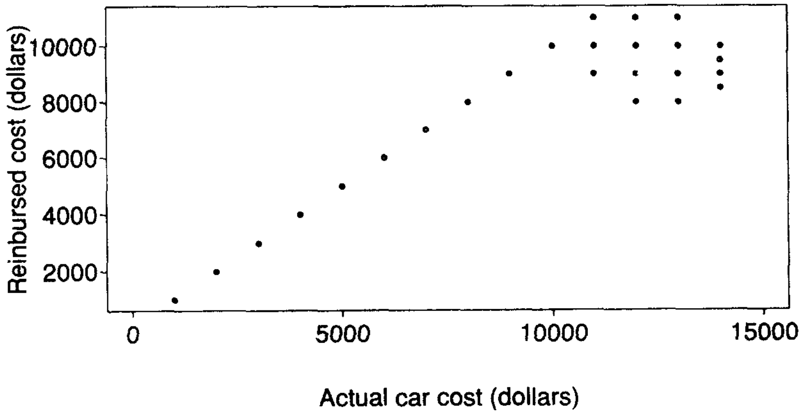


Figure 1.0.2. A bivariate scatterplot with a blip.

important to consider the dotplots and scatterplots. Two further data plots, with a spike and a blip, respectively, are described in Figures 1.0.1 and 1.0.2.

When viewing the data, we should pay careful attention to any outlying observations (see Figure 1.0.3). Outliers are discussed in greater detail by Barnett and Lewis (1978). For example, an outlier can enhance the apparent correlation between two variables that may not otherwise be obviously correlated. Carefully consider the origins and meaning of each outlier and make a careful intuitive decision as to whether or not to include it in the sample. Don't automatically reject outliers, using a "black box" technique,

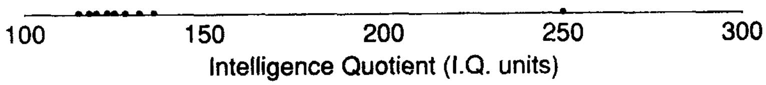


Figure 1.0.3. A dotplot with an outlier.

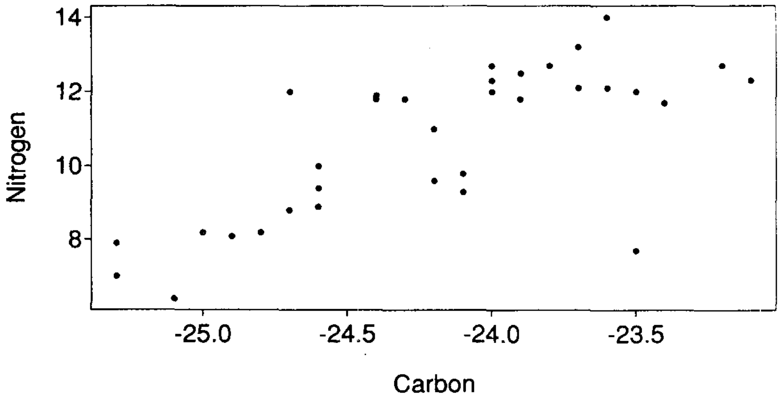


Figure 1.0.4. A scatterplot of the readings on $n = 34$ skeletons.

since they may be quite informative, particularly if they are part of a random sample. Outliers can of course strongly influence any formal analysis, so it is essential to be aware of them. We also regard it as important not to "impute" values for missing data, since the modeling process for the whole data set can then become confused with the imputation process, and the imputed values can exaggerate the information content of the data. Likelihood and Bayesian methods will be able to readily handle missing data problems (just integrate the sampling distribution with respect to the missing observations), without any need to impute the data.

A typical archaeologists' diagram for recording the (transformed) nitrogen and carbon content of skeletons compresses the carbon axis, creating a tendency for the skeletons to be divided into groups, according to nitrogen content alone. In cases where there are two groups, the group of skeletons with the higher nitrogen content is often taken to be Mesolithic, and the group with lower nitrogen content, Neolithic. In Figure 1.0.4, however, we report the entire scatterplot of the readings on $n = 34$ skeletons, at the Lepenski Vir site in the Danube valley, but with the carbon scale substantially broadened when compared with the archaeologists' procedures. Our scatterplot suggests division into several groups rather than two groups. Indeed, McLachlan's package for multivariate mixtures (McLachlan and Basford, 1988) indicates at least six groups. The strong case for at least four or five groups can be confirmed by matching the groups with gender. Further discussions of these data are provided by Bonsall, Lennon, and McSweeney (1997).

For many data sets, it is also of interest either (a) to draw inferences about unknown parameters of interest, for example, the density of a fluid, the recovery rate for patients receiving a particular treatment, or population means, or proportions, or (b) to be able to predict future observations, given the current and previous observations, for example, economic forecasting, the forecasting of the paths of hurricanes, or the prediction of the probability of failure of an engineering design. It is then useful to formalize the random variability or uncertainty in the data, using a mathematical probability model, that is, by taking the numerical observations to be realizations of random variables whose joint distribution comprises the mathematical or "sampling" model. A second "golden rule" is to realize that "true models" are available only in limited situations. In many cases, a variety of different models can be taken to plausibly represent a data set with a finite number n of observations. Which model to use depends partly on statistical technique, but also on the meaning and usefulness of the model in relation to the actual context of the data.

Given a particular sampling model, a key question is, How should the applied statistician use the data to draw inferences about any unknown parameters appearing in the model? In this text, we adhere to the principle, Given the truth of the sampling model, all information in the data is summarized by the likelihood function. The beautiful concept of likelihood links all major philosophies of statistics and provides a cornerstone of the Bayesian paradigm. Its properties and applications are developed in detail throughout this chapter.

It is possible to draw objectively acceptable conclusions from data, when appropriate randomization is performed at the design stage, ideally with further replications of the experiment, to detect unlucky randomizations. For uncontrolled data, an appropriate model can be more difficult to find, and any conclusions are subjective and subject to the effects of "lurking" or "confounding" variables (see Section 1.2 (H)). In general, the conclusions are subject to "shades of subjectivity," depending upon the way the data are collected. For example, Brown et al. (1997) experienced considerable practical difficulties in collecting a random sample while surveying primary-care patients for drug or alcohol abuse. This was mainly the case because the interviewers were under considerable pressure to complete their interviews within time periods agreed upon with the clinics. The conclusions therefore needed to be qualified accordingly. There are also frequently problems with the selective reporting of significant results (see Dawid and Dickey, 1977). Furthermore, the sample size should be chosen with care at the design stage (Donner, 1984).

Both Fisherian and Bayesian statistics depend heavily upon the concept of "probability." What is probability? For a statistical experiment $\mathcal{E}$, with sample space S, mathematicians will tell you that a probability distribution $p(\cdot)$ is a real-valued function defined on all events (strictly speaking, events are constrained to be "measurable subsets") contained in S and satisfying the Kolmogorov axioms (see Exercise 1.1.1). However, philosophically speaking, there are three main types of probabilities:

(A) *Classical probability:* This is defined by an "m over k" rule and is appropriate whenever $S = \{e_1, e_2, \ldots, e_k\}$ possesses k outcomes that are judged to be "equally likely," and when an event A consists of m of these k outcomes. When the equally likely assumption is made objectively, such as when the outcome that occurs has been chosen at random from the k outcomes, or the equally likely assumption has been tested by replicating the experiment numerous times under identical conditions, then the probability $p(A)$ of the event A, defined by $p(A) = m/k$, can be referred to as an "objective classical probability." When the equally likely assumption is made subjectively (e.g., in the absence of evidence to distinguish that any particular outcome is more likely than any other), then $p(A) = m/k$ can be referred to as a "subjective classical probability." See also Exercises 1.1.a and 1.1.b, which tell us that population proportions can be identified with classical probabilities when individuals are chosen at random from the population.

(B) *Frequency probability:* This will be defined by equation (1.1.1). The frequency probability of an event is the long-run proportion of times the event occurs in a large number of replications of the experiment. Objective classical probability provides an example of frequency probability. Therefore, since the objective classical probability that a roulette wheel will give a black number is 9/19, this can also be used to predict the long-run performance of the wheel.

(C) *Subjective probability:* This measures an individual's uncertainty in an event and may vary from individual to individual. You may calibrate your subjective probabilities by judging whether events A are equally likely to events for an objective auxiliary experiment, for example, the spinning pointer of Exercise 1.1.k. In principle, you should assign probabilities to all events $A \subseteq S$ and ensure that your probabilities satisfy the Kolmogorov axioms. An individual who always tries to represent his uncertainty by a subjective probability distribution is referred to as a "Bayesian."

In general, a number between zero and unity can be regarded as a probability only if all other events in the sample space are envisioned, probabilities are assigned to every event, and the laws of probability, as defined by the Kolmogorov axioms, are satisfied by the entire collection of probabilities (referred to as a "probability distribution"). Many "probabilities" quoted in science and the media do not satisfy these conditions. For a sample space with either finitely many outcomes or outcomes that can be arranged in an infinite sequence, it is sufficient to check that the values assigned to the individual outcomes sum to unity.

Consider situations where you possess some information regarding an unknown parameter θ, for values of θ lying in a parametric space Θ. Then a big question is whether or not you can represent this information by a subjective probability distribution on Θ. Some Bayesians say, You should always represent your information by a subjective probability distribution on Θ, since there are some very simple axioms that tell us that if you don't act like this, then you are irrational, incoherent, and moreover, a sure loser! We do not concur with this type of "normative approach," largely because we are unaware of an axiom system that is simple enough, when compared with the Kolmogorov axioms, to justify this viewpoint. Moreover, some information, such as medical knowledge or evidence in a court case, may be too diverse or eclectic to be representable by probabilities. (These views are open to discussion. For a more traditionally Bayesian approach, see Bernardo and Smith, 1994, section 2.3. Many Bayesians believe that the uncertainty in any event is representable by a probability. These aspects are pursued in Exercise 1.1.k and were previously debated by Leonard, 1980, and discussants. In our current chapter, we also debate the likelihood principle. See Section 1.5 (C), Exercises 1.5.c–1.5.f.)

Other topics discussed in the current chapter include Akaike's and Schwarz's information criteria, AIC and BIC, for deciding between different choices of sampling models. These subtract a penalty per parameter from the log-likelihood function (see equations 1.1.5 and 1.1.6). Information criteria are best justified and compared by computer simulation of sets of observations from particular choices of their true sampling model (see Sections 1.2 (F) and (G)). However, it is also important to consider all possible diagnostics, for example, residual analyses for regression models, when comparing models (see Exercise 1.5.l) and also to consider the real-life or scientific reasonability of the candidates.

Any formal statistical procedure, whether for inference about parameters, prediction of future observations, or choice of sampling model, should possess desirable long-run frequency properties (e.g., good mean squared error (MSE) for estimation of parameters, accurate frequency coverage for approximate confidence intervals, high long-run probability of choosing a reasonable model). In situations where these cannot be developed theoretically, computer simulations can produce accurate and meaningful results. Graduate students and research specialists are encouraged to create novel statistical procedures, but then always to check their new ideas by using frequency simulations.

In Section 1.3, procedures are described for obtaining approximate confidence intervals that closely relate to the multivariate normal likelihood approximation (1.3.11) and that refer to the concept of transforming the parameters to achieve possibly better approximate normality. It is particularly important for the research worker to numerically check any theoretical suggestions when using theoretical approximations, since the numerical work may produce some surprises or suggest adjustment terms to the approximations. For example, an "approximately normally distributed random variable X" might not yield values for $p(X < -1.96)$ or $p(X > 1.96)$ that are particularly close to 0.025, as required for an exact result. A variety of practical justifications of the approximations employed in the text are included (e.g., Sections 1.2 (C), 1.4 (A) and (B)). Some key properties of the multivariate normal distribution are developed in Exercise 1.1.n.

A multivariate normal approximation (1.3.11) and related parameter transformations will provide a central theme to a variety of Bayesian ideas developed later in the text, such as the construction of "prior distributions" for several parameters, computational procedures using importance sampling, Laplacian approximations, and rejection sampling. It is more important for the reader to understand the multivariate normal approximation and related approximate confidence intervals than to research the complicated asymptotic theory of maximum likelihood estimators. For any particular model, it is better to check the validity of (1.3.11) computationally and for finite sample sizes.

The works of Sir Ronald Fisher provide excellent background to this chapter. See, for example, Fisher (1925, 1935, 1959) and Bennett (1971–4). Fisher always mixed his techniques with practical common sense.

1.1 Sampling Models and Likelihoods

Numerical data often arise as a result of some statistical experiment $\mathcal{E}$, that is, an occurrence with a random or uncertain outcome. Suppose that on a single repetition of $\mathcal{E}$, you observe n numerical observations $y_1, \ldots, y_n$. Let the sample space S denote the set of all possible realizations of the column vector $\mathbf{y} = (y_1, \ldots, y_n)^T$. Then S is a subset of n-dimensional Euclidean space R^n, and the vector $\mathbf{y}$ consists of the n observations, arranged in a column.

You might be prepared to make the quite strong assumption that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a numerical realization of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ (i.e., the column vector of the random variables $Y_1, \ldots, Y_n$), which possesses some probability distribution P, defined on events in S. For example, if $\mathcal{E}$ can be repeated a large number of times under identical conditions (i.e., *replicated*), then $P = P(\cdot)$ can be defined in terms of the frequency notion of probability. That is, for any event A contained in S,

$$P(A) = \text{prob}(\mathbf{Y} \in A) = \lim_{m \to \infty} r_m(A), \tag{1.1.1}$$

whenever the limit on the right-hand side exists, where the relative frequency $r_m(A)$ denotes the proportion of times that $\mathbf{y} \in A$, during the first m replications of $\mathcal{E}$. However, if $\mathcal{E}$ can be performed only once, then the assumption that $\mathbf{y}$ is a numerical realization of $\mathbf{Y}$ cannot always be made objectively. In some situations where you use random sampling from a population or in other situations where outcomes of the experiment can be regarded as equally likely, it will still be possible to define P in an objective fashion. However, in many cases, part of the modeling process, that is, the specification of P, will need to be performed subjectively and by reference to the scientific or social background of the data. In many cases, P will be incompletely known, even after a variety of modeling assumptions, and it is therefore frequently necessary to infer reasonable choices of P, based upon the vector $\mathbf{y}$ of observations, for a single repetition of $\mathcal{E}$.

For simplicity, assume that $\mathbf{Y}$ is either a continuous random vector with density $p(\mathbf{y})$, for $\mathbf{y} \in S$, or a discrete random vector with probability mass function $p(\mathbf{y})$, for $\mathbf{y} \in S$. Then, following the tenets of *parametric statistical inference*, you might wish to make an assumption of the form

$$p(\mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}) \qquad \left(\mathbf{y} \in S,\ \boldsymbol{\theta} \in \Theta \subseteq R^k\right), \tag{1.1.2}$$

where $f(\mathbf{y} \mid \boldsymbol{\theta})$ is specified as a function of both $\mathbf{y} \in S$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta$. Here $\boldsymbol{\theta}$ is some vector of unknown parameters, and Θ is the parameter space. If $n \geq k$, you can now make inferences about a k-dimensional vector $\boldsymbol{\theta}$ of unknown parameters, rather than an entire function $p(\cdot)$.

Box (1980) distinguishes between this *inference* problem and the problem of statistically *modeling* the choice of functional form for f. Modeling involves both creating an appropriate choice for f in relation to the scientific background and checking the reasonability of this choice against the data. Modeling requires substantial inductive thought, while inference requires deduction, that is, the calculation of mathematical conclusions, given that the functional form of the model is assumed true. This blend of inductive and deductive thought is part of the *inductive synthesis* (Aitken, 1944, p. 3).

Following Birnbaum's (1962) philosophy of "the irrelevance of observations not actually observed" (e.g., why use procedures involving significance probabilities, minimum variance criteria, and confidence statements, which average across the sample space?) and Edwards's famous 1972 treatise on likelihood, it is a reasonable and widely