Bayesian Hierarchical Modeling

David Draper
Department of Applied Mathematics and Statistics
University of California, Santa Cruz
draper@ams.ucsc.edu
http://www.cse.ucsc.edu/~draper


© 2001 David Draper

All rights reserved. No part of this book may be reprinted, reproduced, or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or by an information storage or retrieval system, without permission in writing from the author.

This book was typeset by the author using a PostScript (Adobe Systems, Inc.) based phototypesetter. The figures were generated in PostScript using the S+ data analysis language (Becker et al., 1988), and were directly incorporated into the typeset document. The text was formatted using the \LaTeX{} language (Lamport, 1994), a version of \TeX{} (Knuth, 1984).
TO ANDREA
Contents

Preface ix

1 Introduction to Bayesian modeling 1
   1.1 Quantification of uncertainty about observables 1
   1.2 Discrete outcomes: Exchangeability 7
   1.3 Prior, posterior, and predictive distributions 9
   1.4 Inference and prediction 12
   1.5 Coherence and calibration 13
   1.6 Conjugate analysis 15
   1.7 Comparison with frequentist modeling 18
   1.8 Continuous outcomes 20
   1.9 Additional reading 33
   1.10 Problems 34
   1.11 Notes 38

2 Simulation-based computation 47
   2.1 The need for Markov Chain Monte Carlo methods 47
   2.2 Hastings and Metropolis sampling 53
   2.3 Practical implementation issues 55
   2.4 MCMC monitoring and convergence diagnostics 73
   2.5 Gibbs sampling 83
   2.6 Case study: Measurement of physical constants 93
   2.7 Additional reading 109
   2.8 Problems 110
   2.9 Notes 111

3 Hierarchical models for combining information 123
   3.1 Meta-analysis 123
   3.2 Case study: Can aspirin prevent heart attack mortality? 123
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Approximate fitting of Gaussian HMs: Maximum likelihood and empirical Bayes</td>
<td>123</td>
</tr>
<tr>
<td>3.4</td>
<td>Incorporating study-level covariates</td>
<td>123</td>
</tr>
<tr>
<td>3.5</td>
<td>Case study: Effects of teacher expectancy on pupil IQ</td>
<td>123</td>
</tr>
<tr>
<td>3.6</td>
<td>Additional reading</td>
<td>123</td>
</tr>
<tr>
<td>3.7</td>
<td>Problems</td>
<td>123</td>
</tr>
<tr>
<td>3.8</td>
<td>Notes</td>
<td>123</td>
</tr>
<tr>
<td>4</td>
<td>Hierarchical model diagnostics</td>
<td>125</td>
</tr>
<tr>
<td>4.1</td>
<td>Frequentist-inspired diagnostics</td>
<td>125</td>
</tr>
<tr>
<td>4.2</td>
<td>Predictive validation</td>
<td>125</td>
</tr>
<tr>
<td>4.3</td>
<td>Case study: Dose-response relationships in carcinogenicity assessment of exposure to diesel fumes</td>
<td>125</td>
</tr>
<tr>
<td>4.4</td>
<td>Additional reading</td>
<td>125</td>
</tr>
<tr>
<td>4.5</td>
<td>Problems</td>
<td>125</td>
</tr>
<tr>
<td>4.6</td>
<td>Notes</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>Random-effects and mixed models</td>
<td>127</td>
</tr>
<tr>
<td>5.1</td>
<td>Model-based analysis of cluster samples</td>
<td>127</td>
</tr>
<tr>
<td>5.2</td>
<td>Predictor variables at all levels of the hierarchy</td>
<td>127</td>
</tr>
<tr>
<td>5.3</td>
<td>Comparison between Bayesian and frequentist methods for random-effects and mixed models</td>
<td>127</td>
</tr>
<tr>
<td>5.4</td>
<td>Case study: Quality of care measurement for elderly hospitalized Americans</td>
<td>127</td>
</tr>
<tr>
<td>5.5</td>
<td>Additional reading</td>
<td>127</td>
</tr>
<tr>
<td>5.6</td>
<td>Problems</td>
<td>127</td>
</tr>
<tr>
<td>5.7</td>
<td>Notes</td>
<td>127</td>
</tr>
<tr>
<td>6</td>
<td>Longitudinal data analysis</td>
<td>129</td>
</tr>
<tr>
<td>6.1</td>
<td>Repeated-measures designs</td>
<td>129</td>
</tr>
<tr>
<td>6.2</td>
<td>Growth-curve analysis</td>
<td>129</td>
</tr>
<tr>
<td>6.3</td>
<td>Case study: Effects of maternal speech patterns on infant speech development</td>
<td>129</td>
</tr>
<tr>
<td>6.4</td>
<td>Additional reading</td>
<td>129</td>
</tr>
<tr>
<td>6.5</td>
<td>Problems</td>
<td>129</td>
</tr>
<tr>
<td>6.6</td>
<td>Notes</td>
<td>129</td>
</tr>
<tr>
<td>7</td>
<td>Mixture modeling</td>
<td>131</td>
</tr>
<tr>
<td>7.1</td>
<td>Density estimation</td>
<td>131</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.2</td>
<td>Nonparametric modeling with mixtures of Dirichlet process priors</td>
<td>131</td>
</tr>
<tr>
<td>7.3</td>
<td>Additional reading</td>
<td>131</td>
</tr>
<tr>
<td>7.4</td>
<td>Problems</td>
<td>131</td>
</tr>
<tr>
<td>7.5</td>
<td>Notes</td>
<td>131</td>
</tr>
<tr>
<td>8</td>
<td>Hierarchical modeling as an approach to model selection</td>
<td>133</td>
</tr>
<tr>
<td>8.1</td>
<td>Model expansion</td>
<td>133</td>
</tr>
<tr>
<td>8.2</td>
<td>Bayes factors and Laplace approximations</td>
<td>133</td>
</tr>
<tr>
<td>8.3</td>
<td>The effects of model uncertainty</td>
<td>133</td>
</tr>
<tr>
<td>8.4</td>
<td>Case study: Effects of an intervention to reduce hospitalization rates for elderly people</td>
<td>133</td>
</tr>
<tr>
<td>8.5</td>
<td>Case study: Risk assessment in the Challenger space shuttle disaster</td>
<td>133</td>
</tr>
<tr>
<td>8.6</td>
<td>Additional reading</td>
<td>133</td>
</tr>
<tr>
<td>8.7</td>
<td>Problems</td>
<td>133</td>
</tr>
<tr>
<td>8.8</td>
<td>Notes</td>
<td>133</td>
</tr>
<tr>
<td>9</td>
<td>Discussion and further topics</td>
<td>135</td>
</tr>
<tr>
<td>9.1</td>
<td>Warnings on the unwary use of HMs. Bayes $\neq$ free lunch</td>
<td>135</td>
</tr>
<tr>
<td>9.2</td>
<td>Directions for future research</td>
<td>135</td>
</tr>
<tr>
<td>9.3</td>
<td>Additional reading</td>
<td>135</td>
</tr>
<tr>
<td>9.4</td>
<td>Notes</td>
<td>135</td>
</tr>
<tr>
<td>Appendix 1: Some common prior and likelihood families</td>
<td>137</td>
<td></td>
</tr>
<tr>
<td>Appendix 2: Software details</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A Hastings sampler in S+ for Section 2.2</td>
<td>139</td>
</tr>
<tr>
<td>2</td>
<td>An S+ function to prepare MCMC output for CODA</td>
<td>142</td>
</tr>
<tr>
<td>3</td>
<td>A Hastings sampler in C for Section 2.2</td>
<td>143</td>
</tr>
<tr>
<td>4</td>
<td>A Metropolis sampler in S+ for Section 2.2</td>
<td>150</td>
</tr>
<tr>
<td>5</td>
<td>A Gibbs sampler in S+ for Section 2.5</td>
<td>152</td>
</tr>
<tr>
<td>6</td>
<td>Computing covariance matrices in Maple</td>
<td>155</td>
</tr>
<tr>
<td>7</td>
<td>A generic Metropolis sampler in S+</td>
<td>156</td>
</tr>
<tr>
<td>8</td>
<td>BUGS files for the $t$ example of Section 2.6</td>
<td>159</td>
</tr>
<tr>
<td>9</td>
<td>Metropolis and Gibbs sampling via MLwiN</td>
<td>160</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>161</td>
</tr>
</tbody>
</table>
Index

169
Preface

This book provides an introduction to the formulation, fitting, and checking of [hierarchical] or [multi-level] models, from the Bayesian point of view. Hierarchical models (HMs) arise frequently in five main kinds of applications:

- HMs are common in fields such as health and education, in which data—both outcomes and predictors—are often gathered in a nested or hierarchical fashion: for example, patients within hospitals, or students within classrooms within schools. HMs are thus also ideally suited to the wide range of applications in government and business in which single- or multi-stage cluster samples are routinely drawn, and offer a unified approach to the analysis of random-effects (variance-components) and mixed models.

- A different kind of nested data arises in meta-analysis in, e.g., medicine and the social sciences. In this setting the goal is combining information from a number of studies of essentially the same phenomenon, to produce more accurate inferences and predictions than those available from any single study. Here the data structure is subjects within studies, and as in the clustered case above there will generally be predictors available at both the subject and study levels.

- When individuals—in medicine, for instance—are sampled cross-sectionally but then studied longitudinally, with outcomes observed at multiple time points for each person, a hierarchical data structure of the type studied in repeated-measures or growth curve analyses arises, with the readings at different time points nested within person.

- For simplicity people often try to model data as (conditionally) IID at a fairly high level of aggregation—for instance, by pretending that all the subjects in a sampling experiment are drawn homogeneously from a single population. In fact, heterogene-
ity is often the rule rather than the exception, and frequently the available predictor variables do not "explain" this heterogeneity sufficiently. With recent computational advances it is becoming increasingly straightforward to at least describe such heterogeneity with mixture models that employ latent variables (unobserved predictors) in a hierarchical structure. Examples include density estimation with an unknown number of sub-populations mixed together and Bayesian nonparametric modeling, in which people work with distributions whose sample spaces are themselves sets of distributions instead of (say) real numbers.

- Hierarchical modeling also provides a natural way to treat issues of model selection and model uncertainty with all types of data, not just cluster samples or repeated measures outcomes. For example, in regression, if the data appear to exhibit residual variation that changes with the predictors, you can expand the model that assumes constant variation, by embedding it hierarchically in a family of models that span a variety of assumptions about residual variation. In this way, instead of having to choose one of these models and risk making the wrong choice, you can work with several models at once, weighting them in proportion to their plausibility given the data.

In studying HMs there are two kinds of technical issues that also arise: fully Bayesian computation in HMs requires the use of simulation methods such as those based on Markov Chain Monte Carlo (MCMC) ideas, and—as usual with any class of statistical models—there are questions of model diagnostics.

[Plan of the book.] In the chapters below I describe the principles of Bayesian hierarchical modeling, with emphasis on practical rather than theoretical issues, and I illustrate these principles with analyses of real data drawn from case studies. The material is intended for applied statisticians with an interest in learning more about hierarchical models in general, and the Bayesian analysis of such models in particular. The field of study examined here is surprisingly wide, touching on topics in numerical analysis, high-dimensional integration, and measures on function space (on the mathematical side), the meaning of uncertainty and probability (in philosophy and statistics), and practical issues in Markov chains, time series, and modern nonparametric analysis.
The nine chapters cover the five application areas mentioned above, together with an introductory chapter on Bayesian modeling, one chapter each on MCMC and model diagnostics, and a concluding chapter with discussion and suggestions for future research. An appendix reviews standard probability distributions useful in Bayesian work, and another provides computing details in the environments I used to write the book: the statistical computing and graphics package S, the Gibbs sampling package BUGS, the multi-level modeling package MLwiN, the symbolic computing package Maple, and the high-level programming language C.

An understanding of probability at the level typically required for a master's degree in statistics provides ample mathematical background. I have taught subsets of this material successfully to groups including British final-year undergraduates, American PhD students, and PhD-level researchers enrolled in short courses, and the book has also proven useful for self-study by researchers and graduate students in a variety of disciplines (including statistics).

No previous experience with Bayesian methods is needed—all relevant ideas are covered in a self-contained fashion. If you already know a fair bit about Bayes you can move through Chapter 1 briskly, although there are philosophical and practical issues of potential interest even to seasoned Bayesians there. If you are new to Bayes, a good way to read this book is in conjunction with one or both of the following excellent publications: the Bayesian text by Gelman et al. (1995), and the monograph on MCMC by Gilks et al. (1996) (although the latter is at a more advanced level than the former). A supplementary and complementary perspective on many of the issues covered here can also be obtained by doing some reading in parallel in the excellent book by Carlin and Louis (1996).

Some style and layout conventions to be aware of in the chapters that follow:

- I like to teach and talk about research ideas informally, and the book reflects this. I have tried to write as if you and I were having an extended conversation on the topics covered here. This is natural in a book on applications of the Bayesian approach to probability, and has various advantages, but one possible disadvantage is that the scope of agreement in the statistics community with statements I make may not be immediately clear. So here is a dictionary: sentences including phrases like "You
can show that” and “Evidently” are meant to be expressions of mathematical fact; phrases like “Most people believe that” signal general unanimity (in my view) among (Bayesian) statisticians on the point I’m covering; and phrases like “It seems to me that” precede a personal opinion of mine, which may or may not be shared by other statisticians.

- I am writing in \LaTeX{}, and I don’t like \LaTeX{}’s subsection layout, so [one-line text boxes] act as subsection headings. Multi-line text boxes, in contrast, bring emphasis to definitions, theorems, and summaries of important points.

- The book is dotted with blocks of text that begin **NB**—these highlight things like general notational conventions and pitfalls to be avoided in implementing the ideas I’m discussing at that point.

- **Bold font** is generally reserved for the first appearance of important technical terms, and *italics* signal items of particular emphasis.

- I have tried to write for a fairly diverse audience in terms of mathematical and statistical background. One of the main devices for (I hope) achieving this fairly smoothly is *footnotes*\footnote{1}, which are often too long to be at the bottom of the page where they belong, so I have collected them at the end of each chapter. The naming convention is that, for instance, note\footnote{2} in Chapter 3 will be found as item 3.6 in the Notes section of that chapter. In general, the footnotes supplement the main text by adding historical details, additional mathematical formalism, notices of nonstandard terminology, and the like. The intent is that if you are new to much of this material, you can skip (many or all of) the notes on first reading if you want; whereas if you are fairly experienced in the topics covered here, or you want to dig a bit deeper, you may find that the notes enrich the material and suggest directions for further reading.

- I also offer a somewhat eclectic variety of problems in each chapter: some are data-analytic, others somewhat more theoretical, and they vary widely in difficulty. Problems that use material in the notes begin with the symbol (\textit{Nn}), where \textit{n} refers to the chapter in which the relevant notes may be found. To get the most out of the material, I recommend not only working many or all of the problems but also programming up most or all of
the examples and case studies to see if you get results similar to mine.

I am grateful to Bill Browne, Ryan Cheal, Dimitris Fouskakis, David Freedman, Andrew Gelman, Sander Greenland, Merilee Hurn, Dennis Lindley, Nick Longford, David Madigan, Colin Mallows, Michael Seltzer, and David Williams for comments on earlier versions of this material, and to the UK Engineering and Physical Sciences Research Council, the European Commission, the University of Bath (UK), and the University of California, Santa Cruz for support. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people or institutions responsible for any errors that may be present.

Santa Cruz, California

April 2001

David Draper
1.1 Quantification of uncertainty about observables

Case study 1.1: Hospital-specific prediction of mortality rates.
Let’s say you are interested in measuring the quality of care (e.g.,
Kahn et al., 1990) offered by one particular hospital. I am thinking
of the Royal United Hospital (RUH) in Bath, England, where I
work; you will probably have a different hospital in mind.

As part of this you decide to examine the medical records of
all patients treated at the RUH in one particular time window, say
January 1996–December 1999, for one particular medical condition
for which there is a strong process-outcome link\(^1\), say acute my-
cardial infarction (AMI; heart attack). In the time window you’re
interested in there will be about \( n = 400 \) AMI patients at the RUH.

To keep things simple let’s ignore process for the moment and
focus here on one particular outcome: death status (mortality) as
of 30 days from hospital admission, coded 1 for dead and 0 for
alive. (In addition to process this will also depend on the sickness
at admission of the AMI patients, but let’s ignore that initially
too.) From the vantage point of December 1995, say, what may be
said about the roughly 400 1’s and 0’s you will observe in 1996–99?

The meaning of probability. You are definitely uncertain
about the 0–1 death outcomes \( Y_1, \ldots, Y_n \) before you observe any
of them. Probability is supposed to be the part of mathematics
concerned with quantifying uncertainty\(^2\); how can probability be
used here?

Consider a description \( A \) of some aspect of something about
which you are uncertain. (Here, for example, \( A \) could be \( Y_i = 1 \) =
{patient \( i \) will die} for some \( i \).) Three main approaches to endowing
probabilities with real-world meaning have so far been developed
(e.g., Oakes, 1986\(^3\); Hacking, 1975): classical, frequentist and
Bayesian.
• **Classical:** Enumerate *elemental outcomes* (EOs) in a way that makes them *equipossible* on the basis of symmetry considerations, and compute

$$P_C(A) \equiv \frac{n_A}{n} = \frac{\text{number of EOs favorable to } A}{\text{total number of EOs}}. \quad (1.1)$$

• **Frequentist:** Restrict attention to *attributes* $A$ of *events* (phenomena that are inherently repeatable under "identical" conditions) and define

$$P_F(A) \equiv \lim_{n \to \infty} \frac{\# \text{ of repetitions in which } A \text{ occurs}}{n}. \quad (1.2)$$

• **Bayesian:** Imagine betting with someone about the truth of a *proposition* $A$ (propositions can be anything—not just repeatable phenomena—whose truth value is not (yet) known), and ask yourself what odds $(O_A|B_{\text{you}})$ you would need to give or receive in order that you judge the bet fair, where $B_{\text{you}}$ represents your knowledge and beliefs relevant to the assessment of the odds; then (for you)

$$P_{B;\text{you}}(A) \equiv P_B(A|B_{\text{you}}) \equiv \frac{(O_A|B_{\text{you}})}{1 + (O_A|B_{\text{you}})}. \quad (1.3)$$

**NB** Some notational conventions: (1) In what follows I will usually just write $B$ instead of $B_{\text{you}}$; (2) When it is clear from context that I am talking about a Bayesian probability, I will generally drop the $B$ in $P_B$; and (3) For brevity I will sometimes omit the explicit conditioning on your beliefs $B$ in the notation. This should always be regarded as present, even when not actually printed in the conditional probability expressions.

Each of these probability definitions has general advantages and disadvantages:

• **Classical**
  
  — **Plus:** When relevant, this definition is simple—most people are first taught classical probability, with toy examples like idealized coin-tossing and drawing balls from urns.
  
  — **Minus:** The only way to define "equipossible" without a circular appeal to probability is through the *principle of insufficient reason*—you judge EOs equipossible if you have no
grounds (empirical, logical, or symmetrical) for favoring one
over another—but this leads to paradoxes (for instance, the
assertion of equal uncertainty is not invariant to the choice
of scale on which it is asserted⁴).

- **Frequentist**
  - *Plus*: Mathematical analysis with this approach is relatively
    tractable, which helps to explain the widespread use of fre-
    quentist probability in mathematical statistics over the last
    100 years.
  - *Minus*: But the frequentist definition only applies to inher-
    ently repeatable events: for example, \( P_F(\text{Al Gore will be}
    \text{elected president of the United States in 2000}) \) is (strictly
    speaking) undefined.

- **Bayesian**
  - *Plus*: All forms of uncertainty are inherently quantifiable with
    this approach.
  - *Minus*: There is no guarantee that the answer you get by
    querying yourself about betting odds will retrospectively be
    seen by you or others as “good” (but how should the quality
    of an uncertainty assessment itself be assessed?).

**Application to mortality prediction.** Suppose for the mo-
moment that you did in fact have a variety of process and admission
sickness variables available for a large collection \( \mathcal{P} \) of AMI patients,
and you were trying to assess the probability that a particular
patient—let’s call her \( S \)—with a given process and admission sick-
ness profile will die within 30 days of admission. How would the
three definitions above be applied to this assessment?

If you think about how you would try to quantify this patient’s
risk of dying, you will see that all three approaches require you
to make judgments about the similarity of this patient to other
patients. The English statistician and geneticist Fisher defined the
**recognizable subpopulation** \( \mathcal{P}_S \) to which this patient belongs
as his way of coming to grips with similarity judgments:

**Definition** (Fisher, 1956): The recognizable subpopula-
tion \( \mathcal{P}_S \) for patient \( S \) is the smallest subset to which she
belongs for which the AMI mortality rate differs from
that in the rest of \( \mathcal{P} \) by an amount you judge as signifi-
cant in a practical sense.
Within $\mathcal{P}_S$ you regard the risk of dying as close enough to constant that the differences aren’t worth bothering over, but the differences between mortality rates in $\mathcal{P}_S$ and its complement matter to you. I will address below how you would go about identifying $\mathcal{P}_S$ in practice.

Taking it as given that $\mathcal{P}_S$ has been established, as a classicist you would then (a) use Fisher’s definition to establish equipossibility within $\mathcal{P}_S$, (b) count $n_A = \text{(number of deaths in $\mathcal{P}_S$)}$ and $n = \text{(total number of people in $\mathcal{P}_S$)}$, and (c) compute $P_C(A) = \frac{n_A}{n}$.

As a frequentist, to bring in the idea of repeating something under “identical” conditions, you would have to (a) equate $P(A)$ to $P(\text{a person chosen at random (IID) from $\mathcal{P}_S$ dies})$, (b) imagine repeating this random sampling indefinitely, and (c) conclude that the limiting value of the relative frequency of mortality in these repetitions would be $P_F(A) = \frac{n_A}{n}$. Notice that strictly speaking you can’t talk about $P_F(\text{this patient will die})$—you have to imagine embedding this patient in a repeatable sequence and settle for saying something about the sequence.

As a Bayesian, with the information given here you would regard this patient as exchangeable with all other patients in $\mathcal{P}_S$—meaning informally that you judge yourself equally uncertain about mortality for all the patients in this set—and this judgment, together with the axioms of coherence (a kind of internal consistency requirement; see Note 1.16), would also yield $P_{B,you}(A) = \frac{n_A}{n}$ (although I have not yet said why this is so). I will look at exchangeability and coherence in more detail below.

Note that with the same information base the three approaches in this case have led to the same answer, although the meaning of that answer depends on the approach. For example, frequentist probability describes the process of observing a repeatable event whereas Bayesian probability is an attempt to quantify your uncertainty about something, repeatable or not.

**Subjectivity and “objectivity.”** The classical and frequentist approaches have sometimes been called “objective,” whereas the Bayesian approach is clearly subjective or judgmental. I would argue, however, that in interesting applied problems of realistic complexity, the judgment of similarity (equipossibility, IID, exchangeability) that is evidently central to all three theories makes them all subjective in practice.

Imagine, for instance, that you were given data on death status
in a large group of AMI patients, along with many variables that
might or might not be relevant to predicting their mortality, and
asked to identify $P_s$. You might build a generalized linear model
to estimate $P$(death within 30 days) from the available predictors.
But in building this model you would make many judgment calls,
for example the choice of link function (logit versus complementary
log-log, say) and the “best” subset of predictors to include. The
result could easily be considerable variation in the estimates of
$P$(death) obtained by you and other reasonable analysts working
independently, and the differences between the answers obtained
in this way come entirely from the exercise of modeling judgment.

Thus the assessment of complicated probabilities is inherently
subjective. With this in mind attention in all three approaches
should perhaps shift away from trying to achieve “objectivity” to-
ward the explicit statement of the assumptions and judgments
made in forming probability assessments, so that consumers of
these assessments may judge their plausibility.

Frequentist modeling. I will focus on the approaches with
the most widespread usage—frequentist and Bayesian—in the rest
of the book. How, for instance, can the frequentist definition of
probability be applied to the hospital mortality problem?

As a frequentist, to use probability to quantify your uncertainty
about the 1’s and 0’s, you have to think of them as either literally
a random sample or like a random sample from some population,
either hypothetical or actual.

- An example of a hypothetical population would be all AMI pa-
tients who might have come to the RUH in 1996–99 if the world
had turned out differently in some (unspecified) ways.

- Some actual populations: (1) Assuming sufficient time-homogen-
ey in all relevant factors, you could try to argue that the col-
lection of all 400 AMI patients at the RUH from 1996–99 is like
a random sample of size 400 from the population of all AMI pa-
tients at the RUH from (say) 1993–2002, even though in fact it is
a kind of time-cluster sample in which you got everybody from
1996–99 and nobody from 1993–95 or 2000–02; or (2) Assuming
the RUH to be representative of some broader collection of
hospitals in England and ignoring intracluster correlation, you
could try to argue that a cluster sample of all 400 AMI patients
from the RUH was like a simple random sample of 400 AMI
patients from this larger collection of hospitals.
None of these options is, shall we say, entirely compelling⁷.

If you are willing to pretend the data are like a sample from some population, you could then regard the 400 1's and 0's at the RUH as realizations of random variables and begin to think about a model, for example

\[ Y_i \sim \text{B}(\theta_i), \quad i = 1, \ldots, n, \quad (1.4) \]

where \( \text{B}(\cdot) \) denotes the Bernoulli distribution. (Appendix 1 contains a summary of the distributions used in this book.) In the absence of any sickness or process information, however, you would probably have to treat the 1's and 0's as homogeneous and work with the simpler model

\[ Y_i \sim \text{B}(\theta), \quad i = 1, \ldots, n. \quad (1.5) \]

Interest would then focus on inference about the parameter \( \theta \), the “underlying death rate”: if \( \theta \) were unusually high, that would be \textit{prima facie} evidence of a possible quality of care problem at the RUH⁸.

**Bayesian modeling.** As a Bayesian in this situation, your job is to quantify your uncertainty about the 400 binary observables you will begin to see starting in 1996—in other words, your initial modeling task is \textit{predictive} rather than inferential. There is no samples-and-populations story in this approach, but probability and random variables arise in a different way: quantifying your uncertainty (for the purpose of betting with someone about some aspect of the 1's and 0's, say) requires \textit{eliciting} from yourself a joint probability distribution that \textit{accurately} captures your judgments about what you will see⁹:

\[ P_{\text{B:you}}(Y_1 = y_1, \ldots, Y_n = y_n). \quad (1.6) \]

Notice as before that in the frequentist approach the random variables describe the \textit{process} of observing a repeatable event (the “random sampling” appealed to here), whereas in the Bayesian approach you use random variables to quantify \textit{your uncertainty about observables you haven’t seen yet}¹⁰.

I will argue later (Section 1.5) that the concept of probabilistic \textit{accuracy} has two components: you want your uncertainty assessments to be both \textit{internally} and \textit{externally} consistent, which corresponds to the ideas of \textit{coherence} and \textit{calibration}, respectively.
1.2 Discrete outcomes: Exchangeability

Eliciting a 400-dimensional distribution doesn't sound easy—major simplification is evidently needed. In this case, and many others, this is provided by exchangeability considerations. If (as in the frequentist approach) you have no relevant information that distinguishes one AMI patient from another, your uncertainty about the 400 1's and 0's is symmetric, in the sense that a random permutation of the order in which the 1's and 0's were labeled from 1 to 400 would leave your uncertainty about them unchanged. The Italian statistician de Finetti (1930, 1937/1980) called random variables with this property exchangeable:

Definition (de Finetti, 1930): \( \{Y_i, i = 1, \ldots, n\} \) are exchangeable if the distributions of \( (Y_1, \ldots, Y_n) \) and \( (Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \) are the same for all permutations \( (\pi(1), \ldots, \pi(n)) \).

NB Exchangeability and IID are not the same: exchangeable \( Y_i \) do have identical marginal distributions but are not independent. For example, if you were expecting a priori about 15% 1's, say (that's the 30-day death rate for AMI in England with average-quality care), the knowledge that in the first 50 outcomes 20 of them were deaths would certainly change your prediction of the 51st. In other words, for you \( P_{\theta}(Y_{51} | \sum_{i=1}^{50} Y_i = 20) \neq P_{\theta}(Y_{51}) \) (we will see a bit later that the \( Y_i \) only become independent conditional on the same \( \theta \) that arose in the frequentist approach; Problem 1.1).

de Finetti also defined partial or conditional exchangeability (e.g., Draper et al., 1993): if, for instance, the gender \( X \) of the AMI patients is available, and there is evidence from the medical literature that 1's tended to be noticeably more likely for men than women, then you would probably want to assume conditional exchangeability of the \( Y_i \) given \( X_i \), meaning that the male and female 1's and 0's, viewed as separate collections of random variables, are each unconditionally exchangeable.

[de Finetti's representation theorem for 1's and 0's.]

The judgment of exchangeability still seems to leave the joint distribution of the \( Y_i \) quite imprecisely specified. After defining the concept of exchangeability, however, de Finetti went on to prove a remarkable result: if you are willing to regard the \( \{Y_i, i = 1, \ldots, n\} \) as part of an infinite exchangeable sequence of 1's and 0's (meaning that every finite subsequence is exchangeable), then you can
express your joint distribution in a particularly simple way (e.g., de Finetti, 1930; Bernardo and Smith, 1994):

\textbf{Theorem} (de Finetti, 1930): If \( Y_1, Y_2, \ldots \) is an infinitely exchangeable sequence of 0–1 random quantities with probability measure \( P \), there exists a distribution function \( Q(\theta) \) such that the joint distribution \( p(y_1, \ldots, y_n) \) for \( Y_1, \ldots, Y_n \) is of the form

\[
p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} dQ(\theta),
\]

where

\[
Q(\theta) = \lim_{n \to \infty} P\left(\frac{1}{n} \sum_{i=1}^n Y_i \leq \theta\right)
\]

and \( \theta \overset{P}{=} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n Y_i \). \hfill (1.7)

Leaving aside for a moment the interpretation of \( \theta \), the distribution function \( Q \) will generally be well-behaved enough to have a density: \( dQ(\theta) = p(\theta) d\theta \). In this case de Finetti’s Theorem says

\[
p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} p(\theta) d\theta. \hfill (1.8)
\]

Now by the law of total probability and the definition of conditional probability,

\[
p(y_1, \ldots, y_n) = \int_0^1 p(y_1, \ldots, y_n, \theta) d\theta
\]

\[
= \int_0^1 p(y_1, \ldots, y_n|\theta) p(\theta) d\theta, \hfill (1.9)
\]

and (1.8) and (1.9) together imply that

\[
p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}. \hfill (1.10)
\]

But the right side of (1.10) is just the sampling distribution of \( n \) Bernoulli random variables with common success probability \( \theta \).

Thus, according to de Finetti’s Theorem, under exchangeability it is as if (a) there were a random quantity called \( \theta \), interpretable as the limiting relative frequency of 1’s, (b) conditional on this \( \theta \) the \( Y_i \) are IID \( B(\theta) \), and (c) \( \theta \) itself has a distribution with density \( p(\theta) \).
In other words, a Bayesian whose uncertainty about dichotomous $Y_i$ is exchangeable may as well use the simple model

$$\theta \sim p(\theta)$$

$$(Y_i|\theta) \overset{\text{IID}}{\sim} B(\theta), \ i = 1, \ldots, n.$$  \hspace{1cm} (1.11)

This is an example of the simplest kind of hierarchical model \((\text{HM})^{13}\): a model at the top level for the underlying death rate $\theta$, and then a model below that for the 0–1 mortality indicators $Y_i$ conditional on $\theta$.

1.3 Prior, posterior, and predictive distributions

Notice that to make sense of de Finetti’s Theorem you have to treat $\theta$ as a random variable, even though logically it is a fixed unknown constant. This is the main conceptual difference between the Bayesian and frequentist approaches: as a frequentist the random variables are supposed to capture relevant features of the process of sampling from a population, whereas in the Bayesian approach you use the machinery of random variables to express your uncertainty about unknown quantities.

Q1: What is the real-world meaning of $p(\theta)$ in (1.11)?

\(A_1\): $p(\theta)$ does not involve $Y = (Y_1, \ldots, Y_n)$, and probability is all about uncertainty quantification for Bayesians, so $p(\theta)$ must represent your uncertainty about $\theta$ before the data set $Y$ arrives, which is why everybody calls it your \text{prior distribution}^{14} for $\theta$. I will address how you might go about specifying this distribution below.

NB You don’t need to literally think of $\theta$ as having been sampled from $p(\theta)$; the assumption $\theta \sim p(\theta)$ is just a way of quantifying what (if anything) was known about $\theta$ before $Y$ is observed.

Q2: If $p(\theta)$ represents your uncertainty about $\theta$ before the data arrive, what represents this uncertainty after $Y$ has been observed?

\(A_2\): It has to be $p(\theta|Y)$, the conditional distribution for $\theta$ given how $Y$ came out. It is natural to call this the \text{posterior distribution} for $\theta$ given $Y$.

Q3: How do you get from $p(\theta)$ to $p(\theta|Y)$—in other words, how do you update your uncertainty about the unknown $\theta$ in light of the data?

\(A_3\): Use the definition of conditional probability on $p(\theta|Y)$,

$$p(\theta|Y) = \frac{p(\theta, Y)}{p(Y)},$$  \hspace{1cm} (1.12)
and then use the definition again to force \( p(\theta) \) to appear on the right-hand side:

\[
\frac{p(\theta, Y)}{p(Y)} = \frac{p(\theta) p(Y|\theta)}{p(Y)}. \tag{1.13}
\]

The result is

\[
\text{Theorem (Bayes, 1763), for continuous quantities } \theta \text{ (the unknown) and } Y \text{ (the data):}
\]

\[
p(\theta|Y) = \frac{p(\theta) p(Y|\theta)}{p(Y)}. \tag{1.14}
\]

It may seem, from how easy it is to arrive at this result, that the Rev. Bayes\(^5\) didn’t have to work very hard to achieve his immortality, but he actually did quite a bit more: he helped to put conditional probability on a sound footing for the first time, and he encouraged application of the theorem to social and medical problems, by viewing what I have here called \( \theta \) and \( Y \) as examples of causes and effects, respectively—in other words, he suggested how to pass from the easier problem of predicting the likely effects of known causes to the more difficult task of inferring the causes of observed effects.

To put (1.14) into practice, some interpreting is required. As a Bayesian I want to condition on things I know and believe, in using probability to express my uncertainty: remember the \( B \) in equation (1.3), which I have somewhat lazily been notationally suppressing. After the data vector \( Y \) is observed, I know it—it becomes part of my \( B \)—and so I should condition on the data in applying (1.14). Thus, I am thinking of the left side of (1.14) as a function of \( \theta \) for fixed \( Y \), so that must also be true of the right side. In other words, (a) \( p(Y) \) is just a constant—in fact, you can think of it as the normalizing constant, put into the equation to make the right side of (1.14) integrate to 1; and (b) \( p(Y|\theta) \) may look like the usual frequentist sampling distribution for \( Y \) given \( \theta \) (Bernoulli, in this case), but to use (1.14) I have to think of \( p(Y|\theta) \) as a function of \( \theta \) for fixed \( Y \). When thought of this way—you could denote it \( l(\theta|Y) = p(Y|\theta) \)—Fisher (1922) called it the likelihood function.

\textbf{NB} The roles of \( \theta \) and \( Y \) are completely reversed in the Bayesian approach to inference when compared with the frequentist approach: with my frequentist hat on I regard \( \theta \) as a fixed (unknown) constant and \( Y \) as a random variable, and everything focuses on
imagining what would happen as Y changes randomly from sample
to sample; but with my Bayesian hat on I am thinking of Y as a
fixed (known) constant and $\theta$ as a random variable, and everything
comes down to assessing my uncertainty about $\theta$ after conditioning
on the one and only one Y I'm ever going to see.

From (1.14), Bayes' Theorem can evidently be interpreted as
follows:

$$p(\theta|Y) = c \cdot p(\theta) \cdot l(\theta|Y)$$

posterior = \left( \begin{array}{c}
\text{normalizing} \\
\text{constant}
\end{array} \right) \cdot \text{prior} \cdot \text{likelihood}.

You can also readily construct predictive distributions for
the $Y_i$ before they are observed, or for future $Y_i$ once some of them
are known. For example, the posterior predictive distribution for
$(Y_{m+1}, \ldots, Y_n)$ given $(Y_1, \ldots, Y_m)$—that is, $p(y_{m+1}, \ldots, y_n|y_1, \ldots, y_m)$—is, by a trick similar to that in equation (1.9), just

$$\int_{0}^{1} p(y_{m+1}, \ldots, y_n|\theta, y_1, \ldots, y_m) \cdot p(\theta|y_1, \ldots, y_m) \, d\theta$$

$$= \int_{0}^{1} p(y_{m+1}, \ldots, y_n|\theta) \cdot p(\theta|y_1, \ldots, y_m) \, d\theta$$

$$= \int_{0}^{1} \prod_{i=m+1}^{n} \theta^y \cdot (1-\theta)^{1-y} \cdot p(\theta|y_1, \ldots, y_m) \, d\theta .$$

Notice an important simplification here in going from $p(y_{m+1}, \ldots,$
$y_n|\theta, y_1, \ldots, y_m)$ to $p(y_{m+1}, \ldots, y_n|\theta)$: conditional on $\theta$ the $Y_i$ are
independent—in other words, if you know $\theta$ the individual values
of $Y_1, \ldots, Y_m$ will not help you to predict $Y_{m+1}, \ldots, Y_n$. Two nice
things follow from this: (a) $p(y_{m+1}, \ldots, y_n|\theta, y_1, \ldots, y_m)$ reduces to
$p(y_{m+1}, \ldots, y_n|\theta)$, and then (b) since the $Y_i$ are conditionally inde-
pendent given $\theta$, $p(y_{m+1}, \ldots, y_n|\theta)$ reduces to $\prod_{i=m+1}^{n} p(y_i|\theta)$. The
result—for example, the middle line of (1.16)—is intuitively rea-
sonable: you are trying to construct your predictive distribution
for a bunch of new $Y$'s, and it would sure help to know $\theta$ in doing
so, but $\theta$'s value is not certain. So take a weighted average, or mix-
ture, of conditional predictive distributions given $\theta$, weighted by
your best current information about $\theta$, namely the posterior for $\theta$
given the $Y_i$ you have already seen.

This also brings up a key difference between a parameter like $\theta$
on the one hand and the $Y_i$, before you have observed any data,
on the other: parameters are inherently unobservable. This makes it harder to evaluate the quality of your uncertainty assessments about \( \theta \) than to do so about the observable \( Y \). Once you have the posterior for \( \theta \) given \( Y \), \( p(\theta|Y) \), there is no direct way to check its quality as an uncertainty assessment, because \( \theta \) is (and presumably always will remain) unknown, whereas once you have a predictive distribution \( p(y_{m+1}|y_1, \ldots, y_m) \) for an observable like \( Y_{m+1} \), you can directly check its quality by comparing the actual \( Y_{m+1} \) with your predictive distribution for it.

1.4 Inference and prediction

The de Finetti approach to modeling emphasizes the prediction of observables as a valuable adjunct to inference about unobservable parameters, for at least two reasons:

- Key scientific questions are often predictive in nature: for instance, rather than asking "Is drug A better than B (on average) for lowering blood pressure?" (inference), the ultimate question is "How much more will drug A lower this patient's blood pressure than drug B?" (prediction); and

- Good diagnostic checking is predictive: As noted above, an inference about an unobservable parameter can never be directly verified, but often you can reasonably conclude that inferences about the parameters of a model which produces poor predictions of observables are also suspect. This will serve as the basis of the model diagnostics in Chapter 3.

With the predictive approach parameters diminish in importance, especially those that have no physical meaning—from the Bayesian viewpoint (e.g., Lindley, 1972) such parameters (unlike \( \theta \) above) can be regarded as just place-holders for a particular kind of uncertainty on your way to making good predictions. It is arguable (e.g., Draper, 1995a) that the discipline of statistics, and particularly its applications in the social sciences, would be improved by a greater emphasis on predictive feedback. When was the last time you saw a statistical application, outside of (say) weather-forecasting, in which the investigators made testable predictions based on their inferential conclusions and verified them with new data?

This is not to say that parametric thinking should be abolished. As the calculation in equation (1.16) emphasized, parameters play
an important simplifying role in forming modeling judgments: the single strongest simplifier of a joint distribution is independence of its components, and whereas (for instance) in the mortality example the $Y_i$ are not themselves independent, they become so conditional on $\theta$.

1.5 Coherence and calibration

de Finetti’s Theorem for 0–1 outcomes says informally that if you are trying to make coherent\textsuperscript{16} (internally consistent) probability assessments about a series of 1’s and 0’s that you judge exchangeable, you may as well behave like a frequentist—IID $B(\theta)$—with a prior distribution $p(\theta)$. But where does this prior come from? (NB Coherence doesn’t help in answering this question—it turns out that any prior $p(\theta)$ could be part of somebody’s coherent probability judgments.)

Some people regard the need to answer this question in the Bayesian approach as a drawback, but it seems to me to be a positive aspect\textsuperscript{17}, as follows. From Bayes’ Theorem the prior is supposed to be a summary of what you know (and don’t know) about $\theta$ before the $Y_i$ start to arrive: from previous datasets of which you are aware, from the relevant literature, from expert opinion, and so on—from all “good” sources, if any exist. Such information is almost always present, and should presumably be used when available. The issue is how to do so “well.”

The goal is evidently to choose a prior that you will retrospectively be proud of, in the sense that your predictive distributions for the observables (a) are well-centered near the actual values and (b) have uncertainty bands that correspond well to the realized discrepancies between actual and predicted values. This is a form of calibration of your probability judgments.

There is no guaranteed way to do this, just as there is no guaranteed way to arrive at a “good” frequentist model (see “Where does the likelihood come from?” in Section 1.8).

**Choosing a “good” prior.** Some general comments on arriving at a “good” prior:

- There is a growing literature on methodology for elicitation of prior information (e.g., Kadane et al., 1980; Craig et al., 1997;
Kadane and Wolfson, 1997; O'Hagan, 1997), which brings together ideas from statistics and perceptual psychology. To take just one example from this literature, people turn out to be better at estimating percentiles of a distribution than they are at estimating standard deviations, a fact that has direct consequences for how you should ask experts about variability.

- Bayes' Theorem on the log scale says (apart from the normalizing constant) that

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}); \quad (1.17)$$

in other words, (posterior information) = (prior information) + (data information). This means that close attention should be paid to the information content of the prior, for instance by density-normalizing the likelihood and plotting it on the same scale as the prior. It is possible for small $n$ for the prior to swamp the data, and in general you should not let this happen without a good reason for doing so. Comfort can also be taken from the other side of this coin: with large $n$ (in most situations, at least) (1.17) implies that the data swamp the prior, and prior specification errors become less important.

- When you notice you are quite uncertain about how to specify the prior, you can try sensitivity or (pre-posterior) analysis: exploring the mapping from prior to posterior, before the data are gathered, by (a) generating some possible values for the observables, (b) writing down several plausible forms for the prior, and (c) carrying these forward to posterior distributions. If the resulting distributions are similar ("all reasonable roads lead to Rome"), you have uncovered a useful form of stability in your results; if not you can try to capture the prior uncertainty hierarchically, by, for instance, adding another layer to models like (1.11) above (Problem 7.1).

- Calibration can be estimated by a form of cross-validation: with a given prior you can (a) repeatedly divide the data at random into modeling and validation subsets, (b) update to posterior predictive distributions based on the modeling data, and (c) compare these distributions with the actual values in the validation data. Chapter 3 illustrates some examples of this idea, which I will call predictive validation in what follows.

Note that calibration is inherently frequentist in spirit—it is
CONJUGATE ANALYSIS

based on questions like “What percentage of the time do your 90% central predictive intervals include the actual value?”). This leads to a useful synthesis of Bayesian and frequentist thinking:

Coherence keeps you internally honest; calibration keeps you in good contact with the world.

**Bayes + frequentist, not Bayes vs. frequentist.** People often talk about the so-called Bayesian-frequentist controversy as if it is necessary to choose sides, in a confrontation in which one approach must be right and the other wrong. There is a kind of empirical theorem that shows this attitude must be wrong: intelligent people have been arguing about this topic for almost 250 years, at least since the publication of Bayes (1763), and if the two sides were metaphorical boxers it is clear from current statistical theory and practice that both boxers are still standing in the ring after all of the punching. The implication I draw from this is that everyone should seek a personal synthesis of the best features of both the Bayesian and frequentist ways of looking at the world.

I find in my own applied work, for instance, that it is useful to reason in a Bayesian way when formulating my inferences and predictions, and to reason in a frequentist way when evaluating their quality, through calibration-style comparisons between predictive distributions for observables and the actual observables themselves. Others (e.g., Box 1980, Rubin 1984) have offered similar views; for a more skeptical position see Freedman (1995). After you have gained experience with the methods in this book, you may reach different conclusions—if so I would be interested to hear them.

1.6 Conjugate analysis

**Example: Prior specification in the mortality data.** Let’s say (a) you know that the 30-day AMI mortality rate given average care and average sickness at admission in England is about 15% (which is in fact about right), (b) you know little about care or patient sickness at the RUH, but (c) you would be somewhat surprised (on Central Limit Theorem grounds) if the “underlying rate” at the RUH were much less than 5% or more than 30% (note the asymmetry). To quantify these judgments you seek a flexible family of densities on (0,1), one of whose members has mean 0.15 and (say) 95% central interval (0.05,0.30).
A convenient family for this purpose is the \textbf{beta} distributions, $\text{Be}(\theta|\alpha, \beta) = c \theta^{\alpha-1} (1 - \theta)^{\beta-1}$, for two reasons:

- This family exhibits a wide variety of distributional shapes (e.g., Johnson and Kotz, 1970); and

- The likelihood in this problem comes from the Bernoulli/binomial sampling distribution for the $Y_i$, $p(y_1, \ldots, y_n|\theta) = l(\theta|y) = c \theta^S (1 - \theta)^{n-S}$, where $S = \sum_{i=1}^n y_i$. Thus with this choice of prior, the likelihood and prior (and thus the posterior) have the same distributional form, $\theta^S (1 - \theta)^{n-S}$, which makes life computationally much easier. For this reason the collection of beta prior distributions is said to be \textbf{conjugate} to the Bernoulli/binomial likelihood\textsuperscript{18}.

**Conjugate analysis**—finding conjugate priors for standard likelihoods and restricting attention to them on tractability grounds—is one of only two fairly general methods for getting closed-form answers in the Bayesian approach; the other is \textbf{asymptotic analysis} (e.g., Bernardo and Smith, 1994), about which I won’t have much to say here. The idea in the next few sections is to see how far conjugate analysis can take us and then to switch over to a more general approach to computation, \textbf{Markov Chain Monte Carlo (MCMC)}, in Chapter 2.

In the mortality example, trial and error shows $\alpha = 4.5$ and $\beta = 25.5$ produce approximately the desired mean and central interval—this distribution has mode 0.125 and standard deviation (SD) 0.064. $\alpha$ and $\beta$ are called \textbf{hyperparameters} since they are parameters of the prior distribution for the parameter $\theta$ of central interest. With $(\alpha_0, \beta_0) = (4.5, 25.5)$, written hierarchically the model is

$$(\alpha, \beta) = (\alpha_0, \beta_0) \quad \text{(hyperparameters)}$$

$$(\theta|\alpha, \beta) \sim \text{Be}(\alpha, \beta) \quad \text{(prior)} \quad (1.18)$$

$$(Y_1, \ldots, Y_n|\theta) \overset{\text{iid}}{\sim} \text{B}(\theta) \quad \text{(likelihood)}$$

The conjugacy of the prior leads to a simple closed form for the posterior here: with $y$ as the vector of observed $Y_i$, $i = 1, \ldots, n$, and $S$ as the sum of the $y_i$,

$$p(\theta|y, \alpha, \beta) = \frac{c \ p(y|\theta) \ p(\theta|\alpha, \beta)}{c \ \theta^S (1 - \theta)^{n-S} \ \theta^{\alpha-1} (1 - \theta)^{\beta-1}} = \frac{c \ \theta^{S+\alpha-1} (1 - \theta)^{n-S+\beta-1}}{c \ \theta^{S+\alpha-1} (1 - \theta)^{n-S+\beta-1}},$$

\text{where } c \text{ is a normalizing constant.}
in other words, the posterior for \( \theta \) is \( \text{Be}(\alpha + S, \beta + n - S) \). \textbf{NB} This brings up the Bayesian version of \textit{sufficiency}: A quantity \( f(Y) \) is \textit{sufficient} for the parameter \( \theta \)—informally, \( f(Y) \) is the only function of \( Y \) you need (given the model you are working with) in drawing inferences about \( \theta \)—if the likelihood \( l(\theta|Y) \) depends on \( Y \) only through \( f(Y) \). Here \( S \) is evidently the sufficient statistic for \( \theta \) with the Bernoulli/binomial likelihood.

\begin{quote}
\textbf{Prior effective sample size.} This gives the hyperparameters a direct interpretation in terms of \textit{effective information content of the prior}: it is as if the data—represented by the \( \text{Be}(S + 1, n - S + 1) \) likelihood—were worth \( (S + 1) + (n - S + 1) = n \) observations and the prior (\( \text{Be}(\alpha, \beta) \)) were worth \( (\alpha + \beta) \) observations. This can be used to judge whether the prior is "too informative"—here it is equivalent to \( (4.5 + 25.5) = 30 \) binary observables with a mean of 0.15.
\end{quote}

![Figure 1.1. Prior, likelihood, and posterior distributions for the RUH mortality data.](image)

It turns out, with conjugate models such as (1.18), that this idea can be used to make a precise connection between Bayesian and frequentist analyses of the same data set, as follows. To produce the Bayesian analysis here, it is as if you (a) create a \textbf{prior data set} with sample size \( n^* = 30 \), consisting of 4.5 1's and 25.5 0's (so to speak) and mean 0.15, (b) merge this prior data set with the actual data set, and (c) perform a frequentist analysis on the resulting set.
of \((n^* + n)\) values. This gives a kind of literal interpretation to the idea of prior information.

Suppose the \(n = 400\) observed mortality indicators consist of \(S = 72\) 1’s and \((n - S) = 328\) 0’s. Then the prior is \(\text{Be}(4.5,25.5)\), the likelihood is \(\text{Be}(73,329)\), the posterior for \(\theta\) is \(\text{Be}(76.5,353.5)\), and the three densities plotted on the same graph come out as in Figure 1.1. In this case the posterior and the likelihood nearly coincide, because the data information outweighs the prior information by \(400/30\), which is more than 13 to 1.

The mean of a \(\text{Be}(\alpha, \beta)\) distribution is \(\alpha / (\alpha + \beta)\); with this in mind the posterior mean has a clear interpretation as a weighted average of the prior mean and data mean, with weights determined by the effective sample size of the prior, \((\alpha + \beta)\), and the data sample size \(n\):

\[
\frac{\alpha + S}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{S}{n}
\]

posterior mean = prior mean \cdot prior weight + data mean \cdot data weight

0.178 = 0.070 \cdot 0.15 + 0.93 \cdot 0.18.

Another way to put this is that the combining of prior and data information shrinks the data mean, \(\bar{y} = S/n = 72/400 = 0.18\), toward the prior mean 0.15 by (in this case) a modest amount: the posterior mean is about 0.178, and the shrinkage factor is \(30/(30 + 400) \approx 0.07\). This idea of shrinkage estimation will come up again in Chapter 3.

1.7 Comparison with frequentist modeling

To analyze these data as a frequentist you would probably appeal to the Central Limit Theorem: \(n = 400\) is big enough so that the sampling distribution of \(Y\) is approximately \(N(\theta, \sigma^2 (\theta) n)\), so an approximate 95% confidence interval for \(\theta\) would be centered at \(\hat{\theta} = \bar{y} = 0.18\), with an estimated standard error of \(\sqrt{\sigma^2 (\theta)/n} = 0.0192\), and would run roughly from 0.142 to 0.218. By contrast the posterior for \(\theta\) is also approximately Gaussian\(^10\), with a mean of 0.178 and an SD of \(\sqrt{(\alpha^* + \beta^*)/(\alpha^* + \beta^* + 1)} = 0.0184\), where \(\alpha^*\) and \(\beta^*\) are the parameters of the beta posterior distribution; a 95% central posterior interval for \(\theta\) would thus run from about 0.143 to 0.215. The two approaches give almost the same answers in this
case, a result that is typical of situations with fairly large \( n \) and relatively diffuse prior information (meaning that the prior SD is large relative to the normalized likelihood SD, or equivalently that the prior effective sample size is small relative to the data sample size), as in Figure 1.1.

Note, however, that the interpretation of the two analyses differs somewhat:

- In the frequentist approach \( \theta \) is fixed but unknown and \( \bar{Y} \) is random, with the analysis based on imagining what would happen if the hypothetical random sampling were repeated, and appealing to the fact that across these repetitions \( (\bar{Y} - \theta) \sim N(0, 0.0192^2) \); whereas

- In the Bayesian approach \( \bar{Y} \) is fixed at its observed value and \( \theta \) is treated as random, as a means of quantifying your posterior uncertainty about it: \( (\theta - \bar{Y})|\bar{Y} \sim N(0, 0.0184^2) \).

This means among other things that, while it is not legitimate with the frequentist approach to say that \( P_F(0.14 \leq \theta \leq 0.22) \approx 0.95 \), which is what many users of confidence intervals would like them to mean, the corresponding statement \( P_B(0.14 \leq \theta \leq 0.22 | Y, \text{little or no prior info}) \approx 0.95 \) is a natural consequence of the Bayesian approach. In the case of diffuse prior information this justifies the fairly common practice of computing inferential summaries in a frequentist way and then interpreting them Bayesianly.

When nondiffuse prior information is available and you use it, your answer will differ from a frequentist analysis based on the same likelihood. Assuming that after the fact the likelihood is judged to have been based on an accurate reflection of the sampling realities, if your prior is retrospectively seen to have been well-calibrated you will get a better answer than with the frequentist approach; if poorly calibrated, a worse answer (Samaniego and Reneau, 1994). This may be restated schematically as

<table>
<thead>
<tr>
<th>Bayesian with</th>
<th>frequentist with</th>
<th>Bayesian with</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;bad&quot; prior ( \leq ) with &quot;good&quot; prior</td>
<td>&quot;good&quot; prior ( \leq ) with &quot;good&quot; prior</td>
<td>(1.20)</td>
</tr>
<tr>
<td>and &quot;good&quot; likelihood</td>
<td>and &quot;good&quot; likelihood</td>
<td>likelihood</td>
</tr>
</tbody>
</table>

What you make of this depends on your risk-aversion: Is it better...
to try to land on the right in this box, running some risk of landing on the left, or to steer a middle course? (For myself, I try to use predictive calibration (as in Section 1.5) to end up on the right. **NB**

(1) I will give several examples later in which a Bayesian analysis is better even with diffuse prior information. (2) Expression (1.20) says nothing about analysts, Bayesian or frequentist, with “bad” likelihoods.)

### 1.8 Continuous outcomes

For continuous outcomes there is an analogue of de Finetti’s Theorem that is equally central to Bayesian model-building (e.g., Bernardo and Smith, 1994):

**Theorem** (de Finetti, 1937): If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of real-valued random quantities with probability measure $P$, there exists a probability measure $Q$ over $\mathcal{D}$, the space of all distribution functions on the real line $\mathbb{R}$, such that the joint distribution function of $Y_1, \ldots, Y_n$ has the form

$$P(y_1, \ldots, y_n) = \int_{\mathcal{D}} \prod_{i=1}^{n} F(y_i) \ dQ(F), \quad (1.21)$$

where $Q(F) = \lim_{n \to \infty} P(\hat{F}_n)$ and $\hat{F}_n$ is the empirical distribution function based on $Y_1, \ldots, Y_n$.

In other words, exchangeability of real-valued observables may be taken as equivalent to the HM

$$F \sim p(F) \quad \text{(prior)}$$

$$(Y_1, \ldots, Y_n | F) \overset{\text{iid}}{\sim} F \quad \text{(likelihood)} \quad (1.22)$$

for some prior distribution $p$ on the set $\mathcal{D}$ of all possible distribution functions on $\mathbb{R}$.

This prior makes the continuous form of de Finetti’s Theorem considerably harder to apply: to take the elicitation task seriously is to try to specify a measure on function space ($F$ is in effect an infinite-dimensional parameter). **NB** This task is not unique to Bayesians—you may just as well ask “Where does the likelihood come from?” in frequentist analyses of observational data as to ask “Where does the prior on the parameters come from?” in Bayesian modeling.)
CONTINUOUS OUTCOMES

The field of Bayesian nonparametrics, which began with work by Freedman (1963), Ferguson (1973, 1974), and others, has developed in an effort to put truly rich priors on \( \mathcal{D} \). This approach, however, has been stalled at an insufficiently practical stage until quite recently because of computational difficulties, but MCMC (Chapter 2) is changing that (Walker et al., 1997) at present. I will revisit this topic in Section 7.6.

**Model uncertainty.** Given that Bayesian nonparametrics is still basically at the pure research stage, what most Bayesians say they do in practice is to appeal to considerations that narrow down the field, such as an *a priori* judgment that the \( Y_i \) ought to be symmetrically distributed about a location parameter \( \mu \), and then try to use a plausible parametric family (the most popular is of course the Gaussian) satisfying (for instance) the symmetry restriction as a substitute for all of \( \mathcal{D} \). What most analysts (Bayesian and frequentist) actually do in practice is to look at the data when specifying their models: for example, with data on hospital length of stay for AMI patients, you might (a) make a histogram or kernel density trace of your sample \( y_1, \ldots, y_n \), (b) observe that the sample looks a lot like it follows a lognormal (LN) distribution, and (c) replace the infinite-dimensional elicitation problem in the first line of (1.22) by a two-dimensional elicitation problem on the parameters of the lognormal family. In other words, you would replace (1.22) with the vastly simpler HM

\[
(\mu, \sigma^2) \sim \ p(\mu, \sigma^2) \quad \text{(prior)}
\]

\[
(Y_1, \ldots, Y_n | \mu, \sigma^2) \overset{\text{IID}}{\sim} \ LN(\mu, \sigma^2). \quad \text{(likelihood)} \quad (1.23)
\]

Now the something-for-nothing bell should be going off in your head at this point: aren't we using the data twice with this approach (once to specify the prior on \( \mathcal{D} \), and once to draw inferences and make predictions given this choice of prior) and shouldn't we have to pay some price for doing so? This is the general problem of model uncertainty (e.g., Madigan and Raftery 1994, Draper 1995b), and it is not unique to Bayesians.

There is a real dilemma here: if you employ strategy \( S^* = \{\text{use the data to specify the model and then pretend you knew the resulting model all along}\} \), your conclusions are likely to be miscalibrated, in the direction of underpropagation of uncertainty (in other words, your nominal 90% predictive intervals may in fact only cover the actual observables (say) 65% of the time); but not
using $S^*$ can permit the data to surprise you in ways that would make you want to go back and revise your prior.

**NB** This last point is an example of what Lindley (19xx) calls **Cromwell's Rule**, which reminds us that it is dangerous to place prior probability 0 (or 1, for that matter; Problem 1.2) on anything, because it is then impossible to learn from any future data. For example, with any proposition $A$, setting $P(A) = 0$ forces $P(A | \text{data}) = P(A) \frac{P(\text{data} | A)}{P(\text{data})} = 0$, even if the data are highly likely under $A$ and highly unlikely under (not $A$). The application of this to model uncertainty is unfortunate: in practice people generally put nonzero probability on extremely small subsets of \{all possible models for a given data set\}, and yet doing so without looking at the data in effect forces many things that could easily be possible \((a \ priori)\) in the data to be impossible in your posterior analysis.

I will suggest a (partial) way out of this dilemma in Chapter 3, based on predictive validation. For the rest of the book, *faute de mieux*, I will generally either identify and work with conventional modeling choices, just to show where they lead, or use the $S^*$-plus-predictive-validation strategy.

<table>
<thead>
<tr>
<th>Value</th>
<th>375</th>
<th>392</th>
<th>393</th>
<th>397</th>
<th>398</th>
<th>399</th>
<th>400</th>
<th>401</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>402</th>
<th>403</th>
<th>404</th>
<th>405</th>
<th>406</th>
<th>407</th>
<th>408</th>
<th>409</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>12</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>410</th>
<th>411</th>
<th>412</th>
<th>413</th>
<th>415</th>
<th>418</th>
<th>423</th>
<th>437</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Case study 1.2:** *Measurement of physical constants.* What is now called the National Institute for Standards and Technology (NIST) in Washington, DC conducts extremely high precision measurement of physical constants, such as the actual weight of so-called **check-weights** that are supposed to serve as reference standards (like the official kg). In 1962–63, for example, back when their name was the National Bureau of Standards (NBS), $n = 100$ weighings (Table 1.1) of a block of metal called NB10, which was supposed to weigh exactly 10g, were made under conditions as close to IID as possible (Freedman et al., 1998). Figure 1.2 is a
normal q-q plot of the 100 measurements $y_1, \ldots, y_n$, which have a mean of $\bar{y} = 404.6$ (the units are micrograms below 10g) and an SD of $s = 6.5$.

Some natural questions that arise from the data in Table 1.1 include (a) How much does NB10 really weigh? (b) How certain are you given the data that the true weight of NB10 is less than (say) 405.25? (c) What is the underlying accuracy of the NB10 measuring process? And (d) How accurately can you predict the 101st measurement?

![Normal q-q plot of the NB10 data.](image)

**Figure 1.2. Normal q-q plot of the NB10 data.**

**A simple Gaussian model.** Evidently from Figure 1.2 it is plausible in answering these questions to assume symmetry of the "underlying distribution" $F$ in de Finetti's Theorem. One conventional choice, for instance, is the *Gaussian*:

\[
\begin{align*}
(\mu, \sigma^2) & \sim p(\mu, \sigma^2) \\
(Y_i | \mu, \sigma^2) & \iid N(\mu, \sigma^2).
\end{align*}
\]

(1.24)

**NB** People call the reciprocal of the variance $\sigma^2$ the *precision* of a distribution. In Bayesian work the precision is often the most intuitive scale on which to think about uncertainty or variability, as the results below will demonstrate.

(1.24) is our first example with more than one parameter, and we are still in the world of conjugate analysis because that's the
only computational tool I’ve discussed so far. Before I start in on how to specify the conjugate prior for \( \mu \) and \( \sigma^2 \) in this model, it is helpful to look at what happens in the simpler case in which you pretend that \( \sigma^2 \) is known. The conjugate prior for \( \mu \) (see Note 1.18) turns out (not too surprisingly, I guess) to be Gaussian; with this choice the model becomes

\[
\mu \sim N(\mu_0, \sigma^2_\mu) \\
(Y_1 | \mu) \sim \text{IID } N(\mu, \sigma^2)
\]

(1.25)

From the conjugacy of the prior, the posterior for \( \mu \) is also Gaussian, and it turns out that the posterior mean and variance have particularly simple expressions. Intuitively what is going on is this:

- The prior, considered as a data source, is Gaussian with mean \( \mu_0 \), variance \( \sigma^2_\mu \), and precision \( \frac{1}{\sigma^2_\mu} \).

- Notice from (1.19) that, in the mortality example of Section 1.6, the likelihood and posterior distributions depend on the data only through the sufficient statistic for the Bernoulli/binomial sampling distribution. In the same way you can (as people say) reduce by sufficiency here as well: the sufficient statistic for \( \mu \) in the Gaussian model with known variance is the sample mean \( \bar{Y} \). So to work out the form of the likelihood you just consider the sampling distribution of \( \bar{Y} \)—which is Gaussian with mean \( \mu \), variance \( \frac{\sigma^2}{n} \), and precision \( \frac{n}{\sigma^2} \)—as a function of \( \mu \) for fixed \( \bar{Y} \).

This distribution can be written \( c_1 \exp\left[ -c_2(\bar{Y} - \mu)^2 \right] \), and from this you can see that \( \bar{Y} \) and \( \mu \) play a symmetric role in it. So if I interchange the role of \( \mu \) and \( \bar{Y} \), I just get a Gaussian with the same variance and precision—\( \frac{\sigma^2}{n} \) and \( \frac{n}{\sigma^2} \), respectively—but now it’s a distribution for \( \mu \) with mean \( \bar{Y} \).

- In the mortality example the posterior mean was a weighted average of the prior mean and the data mean, with weights given by the prior effective sample size \( n^* \) and the data sample size \( n \). This turns out to be a general result with conjugate analysis, so the same trick applies here, but what is \( n^* \) in this case? You can show that the right weights in the weighted average are given by the precisions of the prior and likelihood data sources:

\[
E(\mu | \bar{y}) = \frac{\left( \frac{1}{\sigma^2_\mu} \right) \mu_0 + \left( \frac{n}{\sigma^2} \right) \bar{y}}{\left( \frac{1}{\sigma^2_\mu} \right) + \left( \frac{n}{\sigma^2} \right)} = \frac{\left( \frac{\sigma^2}{\sigma^2_\mu} \right) \mu_0 + n \bar{y}}{\left( \frac{\sigma^2}{\sigma^2_\mu} \right) + n}.
\]

(1.26)
This also demonstrates along the way that \( n^* = \frac{\sigma_\mu^2}{\sigma_\mu^2} \).

- Finally, what about the posterior variance \( V(\mu|y) \)? Based on what happened with the prior mean, you can guess that the posterior variance would be driven by the prior and likelihood precisions, and in fact it turns out that on the precision scale the accuracy of the information sources is additive (which is why Bayesians like the idea of precision so much):

\[
\begin{pmatrix}
\text{posterior precision} \\
\text{precision}
\end{pmatrix}
= \begin{pmatrix}
\text{prior precision} \\
\text{precision}
\end{pmatrix}
+ \begin{pmatrix}
\text{likelihood precision} \\
\text{precision}
\end{pmatrix},
\]

from which

\[
V(\mu|y) = \frac{1}{\left(\frac{1}{\sigma_\mu^2}\right) + \left(\frac{3}{\sigma_\nu^2}\right)} = \frac{\sigma^2}{n^* + n}.
\]

Some unpleasant algebra, with which I will not burden you, verifies all of the above intuition. A few points to note:

- The idea, from Section 1.6, of the prior being equivalent to a data set works again in this case: it is as if a data set with \( n^* \) observations and mean \( \mu_0 \) were merged with the observed data set \( y \) and a frequentist analysis were conducted on the merged data. This is also a general feature of conjugate analysis.

- The concept of little or no prior information here corresponds to the prior SD \( \sigma_\mu \) being large, or equivalently the prior precision \( \frac{1}{\sigma_\mu^2} \) being small. In the limit as \( \sigma_\mu \to \infty \) the prior sample size would go to 0, and you can see from (1.26) and (1.28) that the Bayesian results would coincide with the usual frequentist answers.

NB I have been using the term diffuse to convey the idea of a prior distribution embodying little or no information about the parameter in question. Many other Bayesians talk about non-informative priors in this situation, but I don’t like this terminology, because every choice of prior (diffuse or not) conveys information, namely your choice—which needs to be defended in each case—for the appropriate effective prior sample size. I will stick with diffuse in what follows.

Bayesian inference with multivariate \( \theta \). Returning now to (1.24) with \( \sigma^2 \) unknown, this model has a \((p = 2)\)-dimensional parameter, \( \theta = (\mu, \sigma^2) \). When \( p > 1 \) you can still use Bayes’ Theorem
directly to obtain the joint posterior distribution,

\[
p(\theta \mid y) = c \cdot p(\theta) l(\theta \mid y) = p(\mu, \sigma^2 \mid y) = c \cdot p(\mu, \sigma^2) l(\mu, \sigma^2 \mid y),
\]

where \( y = (y_1, \ldots, y_n) \), although making this calculation directly requires a \( p \)-dimensional integration to evaluate \( c \)—for example, in this case

\[
c = [p(y)]^{-1} = \left( \int \int p(\mu, \sigma^2, y) \, d\mu \, d\sigma^2 \right)^{-1} = \left( \int \int p(\mu, \sigma^2) l(\mu, \sigma^2 \mid y) \, d\mu \, d\sigma^2 \right)^{-1}.
\]

Usually, however, you will be more interested in the marginal posterior distributions, in this case \( p(\mu \mid y) \) and \( p(\sigma^2 \mid y) \). Obtaining these requires \( p \) integrations, each of dimension \( (p-1) \), a process that people refer to as marginalization or integrating out the nuisance parameters. For example,

\[
p(\mu \mid y) = \int p(\mu, \sigma^2 \mid y) \, d\sigma^2.
\]

**Predictive** distributions also involve a \( p \)-dimensional integration: for example, with \( y = (y_1, \ldots, y_n) \),

\[
p(y_{n+1} \mid y) = \int \int p(y_{n+1}, \mu, \sigma^2 \mid y) \, d\mu \, d\sigma^2
\]

\[
= \int \int p(y_{n+1} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid y) \, d\mu \, d\sigma^2.
\]

And, finally, if you are interested in a function of the parameters, you have some more hard integrations ahead of you. For instance, suppose you wanted the posterior distribution for the coefficient of variation \( \lambda = g_1(\mu, \sigma^2) = \frac{\mu}{\sqrt{\sigma^2}} \) in model (1.24). Then one fairly direct way to get this posterior (e.g., Bernardo and Smith, 1994) is to (a) introduce a second function of the parameters, say \( \eta = g_2(\mu, \sigma^2) \), such that the mapping \( f = (g_1, g_2) \) from \((\mu, \sigma^2)\) to \((\lambda, \eta)\) is invertible; (b) compute the joint posterior for \((\lambda, \eta)\) through the usual change-of-variables formula

\[
p(\lambda, \eta \mid y) = p_{\mu, \sigma^2}[f^{-1}(\lambda, \eta) \mid y] \cdot |J_{f^{-1}}(\lambda, \eta)|,
\]

where \( p_{\mu, \sigma^2}(\cdot \mid y) \) is the joint posterior for \( \mu \) and \( \sigma^2 \) and \(|J_{f^{-1}}|\) is the determinant of the Jacobian of the inverse transformation;
and (c) marginalize in $\lambda$ by integrating out $\eta$ in $p(\lambda, \eta|y)$, in a manner analogous to (1.31). (Here, for instance, $\eta = g_\delta(\mu, \sigma^2) = \sqrt{\sigma^2}$ would create an invertible $f$, with inverse defined by $(\mu = \lambda \eta, \sigma^2 = \eta^2)$; the Jacobian determinant comes out $2\lambda \eta$ and (1.33) becomes $p(\lambda, \eta|y) = 2\lambda \eta p_{\mu, \sigma^2}(\lambda \eta, \eta^2|y)$. This process involves two integrations, one to get the normalizing constant that defines (1.33) and one to get rid of $\eta$.

You can see that when $p$ is a lot bigger than 2 all these integrals may create severe computational problems—this has been the big stumbling block for applied Bayesian work for a long time.

More than 200 years ago Laplace (1774)—perhaps the second applied Bayesian in history (after Bayes himself)—developed, as one avenue of solution to this problem, what people now call Laplace approximations to high-dimensional integrals of the type arising in Bayesian calculations (see, e.g., Tierney and Kadane, 1986). I will cover Laplace approximations only briefly in this book, in Chapter 8; Chapter 2 details how MCMC may be used as an alternative solution to the integration problem.

**The full Gaussian case.** The conjugate prior for $(\mu, \sigma^2)$ in the model (1.24) (e.g., Gelman et al., 1995) turns out to be most simply described hierarchically:

$$\sigma^2 \sim SI\chi^2(\nu_0, \sigma_0^2)$$

$$\mu | \sigma^2 \sim N\left(\mu_0, \sigma^2 / \kappa_0\right) \quad (1.34)$$

Here saying that $\sigma^2 \sim SI\chi^2(\nu_0, \sigma_0^2)$, where SI stands for scaled inverse, amounts to saying that $\tau^2 \equiv 1 / \sigma^2$ follows a scaled $\chi^2$ distribution with parameters $\nu_0$ and $\sigma_0^2$ (see Appendix 1 for details). The scaling is chosen so that $\sigma_0^2$ can be interpreted as a prior estimate of $\sigma^2$, with $\nu_0$ the prior effective sample size of this estimate (in other words, as in the beta-Bernoulli/binomial model of Section 1.6, think of a prior data set with $\nu_0$ observations and sample variance $\sigma_0^2$). The parameters $\mu_0$ and $\kappa_0$ in the second level of the prior model (1.34) have simple parallel interpretations to those of $\sigma_0^2$ and $\nu_0$: $\mu_0$ is the prior estimate of $\mu$, and $\kappa_0$ is the prior effective sample size of this estimate.

In the Gaussian model (1.24, 1.34), which I will abbreviate $G$, the integrations may be done analytically (e.g., Gelman et al., 1995), yielding
\[
\begin{align*}
(\sigma^2 \mid y, G) & \sim SI - \chi^2(\nu_n, \sigma_n^2) \\
(\mu \mid y, G) & \sim t_{\nu_n} \left( \mu_n, \frac{\sigma_n^2}{\kappa_n} \right), \text{ where} \\
\nu_n & = \nu_0 + n, \quad \kappa_n = \kappa_0 + n, \\
\nu_n \sigma_n^2 & = \nu_0 \sigma_0^2 + (n - 1)s^2 + \frac{\kappa_0n}{\kappa_n} (\bar{y} - \mu_0)^2, \\
\mu_n & = \frac{\kappa_0}{\kappa_n} \mu_0 + \frac{n}{\kappa_n} \bar{y}.
\end{align*}
\]

Here \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) and \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \) are the usual sample mean and variance of \( y \), and \( t_{\nu} (\mu, \sigma^2) \) is a scaled version of the usual \( t_{\nu} \) distribution (Appendix 1): \( W \sim t_{\nu} (\mu, \sigma^2) \) just means that \( \frac{W - \mu}{\sigma} \sim t_{\nu} \). Once again, from the conjugacy, the posterior mean for \( \mu \) in (1.35) is a weighted average of the prior mean and data mean, with weights determined by the effective prior sample size and the data sample size.

**NB10 Gaussian results.** (1.35) may be used to answer the four questions listed below Table 1.1, as follows.

**Question (a):** I don’t know anything *a priori* about what NB10 is supposed to weigh (down to the nearest microgram) or about the accuracy of the NBS’s measurement process, so I want to use a diffuse prior for \( \mu \) and \( \sigma^2 \). Considering the meaning of the hyperparameters, to provide little prior information I want to choose both \( \nu_0 \) and \( \kappa_0 \) close to 0. Making them exactly 0 would produce an improper prior distribution (which doesn’t integrate to 1), but choosing positive values as close to 0 as you like yields a proper and highly diffuse prior.

You can see from (1.35) that the result for large \( n \) is then
\[
(\mu \mid y, G) \sim t_n \left( \bar{y}, \frac{(n - 1)s^2}{n^2} \right) \stackrel{d}{=} N \left( \bar{y}, \frac{s^2}{n} \right);
\]

in other words, with diffuse prior information (and as with the Bernoulli model in Section 1.6) the 95% central Bayesian interval virtually coincides with the usual frequentist 95% confidence interval \( \bar{y} \pm t_{n-1}^{0.975} \frac{s}{\sqrt{n}} \approx 404.6 \pm 1.98 \cdot 0.647 = (403.3, 405.9) \). Thus both \{frequentists who assume \( G \)\} and \{Bayesians who assume \( G \) with a diffuse prior\} conclude that NB10 weighs about 404.6\( \mu g \) below 10g, give or take about 0.65\( \mu g \).
CONTINUOUS OUTCOMES

Figure 1.3. Posterior distribution for $\mu$ in the NB10 example, with the \{\mu < 405.25\} region shaded.

Question (b). If interest focuses on whether NB10 weighs less than some value like 405.25, when reasoning in a Bayesian way you can answer this question directly: the posterior distribution for $\mu$ is shown in Figure 1.3, and $P_B(\mu < 405.25 \mid y, G, \text{diffuse prior}) \approx 0.85$. In other words, with these assumptions there are pretty good betting odds—about 5.5 to 1—in favor of the proposition that $\mu < 405.25$.

With your frequentist hat on, $P_F(\mu < 405.25)$ is undefined; about the best you can do is to test $H_0$: $\mu < 405.25$, for which the $p$-value would (approximately) be $p = P_{F, \mu=405.25}(\bar{y} > 405.59) = 1 - 0.85 = 0.15$. This would constitute “insufficient evidence to reject $H_0$ at the usual significance levels,” leaving an inferential impression that contrasts with the reasonably clear Bayesian betting odds. **NB** (1) The significance test tries to answer a different question: in Bayesian language it looks at $P(\bar{y} \mid \mu)$ instead of $P(\mu \mid \bar{y})$. (2) You can see that as with confidence intervals, when a diffuse prior seems appropriate, there is a direct relationship—at least with one-sided tests$^{21}$—between frequentist and Bayesian results: for testing $H_0$: $\mu < c$, the $p$-value is just $p = 1 - P_B(\mu < c \mid y, \text{diffuse prior})$. Thus there is a certain justification in one-sided testing problems for the conclusion, which people sometimes wish to draw, that the $p$ value is the probability that the null hypothesis is false.
**Question (c).** The conjugacy of (1.34) means that the assumption of a scaled inverse $\chi^2$ prior for $\sigma^2$, with hyperparameters $\nu_0$ and $\sigma_0^2$, also produces a $SI-\chi^2$ posterior for $\sigma^2$, with the following parameters (see, e.g., Gelman et al., 1995):

$$
\begin{align*}
(\sigma^2|y, G) & \sim SI-\chi^2(\nu_n, \sigma_n^2), \quad \text{where} \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \frac{1}{\nu_n} \left[ \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n} (\bar{y} - \mu_0)^2 \right].
\end{align*}
\tag{1.37}
$$

The form of $\sigma_n^2$, which acts (for large $\nu_n$, at least) like a posterior estimate of $\sigma^2$, is interesting: the first two terms in $\sigma_n^2$ are (almost) a weighted average of the prior and sample estimates $\sigma_0^2$ and $s^2$ of $\sigma^2$, and there is also a contribution arising from the discrepancy, if any, between the sample mean $\bar{y}$ and the prior mean $\mu_0$.

As in the answer to question (a), a diffuse prior would correspond to choosing $\nu_0$ and $\kappa_0$ close to 0, which would produce the result

$$
(\sigma^2|y, G, \text{diffuse prior, large } n) \sim SI-\chi^2(n, \frac{n-1}{n}s^2). \tag{1.38}
$$

Now you can also show (Problem 1.8) that (1.38) is equivalent to saying that the posterior distribution for the precision $\frac{1}{\sigma^2}$ is gamma with parameters $\frac{n}{2}$ and $\frac{n-1}{2}s^2$, which I will denote $\Gamma\left(\frac{n}{2}, \frac{n-1}{2}s^2\right)$.

This means that the posterior for $\frac{(n-1)s^2}{\sigma^2}$ is $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, which is another way of writing the $\chi^2$ distribution with $(n-1)$ degrees of freedom. But this is just a Bayesian interpretation of the usual frequentist inference for $\sigma^2$ in the Gaussian model with both $\mu$ and $\sigma^2$ unknown (e.g., Snedecor and Cochran, 1980): the sampling distribution of $\frac{(n-1)s^2}{\sigma^2}$ in this model, viewing $\sigma^2$ as fixed and $s^2$ as random, is $\chi^2_{n-1}$. Thus, as in the answer to question (a), {Bayesians with diffuse prior information} and {frequentists} would get the same 95% (central) intervals for $\sigma^2$ and $\sigma$: the NB10 sample SD of $s = 6.5$ produces the 95% interval estimates (32.2, 56.4) and (5.68, 7.51) of $\sigma^2$ and $\sigma$, respectively.

**An informative prior in the Gaussian model.** Simply for illustration, suppose that information from other studies at the NBS before the NB10 data were collected—taking suitable account of any differences between the previous studies and the present measurement method—had suggested that $\sigma$ should be around $\sigma_0 = 10$, with (say) 90% *a priori* limits of roughly $\sigma_{10} = 6, \sigma_{hi} = 34$, and that $\mu$ should be around $\mu_0 = 403$ with 90% prior limits of
CONTINUOUS OUTCOMES

approximately \((\mu_0 = 396, \mu_{hi} = 410)\). To fit this information into the conjugate structure (1.34), I have to find the corresponding prior effective sample sizes \(\nu_0\) and \(\kappa_0\).

Considering \(\nu_0\) first, the fact noted above that \(\sigma^2 \sim SI-\chi^2(\nu_0, \sigma_0^2)\) iff \(\frac{\nu_0\sigma_0^2}{\sigma^2} \sim \chi^2_{\nu_0}\) makes me want to work with the precision instead of the SD in setting up an equation to determine \(\nu_0\), since I have a \(\chi^2\) CDF handy in S+:

\[
0.9 = P(\sigma_{\nu_0} < \sigma < \sigma_{hi}) = P \left( \frac{\nu_0 \sigma_0^2}{\sigma_{hi}^2} < \frac{\nu_0 \sigma_0^2}{\sigma^2} < \frac{\nu_0 \sigma_0^2}{\sigma_{lo}^2} \right) = F_{\chi^2_{\nu_0}} \left( \frac{100 \nu_0}{1156} \right) - F_{\chi^2_{\nu_0}} \left( \frac{100 \nu_0}{36} \right),
\]

where \(F_{\chi^2_{\nu_0}}\) is the \(\chi^2_{\nu_0}\) CDF. Trial and error with this CDF now shows that \(\nu_0 \approx 2.5\).

In specifying \(\kappa_0\) it is helpful to appeal to a fact about the distribution of the \(\mu\) in the conjugate prior specification (1.34). You can show (Problem 1.10) that if \(\sigma^2\) is \(SI-\chi^2\) and \((\mu|\sigma^2)\) is Gaussian then the marginal distribution of \(\mu\) is scaled \(t\):

\[
(1.34) \text{ implies that } \mu \sim t_{\nu_0} \left( \mu_0, \frac{\sigma_0^2}{\kappa_0} \right). \tag{1.40}
\]

From the CDF of the standard \(t\) distribution with \(\nu_0 = 2.5\) degrees of freedom,

\[
P \left( \left| \frac{\mu - \mu_0}{\sqrt{\sigma_0^2/\kappa_0}} \right| \leq 2.56 \right) = 0.9, \text{ yielding}
\]

\[
\kappa_0 = \frac{2.56^2 \sigma_0^2}{(\mu_{hi} - \mu_0)^2} \approx 13. \tag{1.41}
\]

With these values of \(\nu_0\) and \(\kappa_0\) and the NB10 data, (1.37) produces the posterior parameters \(\nu_n = 102.5\) and

\[
\sigma_n^2 = \frac{250 + 4140.19 + 29.08}{102.5} \approx 43.11. \tag{1.42}
\]

This is a bit bigger than the sample variance \(s^2 = 41.82\), both because the prior estimate of \(\sigma^2\) (100) is a lot bigger than \(s^2\) and because of the modest discrepancy between the prior and data means. The prior-to-posterior analysis here is not far from the simple up-
dating rule

\[
p(\sigma^2, \mathcal{G}) = \text{SI-\chi}^2(\nu_0, \sigma_0^2), \quad \text{(prior)}
\]

\[
l(\sigma^2 | y, \mathcal{G}) = \text{SI-\chi}^2(n, s^2) \quad \text{(likelihood)} \quad (1.43)
\]

\[
p(\sigma^2 | y, \mathcal{G}) = \text{SI-\chi}^2(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + ns^2}{\nu_0 + n}) \quad \text{(posterior)},
\]

which would have been exact in the Gaussian model if \( \mu \) had been known (Problem 1.11). The three distributions in (1.43) are plotted in Figure 1.4; you can see that the prior information has tugged the posterior to the right of the data information, but not very much because the prior effective sample sizes were small.

![Graph showing prior, likelihood, and posterior distributions for \( \sigma^2 \)]

**Figure 1.4.** Prior, likelihood, and posterior distributions for \( \sigma^2 \) with the Gaussian model (1.24, 1.34) and an informative prior applied to the NB10 data.

**Question (d).** Analytic integration in the Gaussian model (e.g., Bernardo and Smith, 1994) yields

\[
(y_{n+1} \mid y, \mathcal{G}) \sim t_{\nu_n} \left( \mu_n, \frac{\kappa_n + 1}{\kappa_n} \sigma_n^2 \right), \quad (1.44)
\]

and for \( n \) large and \( \nu_0 \) and \( \kappa_0 \) close to 0 this is \( (y_{n+1} \mid y) \sim N(\bar{y}, s^2) \) (the basis of the usual frequentist answer), yielding a 95% posterior predictive interval for \( y_{n+1} \) of (392, 418).
Model expansion. A standardized version of the predictive distribution (1.44) for the NB10 data is plotted in Figure 1.5, with the standardized data values superimposed. It is evident from this plot (and also from Figure 1.2) that the Gaussian model provides a poor fit for these data—the two most extreme points in the data set in standard units are $-4.6$ and $5.0$. With the symmetric heavy tails indicated in these plots, in fact, the empirical CDF looks quite a bit like that of a $t$ distribution with a rather small number of degrees of freedom $\nu$. This suggests revising the previous model by expanding it: embedding the Gaussian in the $t$ family and adding the parameter $\nu$ for tail-weight.

Unfortunately there is no standard closed-form conjugate choice for the prior on $\nu$. A more flexible approach to computing is evidently needed: this is the subject of Chapter 2.

1.9 Additional reading

[xx finish this]
1.10 Problems

[xx this section is still quite rough]

1.1 Consider the model $\theta \sim p(\theta)$, $(Y_i|\theta) \overset{\text{iid}}{\sim} B(\theta)$ for some prior $p(\theta)$ for which the prior mean $E(\theta)$ and variance $V(\theta)$ are both nonzero. Show in this model that $P(Y_2 = 1) < P(Y_2 = 1|Y_1 = 1)$, thereby concretely demonstrating that without conditioning on $\theta$ the $Y_i$ are dependent. Thus iid $\not\rightarrow$ exchangeability but not conversely.

1.2 I said in Section 1.3 that "...with my frequentist hat on I regard $\theta$ as a fixed (unknown) constant and $Y$ as a random variable, and everything focuses on imagining what would happen as $Y$ changes randomly from sample to sample." This is actually the logical position for frequentists before the data arrive. What about after the data arrive—once $Y$ is observed, in the frequentist approach is it still random, or is it now fixed? If it’s still random, then is it fair to say that frequentists don’t condition on the data? If it’s now fixed, then both $\theta$ and $Y$ are fixed, and where does probability come in? Discuss.

1.3 The Enzyme-Linked ImmunoSorbent Assay (ELISA) test was approved by many countries around the world in the mid-1980s to screen donated blood for the presence of the AIDS virus HIV. The test works by detecting antibodies—substances that the body produces when the virus is present—but, as with any screening test, in practice it makes some mistakes. ELISA was designed so that when a given blood sample does in fact contain a clinically meaningful concentration of HIV, the test gives a positive result (that is, ELISA reports that in its opinion this blood sample has HIV in it) $\alpha = 98\%$ of the time: this is referred to as ELISA’s sensitivity. Moreover, when the blood being tested is not contaminated with the virus ELISA will announce a negative result $\beta = 93\%$ of the time: this is ELISA’s specificity. The prevalence of HIV-positivity in the population of people who donate blood to blood banks is thought to be about $\pi = 1\%$.

(a) Letting $A = \{\text{person is HIV-positive}\}$ and $+ = \{\text{ELISA positive}\}$, express the three numerical facts above in unconditional and conditional probability terms, and use Bayes’ Theorem to show that if someone donates blood and the ELISA test comes out negative, the probability the person is not in
fact HIV-positive given this negative result is virtually 100%, but if ELISA comes out positive the probability the person actually is HIV-positive is only $p = \frac{95}{93} \approx 12\%$. Explain these results by (i) exploring symbolically and numerically how $p$ depends on $\alpha$, $\beta$, and $\pi$, and noting what it is about the given values of these three quantities that has made $p$ so low; and (ii) identifying the two kinds of mistakes ELISA could make and discussing their implications from the blood bank’s point of view.

(b) In practice it is possible to “tune” screening tests like ELISA by changing the threshold of antibodies required to announce a positive result, which will act on the 98% sensitivity and 93% specificity values mentioned above in a tug-of-war fashion: you can increase the sensitivity, for instance, but only by allowing the specificity to decrease (and vice versa). If ELISA were to be made available as a screening test to the general population (for instance, suppose that people were able to send a blood sample to a private lab confidentially and get back the ELISA diagnosis for a fee), which would it be better to increase: ELISA’s sensitivity or specificity? What would happen if ELISA, with its present $\alpha$ and $\beta$, were used as a public health tool in a mass screening program of all Americans, as some members of the US Congress suggested back in the 1980s? Explain.

1.4 Prove the other part of Cromwell’s Rule (Section 1.8): With any proposition $A$, setting $P(A) = 1$ forces $P(A \mid \text{data}) = 1$ no matter how likely the data are under $A$ and (not $A$).

1.5 I used to work at the University of California in Los Angeles, and I like to drink tea while I’m working. For the first few weeks after starting work I didn’t have access to any facilities for making tea, so I would go down twice a day to a vending machine and pay 25 cents for a cup of brown liquid that the machine claimed was “tea.” On the front of the machine there was a bright yellow label that said something like, “Maybe you’ll be lucky!! Every now and then, at random, this machine will give you your quarter back, and your beverage will be free!!” At the end of almost two months, having spent $n = 78$ quarters without getting “lucky,” it occurred to me that the company that owned the machine may have just decided it was cheaper
to put the yellow label on the front than to install any randomization device inside that would actually make refunds. Let $Y_i = 1$ if I got a free cup of “tea” on occasion $i$ and assume the Bernoulli model (1.11) with a prior that appropriately reflects the company’s desire to make money (for example, do you really believe that $\theta = 0.5$ is as likely a priori as $\theta = 0.008$?). Use this model to discriminate between the two possible explanations $\{\theta = 0\}$ and $\{\theta > 0\}$ by calculating their posterior probabilities given the data, and explain why the frequentist $p$ value for testing $H_0: \theta = 0$ would be completely useless in this situation.

1.6 Consider a univariate parameter $\theta$ and a data set $Y = (Y_1, \ldots, Y_n)$ that is IID from some sampling distribution given $\theta$. Learning about $\theta$ from $Y$ in the Bayesian approach involves updating from $p(\theta)$ to $p(\theta|Y)$, and it is interesting, from an experimental design point of view, to examine what may be said in general about the relationship between these two distributions before $Y$ is observed.

(a) The first part of the double expectation theorem from introductory probability says that the prior mean $E(\theta)$ and the posterior mean $E(\theta|Y)$ are related by

$$E(\theta) = E_Y[E(\theta|Y)],$$

(1.45)

where the right side of (1.45) involves averaging over possible data sets $Y$. Explain what this implies—if you were planning on quoting the mean as a point estimate of $\theta$—about the effect you expect the data to have on your prior point estimate. Is this intuitively reasonable? Explain.

(b) The second part of the double expectation theorem says that the prior and posterior variances $V(\theta)$ and $V(\theta|Y)$, respectively, are related by

$$V(\theta) = V_Y[E(\theta|Y)] + E_Y[V(\theta|Y)].$$

(1.46)

Show that this means that, averaging over possible $Y$, you expect to learn about $\theta$, in the sense that you expect the posterior variance to be no larger than the prior variance. However, by creating an explicit example (prior and data set) in the beta/Bernoulli model (1.18) or the Gaussian model (1.24, 1.34), also show that it is possible for you to “know less” after you see $Y$ than before, in that the posterior variance can


be larger than the prior variance. Explain in concrete terms what feature of the relationship between the prior and data information causes this to happen.

1.7 Continuing Problem 1.6, show in the simple Gaussian model (1.25) with known \( \sigma^2 \) that, no matter how discrepant the prior mean and the data mean are, the posterior variance will always be smaller than either the prior variance or the “data variance” (the variance of the density-normalized likelihood). What is it about (1.25) that produces this undesirable result, and what is it about the full Gaussian model (1.24, 1.34) that remedies the defect? Explain.

1.8 Relationships between gamma, inverse gamma, \( \chi^2 \), inverse \( \chi^2 \), scaled inverse \( \chi^2 \) [xx to be finished].

1.9 (A1) Poisson-gamma; negative binomial [xx to be finished].

1.10 Show that if \( \sigma^2 \) is \( SI-\chi^2 \) and \( (\mu|\sigma^2) \) is Gaussian then the marginal distribution of \( \mu \) is scaled \( t \) [xx to be finished].

1.11 Prove (1.43) in the normal model with known mean [xx to be finished].

1.12 Consider the simple Gaussian model (1.25) and make it even simpler by taking \( n = 1 \): \( \theta \sim N(\mu_0|\sigma^2) \), \( (Y|\theta) \sim N(\mu, \sigma^2) \) for known \( \sigma^2 \). Before you have seen \( Y \) this is like a bivariate sampling model for \( (\theta, Y) \): \( \theta \) is drawn from a Gaussian, and then conditional on \( \theta \), \( Y \) is drawn from another Gaussian. This makes me think of a elliptical (why?) scatter plot for \( (\theta, Y) \) and brings up the idea of regression as an alternative way to understand how Bayes’ Theorem works in this model.

It follows from the assumptions so far that both the marginal distribution of \( Y \) and the conditional distribution of \( \theta \) given \( Y \) are also Gaussian (this was Galton’s original way of thinking about regression more than 100 years ago, in fact; see Stigler, 1986). Use the double expectation theorem and anything you know about regression to derive the posterior mean and variance of \( \theta \) given \( Y \).

1.13 Sequential updating: show that you get the same thing when you sequentially absorb \( y_1, \ldots, y_n \) as when you simultaneously absorb them.

[I will supply more problems later, and the new problems will be more interesting and data-oriented than, e.g., 1.10 and 1.11 ... xx to be continued].
1.11 Notes

1.1 *Process* is what health care providers do on behalf of patients; *outcomes* are what happens as a result of that care. Saying that a disease has a strong process-outcome link just means that research has demonstrated for that disease that good process leads to good outcomes and bad process to bad outcomes.

1.2 In the history of ideas this branch of mathematics is relatively new—the ancient Greeks, for example, had no notion of probability. The subject seems to have come into focus fairly suddenly in about 1660, in the independent work of a variety of people including Leibniz (Germany), Pascal (France), Huygens (Holland), and Graunt (England); see Hacking (1975) and Stigler (1986).

1.3 I strongly recommend this excellent book to anyone interested in the foundations of probability and statistics. Oakes presents a devastating critique of significance testing, an interesting account of the various ways people have tried to connect probability with the real world, a comparative evaluation of the leading schools of statistical inference, and a discussion of the role of statistics in the social sciences.

1.4 For example, suppose you are trying to quantify your uncertainty about the probability \( \theta \) of something happening, and you want to express the judgment that any value for \( \theta \) from 0 to 1 (inclusive) is equally plausible (the *principle of insufficient reason*). OK, but what if you had asked yourself the same question about \( f(\theta) \) for some monotone \( f \), like \( \theta^2 \)? You cannot claim that any value for \( \theta \) from 0 to 1 (inclusive) is equally plausible at the same time as you are claiming that any value for \( \theta^2 \) from 0 to 1 (inclusive) is equally plausible. This was what bothered Fisher (1922) about the Bayesian need to specify a prior. However, as Lindley (19xx) points out [Dennis, please help me with a relevant reference], in the language of Note 1.17 below this is actually a feature, not a bug: even if you are pretty unsure of the value of \( \theta \), you are pretty darn sure that \( \theta^{1000} \) is close to 0.

1.5 This is especially likely in sparse cells in the *equivalence grid* formed conceptually by crossing categorical versions of the predictor variables with each other.

1.6 To a Bayesian saying that \( P_\theta(A) \) is “objective” just means that lots of people more or less agree on its value.
1.7 It has bothered me for a long time that almost nobody talks about modeling issues like this in the frequentist approach—instead, people rush directly to "Let $y_1, \ldots, y_n$ be IID," without saying anything about why random variables are appropriate with samples of convenience and data from observational studies. A few exceptions: the lovely introductory statistics book by Freedman et al. (1998), and Mallows (1998).

1.8 Of course, in practice nobody would treat this argument seriously until you compared the observed mortality at the RUH with its expected mortality given how sick its AMI patients were on admission; see, e.g., Keeler et al. (1990).

1.9 When necessary I will use the standard convention of writing random variables in upper case and the values they take on in lower-case.

1.10 I have repeated this observation several times to emphasize, for people who have so far only thought about probability from the frequentist viewpoint, that something fundamentally different is going on here with the random variables in the Bayesian approach.

1.11 Finite versions of de Finetti's Theorem are available (Diaconis and Freedman, 1980; Bernardo and Smith, 1994): call an exchangeable sequence $\{y_i, i \leq n\}$ $N$-extendable if it is the first $n$ elements of a longer exchangeable sequence $\{y_i, i \leq N\}$. (Infinite exchangeability, as in Theorem 1.7, amounts to assuming $N$-extendability for all $N > n$.) Then (1.7) is a good approximation to (1.6) when $N >> n$. In practice this means that you regard the process of observing 1's and 0's to be time-homogeneous across a horizon that is considerably broader than the first $n$ observations—in other words, we are back in effect to the frequentist difficulty of having to define a population. There is no free lunch with de Finetti's Theorem.

1.12 I used to think that de Finetti's Theorem says that if your uncertainty about the $Y_i$ is exchangeable then you must express your predictive distribution (1.6) for the $Y_i$ in the form (1.7), but (as Sander Greenland pointed out to me) in fact all the theorem says is that you can express it in this form. In practice, however, de Finetti's representation is so straightforward to work with that most people just move directly from the exchangeability judgment to (1.7).
1.13 This is a slightly nonstandard use of the term hierarchical model; many people think of HMs as models for situations with data at all levels of the hierarchy, although in such fields as meta-analysis (Chapter 3) even this convention is more honored in the breach than in the observance.

1.14 Since we may as well model $\theta$ as a continuous quantity between 0 and 1, $p(\theta)$ is an ordinary continuous probability density, just like any frequentist-style sampling distribution on (0,1).

1.15 Thomas Bayes (1701?–1761) was an English cleric, philosopher, and mathematician, interested in the foundations of probability and (what we would now think of as) statistics, who managed to make a place for himself in history without a single mathematical publication in his lifetime. Stigler (1986) has a lot of interesting material on what Bayes actually did and did not do. For instance, in the famous essay that he did not allow to be published until after his death, Bayes (1763) posed and solved the following problem, in present-day notation: if $\theta \sim U(0,1)$ and $(Y|\theta) \sim \text{bin}(n,\theta)$ then compute $P(a < \theta < b|Y)$ for any $a$ and $b$. The main controversy concerned the universal appropriateness of his choice of a uniform prior distribution for $\theta$ in real-world problems.

1.16 The formalism of coherence is best understood within the context of Bayesian decision theory. Axiomatic approaches to rational decision-making date back to Ramsay (1931/1980), with von Neumann and Morgenstern (1944) and Savage (1954) also making major contributions. The ingredients of a general decision problem (e.g., Bernardo and Smith, 1994) include

- A set $\{a_i, i \in I\}$ of available actions, one of which you will choose;
- For each action $a_i$, a set $\{E_j, j \in J\}$ of uncertain outcomes describing what will happen if you choose action $a_i$;
- A set $\{c_j, j \in J\}$ of consequences corresponding to the outcomes $\{E_j, j \in J\}$; and
- A preference relation $\leq$, expressing your preferences between pairs of available actions ($a_1 \leq a_2$ means "$a_1$ is not preferred by you to $a_2$"). Define $a_1 \sim a_2$ ("$a_1$ and $a_2$ are equivalent" to you) iff $a_1 \leq a_2$ and $a_2 \leq a_1$.

This preference relation induces a qualitative ordering of the uncertain outcomes ($E \leq F$ means "$E$ is not more likely than
Because if you compare two dichotomized possible actions, involving the same consequences and differing only in their uncertain outcomes, the fact that you prefer one action to another means that you must judge it more likely that if you take that action the preferred consequence will result.

Within this framework you have to make further assumptions—the coherence axioms—to ensure that your actions are internally consistent. Informally these are:

- An axiom insisting that you be willing to express preferences between simple dichotomized possible actions \( \{a, \text{not } a\} \);
- A transitivity axiom in which (for all actions \( a, a_1, a_2, a_3 \)) \( a \leq a \), and if \( a_1 \leq a_2 \) and \( a_2 \leq a_3 \) then \( a_1 \leq a_3 \); and
- An axiom based on the sure-thing principle (Savage, 1954): if, in two situations, no matter how the first comes out the corresponding outcome in the second is preferable, then you should prefer the second situation overall.

This puts \( \leq \) on a sound footing for qualitative uncertainty assessment, but does not yet imply how to quantify—it’s like being able to say that one thing weighs less than another but not to say by how much. To go further requires a fourth assumption, analogous to the existence of a set of reference standards (for example, an official kg weight, half-kg, and so on) and the ability to make arbitrarily precise comparisons with these standards:

- An axiom guaranteeing that for each outcome \( E \) there exists a standard outcome \( S \) (for instance, “idealized coin lands heads”) such that \( E \sim S \).

This framework implies the existence and uniqueness of a (personal) probability \( P_{B,you} \) (abbreviated \( P \)), mapping from outcomes \( E \) to \([0,1]\) and corresponding to the judgments in your definition of \( \leq \), and a utility function \( U_{you} \) (abbreviated \( U \); large values preferred, say), mapping from consequences \( c \) to the real line and quantifying your preferences.

This has all been rather abstract. Four concrete results arising from this framework may make its implications clearer:

- Bayes’ original definition of personal probability is helpful in thinking about how to quantify uncertainty. Pretending that consequences are monetary (for instance, in US$), to say that \( P_{B,you} (E) = p \) for some uncertain outcome \( E \) whose truth
value will be known in the future is to say that you are indifferent between (a) receiving $p \cdot m$ for sure (for some hypothetical amount of money $m$) and (b) betting with someone in such a way that you will get $m$ if $E$ turns out to be true and $0$ if not.

- Any coherent set of probability judgments must satisfy the standard axioms and theorems of a finitely additive probability measure:

  - $0 \leq P(E) \leq 1$ and $P(E^c) = 1 - P(E)$;
  - $P(E_1 \text{ or } \ldots \text{ or } E_J) = \sum_{j \in J} P(E_j)$ for any finite collection $\{E_j, j \in J\}$ of disjoint outcomes;
  - $P(E \text{ and } F) = P(E) \cdot P(F)$ for any two independent outcomes (informally, $E$ and $F$ are independent if your uncertainty judgments involving one of them are unaffected by information about the other); and
  - Conditional probability has a natural definition in this setup, corresponding to the updating of your uncertainty about $E$ in light of $F$, and with this definition $P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$.

Otherwise (de Finetti, 1937/1980) someone betting with you on the basis of your probability judgments can make Dutch book against you, which is to say this person can get you to agree to a series of bets that are guaranteed to lose you money. Thus coherent Bayesian probability obeys the same laws as with the classical and frequentist approaches.

- Nothing so far has said clearly what choice to make in a decision problem if you wish to avoid incoherence. If the outcomes were certain you would evidently choose the action that maximizes your utility function, but since they are not the best action must involve a weighing both of your probabilities for the uncertain outcomes and the utilities you place on their consequences. It is a direct implication of the framework here that the form this weighing should take is simple and clear:

  \textbf{Maximization of expected utility (MEU):} Given your utility and probability judgments, your decision-making is coherent if for each action $a_i$, with associated uncertain outcomes $\{E_j, j \in J\}$ and consequences $\{c_j, j \in J\}$, you compute the expected utility $EU_i = \sum_{j \in J} U(c_j)P(E_j)$ and choose the action that maximizes $\{EU_i, i \in I\}$.

This is the basis of rational choice theory in economics (e.g.,
von Neumann and Morgenstern, 1944). It has been shown (ref. 19xx) if anybody can supply one of these references for me, I would be grateful] that in practice people sometimes act roughly like expected utility maximizers and sometimes they do not. Economists have a simple way out of this: utility is very hard to measure accurately, maybe there is nothing wrong with the theory, we just got their utility functions wrong. Or maybe the theory is incomplete: I recall an interesting talk given at Rand in 1989 by Howard Raiffa, one of the leaders of his generation in Bayesian decision theory, in which he was asked if he followed MEU in his own personal decision-making. He said, “ Heck, no, the choices my wife and I were making [about which jobs to take, where to live, and so on] were far too important to leave to MEU.” (!) He also said, though (and this accords with my own experiences), that he found laying out the ingredients of an MEU calculation—the possible actions, the values you would give to possible consequences, some rough idea of the relative likelihood of the uncertain outcomes—to be invaluable in making personal choices.

1.17 Computer scientists have terminology for an aspect of a computer program that some people regard as undesirable and others think is good: the former call it a bug, the latter a feature. For non-Bayesians having to specify a prior is a bug; for Bayesians it’s a feature.

1.18 The idea of conjugacy is at its most general in the exponential family of parametric probability distributions:

Definition (e.g., Bernardo and Smith, 1994): Given data $y = (y_1, \ldots, y_n)$ and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the sampling distribution $p(y|\theta)$ belongs to the $k$-dimensional exponential family if it can be expressed in the form

$$p(y|\theta) = c f(y) g(\theta) \exp \left[ \sum_{i=1}^{k} \phi_i(\theta) h_i(y) \right]. \quad (1.47)$$

In this case $\{\sum_{i=1}^{n} h_1(y_i), \ldots, \sum_{i=1}^{n} h_k(y_i)\}$ is the set of sufficient statistics for $\theta$ under $p(y|\theta)$.

As noted less formally in Section 1.6, $\{h_1, \ldots, h_k\}$ is sufficient
for $\theta$ under $p(y|\theta)$ if the likelihood $l(\theta|y)$ depends on $y$ only through the values of $\{h_1, \ldots, h_k\}$.

I bring up the exponential family because, if the likelihood $l(\theta|y)$ is of the form (1.47), then in searching for a conjugate prior $p(\theta)$—that is, a prior of the same functional form as the likelihood—you can see directly what will work:

$$p(\theta) = c \, g(\theta)^{\tau_0} \exp \left[ \sum_{i=1}^{k} \phi_i(\theta) \tau_i \right], \quad (1.48)$$

for some $\tau = (\tau_0, \ldots, \tau_k)$. With this choice the posterior for $\theta$ will be

$$p(\theta|y) = c \, g(\theta)^{1+\tau_0} \exp \left\{ \sum_{i=1}^{k} \phi_i(\theta) [h_i(y) + \tau_i] \right\}, \quad (1.49)$$

which is indeed of the same form (in $\theta$) as (1.48).

As a first example, with $S = \sum_{i=1}^{n} y_i$, the Bernoulli/binomial likelihood in (1.18) can be written

$$l(\theta|y) = \theta^S (1-\theta)^{n-S}$$

$$= (1-\theta)^n \left( \frac{\theta}{1-\theta} \right)^S$$

$$= (1-\theta)^n \exp \left[ S \log \left( \frac{\theta}{1-\theta} \right) \right], \quad (1.51)$$

which shows (a) that this sampling distribution is a member of the exponential family with $k = 1$, $g(\theta) = (1-\theta)^n$, $\phi_1(\theta) = \log \left( \frac{\theta}{1-\theta} \right)$ (NB the basis of logistic regression), and $h_1(y_i) = y_i$, and (b) that $h_1(y) = \sum_{i=1}^{n} h_1(y_i) = S$ is sufficient for $\theta$. Then (1.48) says that the conjugate prior for the Bernoulli/binomial is

$$p(\theta) = c (1-\theta)^{\alpha \tau_0} \exp \left[ \tau_1 \log \left( \frac{\theta}{1-\theta} \right) \right]$$

$$= c \, \theta^{\alpha - 1} (1-\theta)^{\beta - 1} = \text{Be}(\alpha, \beta) \quad (1.52)$$

for some $\alpha$ and $\beta$, as it should be.

For an example with $p > 1$, take $\theta = (\mu, \sigma^2)$ with the Gaussian likelihood:
\[ l(\theta|y) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ \frac{-1}{2\sigma^2} (y_i - \mu)^2 \right] \]
\[ = \sigma^{-n} (2\pi)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right] \]

This is of the form (1.47) with \( k = 2, \ c = (2\pi)^{-\frac{n}{2}}, \ f(y) = 1, \ g(\theta) = \sigma^{-n} \exp \left( -\frac{n\mu^2}{2\sigma^2} \right), \phi_1(\theta) = -\frac{1}{2\sigma^2}, \phi_2(\theta) = \frac{\mu}{\sigma^2}, \ h_1(y_i) = y_i^2 \), and \( h_2(y_i) = y_i \), which shows that \([h_1(y) = \sum_{i=1}^{n} y_i^2, h_2(y) = \sum_{i=1}^{n} y_i]\) or equivalently \((\bar{y}, \bar{y}^2)\) is sufficient for \(\theta\). Looking ahead a bit in the text, some very unpleasant algebra then demonstrates that (1.34) is conjugate for the Gaussian likelihood when both \(\mu\) and \(\sigma^2\) are unknown.

1.19 The Be(\(\alpha, \beta\)) distribution converges to the Gaussian as \(\alpha + \beta \to \infty\).

1.20 Named in honor of a letter sent by Oliver Cromwell to the elders of the Church of Scotland in 1xxx, at a moment in history when the Church leaders had already firmly made up their minds about [Dennis: please help me with details here, and a reference (Lindley, 19xx) where Cromwell’s Rule is stated]; Cromwell wrote, “I beseech you, in the bowels of Christ, think it possible that you may be wrong.”

1.21 The situation with a sharp null like \(H_0: \mu = 405.25\) is less pleasant: for Bayesians to make sense of such a hypothesis, there must be a blob of probability exactly at 405.25 in the prior, making both the prior and posterior a funny mixture of discrete and continuous distributions. For example, somebody who bought into this framework might construct a prior by putting probability 0.4 precisely on \(\mu = 405.25\) and spreading the other 0.6 out with a normal distribution scaled to integrate to 0.6. In practice your uncertainty about parameters like \(\mu\) is typically considerably smoother than that, which would seem to call into question the whole enterprise of testing sharp nulls (but see Problem 1.5 for a counterexample). In general I find it better to pass right by the entire enterprise of hypothesis testing in fa-
vor of more informative posterior summaries such as (say) 90% central intervals.

1.22 It spoils a good story, but in fairness I have to report that on my 79th trip to the machine I got a free cup of "tea."

1.23 Carl Morris regards the two parts of the double expectation theorem as so important for applied statisticians that he refers to them as "Adam and Eve."
References


REFERENCES


REFERENCES


REFERENCES


Stander J, Silverman BW (1994). [xx supply this]


