BAYESIAN ANALYSIS FOR THE SOCIAL SCIENCES
BAYESIAN ANALYSIS FOR THE SOCIAL SCIENCES

Simon Jackman
Stanford University

A JOHN WILEY & SONS, INC., PUBLICATION
To XXXXX
CONTENTS

List of Figures xi
List of Tables xiii
Foreword xv
Preface xvii
Acknowledgments xix
Acronymsxxi
Introduction xxiii

PART I INTRODUCING BAYESIAN STATISTICS

1 The Foundations of Bayesian Inference 1
  1.1 What is Probability? 2
    1.1.1 Probability in classical statistics 3
    1.1.2 Subjective Probability 4
  1.2 Subjective Probability in Bayesian Statistics 6
  1.3 Bayes’ Theorem, Discrete Case 8
  1.4 Bayes’ Theorem, Continuous Parameter 12
    1.4.1 Cromwell’s Rule 18
    1.4.2 Bayesian Updating as Information Accumulation 21
  1.5 Parameters as Random Variables, Beliefs as Distributions 23
CONTENTS

1.6 Communicating the Results of a Bayesian Analysis 24
  1.6.1 Bayesian Point Estimation 25
  1.6.2 Credible Regions 29
1.7 Asymptotic Properties of Posterior Distributions 36
1.8 Bayesian Hypothesis Testing 39
  1.8.1 Model Choice 43
  1.8.2 Model Averaging 45
  1.8.3 Bayes Factors 45
1.9 From Subjective Beliefs to Parameters and Models 46
  1.9.1 Exchangeability 47
  1.9.2 Implications and Extensions of de Finetti’s Representation Theorem 50
  1.9.3 Finite Exchangeability 51
  1.9.4 Exchangeability and Prediction 52
  1.9.5 Conditional Exchangeability and Multiparameter Models 53
  1.9.6 Exchangeability of Parameters: Hierarchical Modeling 54
1.10 Historical Note 55
1.11 Summary 56
Problems 56

2 Getting Started: Bayesian Analysis for Simple Models 57
  2.1 Learning about Probabilities, Rates and Proportions 58
    2.1.1 Priors for Probabilities, Rates and Proportions 59
    2.1.2 Bayes Estimates as Convex Combinations of Priors and Data 69
    2.1.3 Parameterizations and Priors 72
    2.1.4 The Variance of the Posterior Density 75
  2.2 Associations Between Binary Variables 79
  2.3 Learning about a Normal Mean 89
    2.3.1 Variance Known 89
    2.3.2 Mean and Variance Unknown 92
    2.3.3 An Improper, Reference Prior 95
  2.4 Learning about a Correlation 96
  2.5 Learning about Parameters in a Regression Model 96
  2.6 Summary 96
Problems 96

3 Why be Bayesian? 99

PART II SIMULATION BASED BAYESIAN ANALYSIS

4 Bayesian Computation via Simulation Methods 103

5 Bayesian Analysis of Extended Regression Models 105
6 Combining Information: Bayesian Hierarchical Modeling 107

PART III  ADVANCED APPLICATIONS IN THE SOCIAL SCIENCES

7 Bayesian Analysis of Choice Making 111

8 Bayesian Approaches to Measurement 113

9 Bayesian Analysis of Experiments 115

Appendix A: Probability Review 117

Appendix B: Proofs of Selected Propositions 123

References
LIST OF FIGURES

1.1 Posterior Probability of Class Membership, Congressional Districts. 13
1.2 Priors, Likelihoods and Posterior Densities 16
1.3 Priors, Likelihoods and Posterior Densities for Non-Standard Cases 17
1.4 Discontinuous Priors and Posteriors 19
1.5 Standard normal distribution and 95% HPD interval 31
1.6 $\chi^2_4$ distribution, with 50% HPD interval 32
1.7 Bimodal Posterior Density for a Correlation Coefficient 34
1.8 Uniform Distribution and 25% HPDs 35
1.9 Sequence of Posterior Densities (1) 37
1.10 Sequence of Posterior Densities (2) 38
1.11 Posterior Distribution and Sampling Distribution, for Example 1.9. 42
2.1 Florida Election Polling 66
2.2 Posterior Mean as a Convex Combination of Prior and Data
2.3 The Beta(\(\frac{1}{2}, \frac{1}{2}\)) density.
2.4 Prior and Posterior Densities for Example 2.6
2.5 Posterior Densities for Example 2.8
2.6 Error in Using a Normal Approximation to a Beta Density
LIST OF TABLES

1.1. Twelve Observations from a Bivariate Normal Distribution. 33
2.1. Two-by-Two Cross Table 81
2.2. Young Adults Sexual Identity, by Sexual Orientation of Mother 84
FOREWORD

This is the foreword to the book.
PREFACE

This is an example preface. This is an example preface. This is an example preface. This is an example preface.

This is a preface section

This is an example of a preface. This is an example preface. This is an example preface. This is an example preface. This is an example preface. This is an example preface.

XXXX, X. X.

XXX, XXX

September, 200X
ACKNOWLEDGMENTS
Generally, “Bayesian analysis” refers to the statistical analysis of data that relies on Bayes’ Theorem, presented below. Bayes’ Theorem tells us how to update prior beliefs about parameters or hypotheses in light of data, to yield posterior beliefs. Or, even more simply, Bayes’ Theorem tells us how to learn rationally about parameters from data. As we shall see, Bayesian analysis is often more easily said than done, or at least this was the case up until recently. In the 1990s there was a veritable explosion of interest in Bayesian analysis in the statistics profession, that has now crossed over into quantitative social science. The mathematics and computation underlying Bayesian analysis has been dramatically simplified via a suite of algorithms known collectively as Markov chain Monte Carlo (MCMC), to be discussed in Chapter 4. The combination of the popularization of MCMC and vast increases in the computing power available to social scientists means that Bayesian analysis is now well and truly part of the mainstream of quantitative social science, as we will see in some detail in later chapters.

XXXX
PART I

INTRODUCING BAYESIAN STATISTICS
In this chapter I provide an introduction to Bayesian statistical inference. I begin by reviewing the fundamental role of probability in statistical inference. In the Bayesian approach, probability is usually interpreted in subjective terms, as a formal, mathematically rigorous characterization of beliefs. I contrast the subjective notion of probability from the classical, objective or frequentist approach, before stating Bayes Theorem in various forms of use in statistical settings. I then review, in quite general terms, how Bayesian data analyses proceed. At a high level of abstraction, Bayesian data analysis is extremely simple, following the same, basic recipe: via Bayes Rule, we use the data to update prior beliefs. Of course, there is much to be said on the implementation of this procedure in any specific application. In this chapter I deal with some general issues. For instance, how does Bayesian inference differ from classical inference? Where do priors come from? What is the result of a Bayesian analysis, and how does one report those results?
1.1 WHAT IS PROBABILITY?

As a formal, mathematical matter, the question “what is probability?” is utterly uncontroversial. The following axioms, known as the Kolmogorov (1933) axioms, constitute the conventional, modern, mathematical definition of probability, which I reproduce here (with measure-theoretic details omitted; see the Appendix for a more rigorous set of definitions). If $\Omega$ is a set of events, and $P(A)$ is a function that assigns real numbers to events $A \subset \Omega$, then $P(A)$ is a probability measure if

1. $P(A) \geq 0, \forall A \subset \Omega$ (probabilities are non-negative)
2. $P(\Omega) = 1$ (probabilities sum to one)
3. If $A$ and $B$ are disjoint events, then $P(A \cup B) = P(A) + P(B)$ (the joint probability of disjoint events is the sum of the probabilities of the events).

On these axioms rests virtually all of contemporary statistics, including Bayesian statistics. This said, one of the ways in which Bayesian statistics differs from classical statistics is in the interpretation of probability. The very idea that probability is a concept open to interpretation might strike you as odd. Indeed, Kolmogorov himself ruled out any questions regarding the interpretation of probabilities:

The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations (Kolmogorov 1956, 1).

Nonetheless, for anyone actually deploying probability in a real-world application, Kolmogorov’s insistence on a content-free definition of probability is quite unhelpful. As Leamer (1978, 24) points out:

These axioms apply in many circumstances in which no one would use the word probability. For example, your arm may contain 10 percent of the weight of your body, but it is unlikely that you would report that the probability of your arm is .1.

Thus, for better or for worse, probability is open to interpretation, and has been for a long time. Differences in interpretation continue to be controversial (although less so now than, say, 30 years ago), are critical to the distinction between Bayesian and non-Bayesian statistics, and so no book-length treatment of Bayesian statistics can ignore it.
Most thorough, historical treatments of probability point to at least four interpretations of probability. For our purposes, the most important distinction is between probability as it was probably (!) taught to you in your first statistics class, and probability as interpreted by most Bayesian statisticians.

1.1.1 Probability in classical statistics

Classical statistics usually interprets probability as a property of the phenomenon being studied: for instance, the probability that a tossed coin will come up heads is a characteristic of the coin. With enough data, say, by tossing the coin many times under more or less identical conditions, and noting the result of each toss, we can estimate the probability of a head, with the precision of the estimate monotonically increasing with the number of tosses. In this view, probability is the limit of a long-run, relative frequency; i.e., if \( A \) is an event of interest (e.g., the coin lands heads up) then

\[
\Pr(A) = \lim_{n \to \infty} \frac{m}{n}
\]

is the probability of \( A \), where \( m \) is the number of times we observe the event \( A \) and \( n \) is the number of events. Given this definition of probability, we can understand why classical statistics is sometimes referred to as

1. frequentist, in the sense that it rests on a definition of probability as the long-run relative frequency of an event;

2. objectivist, in the sense that probabilities are characteristics of objects or things (e.g., the staples of introductory statistics, such as cards, dice, coins, roulette wheels); this position will be contrasted with a subjectivist interpretation of probability.

The fact that “games of chance” are used to introduce the classical conception of probability is no accident. Although games of chance have been with us for millenia, the first recognizably frequentist treatment of probability is widely believed to be that the Italian mathematician Gerolamo Cardano, dating to around 1526. Cardano enumerated the chances associated with the outcomes arising from multiple casts of a six-sided die, under the assumption that the die is fair. Cardano referred to the “chances” of a particular outcome in frequentist terms: i.e., how often we would observe a particular event divided by the total number of outcomes. Later, famous names such as Galileo, Pascal, Fermat and Huygens all solved gambling problems (e.g., David 1955), helping to shape the view that games of chance were the motivating, empirical problem behind probability theory, at least up until
the late seventeenth century. As Leamer (1978, 24) points out, Cardano’s formulation may not have been intended as a definition of probability (for one thing, Cardano’s discussion of the outcomes of games of chance was predicated on the assumption that outcomes were equally likely), but only as a device for computing “chances”. Nonetheless, de Moivre and Laplace adopted the definition, and it found widespread use and acceptance. One of the strongest statements of the frequentist position comes from Richard von Mises:

...we may say at once that, up to the present time [1928], no one has succeeded in developing a complete theory of probability without, sooner or later, introducing probability by means of the relative frequencies in long sequences.

Further,

The rational concept of probability, which is the only basis of probability calculus, applies only to problems in which either the same event repeats itself again and again, or a great number of uniform elements are involved at the same time... [In] order to apply the theory of probability we must have a practically unlimited sequence of observations.

As we shall see, alternative views long pre-date von Mises’ 1928 statement and it is indeed possible to apply the theory of probability without a “practically unlimited” sequence of observations. This is just as well, since many statistical analyses in the social-sciences are conducted without von Mises’ “practically unlimited” sequence of observations.

### 1.1.2 Subjective Probability

Most introductions to statistics are replete with the examples from games of chance, and the naïve view of the history of statistics is that games of chance spurned the development of probability (e.g., Todhunter 1865), and the frequentist interpretation. But historians of science stress that at least two notions of probability were under development from the late 1600s onwards: the objectivist view described above, and a subjectivist view. According to Ian Hacking, the former is “statistical, concerning itself with stochastic laws of chance processes”, while the other notion is “epistemological, dedicated to assessing reasonable degrees of belief in propositions quite devoid of statistical background” (Hacking 1975, 12).

As an example of the latter, consider Locke’s Essay Concerning Human Understanding (1698). Book IV, Chapter XV of the Essay is titled “On Probability”, in which Locke notes that “...most of the propositions we think, reason, discourse — nay, act upon, are such that we cannot have undoubted knowledge of their truth.” Moreover, there are “degrees” of belief, “...from the very neighbourhood of certainty and demonstration, quite down
WHAT IS PROBABILITY?

“For Locke, “Probability is likeliness to be true,” a definition in which (repeated) games of chance play no part. The idea that one might hold different degrees of belief over different propositions has a long lineage, and was apparent in the theory of proof in Roman and canon law, in which judges were directed to employ an “arithmetic of proof”, assigning different weights to various pieces of evidence, and to draw distinctions between “complete proofs” or “half proofs” (Daston 1988, 42-43). Scholars became interested in the possibility of making these notions more rigorous, with Leibniz perhaps the first to make the connection between the qualitative use of probabilistic reasoning in jurisprudence with the mathematical treatments being generated by Pascal, Huygens, and others. Perhaps the most important and clearest statement linking this form of jurisprudential “reasoning under uncertainty” to “probability” is Jakob Bernoulli’s posthumous Ars conjectandi (1713). In addition to developing the theorem now known as the weak law of large numbers, in Part IV of the Ars conjectandi Bernoulli declares that “Probability is degree of certainty and differs from absolute certainty as the part differs from the whole”, it being unequivocal that the “certainty” referred to is a state of mind, but, critically, (1) varied from person to person (depending on one’s knowledge and experience) and (2) was quantifiable. For example, for Bernoulli, a probability of 1.0 was an absolute certainty, a “moral certainty” was nearly equal to the whole certainty (e.g., 999/1000, and so a morally impossible event has only 1 - 999/1000 = 1/1000 certainty), and so on, with events having “very little part of certainty” still nonetheless being possible.

In the early-to-mid twentieth century, the competition between the frequentist and subjectivist interpretations intensified, in no small measure reflecting the competition between Bayesian statistics and the then newer, frequentist statistics being championed by R. A. Fisher. Venn (1866) and later von Mises (1957) made a strong case for a frequentist approach, apparently in reaction to “...a growing preoccupation with subjective views of probability” (Barnett 1999, 76). During this period, both the objective/frequentist and subjective interpretations of probability were formalized in modern, mathematical terms — von Mises formalizing the frequentist approach, and Ramsey (1931) and de Finetti (1974, 1975) providing the formal links between subjective probability and decisions and actions. Ramsey and de Finetti, working independently, showed that subjective probability was, properly, not just any set of subjective beliefs, but beliefs that conformed to the axioms of probability. The Ramsey-de Finetti Theorem states that given \( p_1, p_2, \ldots \), a set of betting quotients on hypotheses \( h_1, h_2, \ldots \), if the \( p_j \) do not satisfy the probability axioms, then there is a betting strategy and a set of stakes such that whoever follows this betting strategy...”
will lose a finite sum whatever the truth values of the hypotheses turn out to be (Howson and Urbach 1993, 79); this theorem is also known as the Dutch Book Theorem, a Dutch book being a bet (or a series of bets) in which the bettor is guaranteed to lose. In de Finetti’s terminology, subjective probabilities that fail to conform to the axioms of probability are incoherent or inconsistent. Thus, subjective probabilities are whatever a particular person believes, provided they satisfy the axioms of probability. In particular, the Dutch book results extend to the case of conditional probabilities, meaning that if I do not update my subjective beliefs in light of data in a manner consistent with the probability axioms, and you can convince me to gamble with you, you have the opportunity to take advantage of my irrationality, and are guaranteed to profit at my expense. That is, while probability may be subjective, Bayes Rule governs how rational people should update subjective beliefs.

1.2 SUBJECTIVE PROBABILITY IN BAYESIAN STATISTICS

Of course, it should come as no surprise that the subjectivist view is almost exclusively adopted by Bayesians. To see this, recall the proverbial coin tossing experiment of introductory statistics. And further, recall the goal of Bayesian statistics: to update probabilities in light of evidence, via Bayes’ Theorem. But which probabilities? The objective sense (probability as a characteristic of the coin) or the subjective sense (probability as degree of belief)? Well, almost surely we do not mean that the coin is changing; it is conceivable that the act of flipping and observing the coin is changing the tendency of the coin to come up heads when tossed, but unless we are particularly violent coin-tossers this kind of physical transformation of the coin is of an infinitesimal magnitude. Indeed, if this occurred then both frequentist and Bayesian inference gets complicated (multiple coin flips no longer constitute an independent and identically distributed sequence of random events). No, the probability being updated here can only be a subjective probability, the observer’s degree of belief about the coin.

Bayesian probability statements are thus about states of mind over states of the world, and not about states of the world per sé. Indeed, whatever one believes about determinism or chance in social processes, the meaningful uncertainty is that which resides in our brains, upon which we will base decisions and actions. Again, consider tossing a coin. As Emile Borel apparently remarked to de Finetti, one can guess the outcome of the toss while the coin is still in the air and its movement is perfectly determined, or even after the coin has landed but before one reviews the result; that is, subjective uncertainty obtains irrespective
of “objective uncertainty (however conceived)” (de Finetti 1980b, 201). Indeed, in one of the more memorable and strongest statements of the subjectivist position, de Finetti writes

PROBABILITY DOES NOT EXIST

The abandonment of superstitious beliefs about...Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is not less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs. In investigating the reasonableness of our own modes of thought and behaviour under uncertainty, all we require, and all that we are reasonably entitled to, is consistency among these beliefs, and their reasonable relation to any kind of relevant objective data (“relevant” in as much as subjectively deemed to be so). This is Probability Theory (de Finetti 1974, 1975, x).

The use of subjective probability also means that Bayesians can report probabilities without an “practically unlimited” sequence of observations. For instance, a subjectivist can attach probabilities to the proposition “Andrew Jackson was the eighth president of the United States” (e.g., Leamer 1978, 25), reflecting his or her degree of belief in the proposition. Contrast the frequentist position, in which probability is defined as the limit of a relative frequency. Then, what, precisely, is the frequentist probability of the truth of the proposition “Jackson was the eighth president”? Since there is only one relevant experiment for this problem, the frequentist probability is either zero (if Jackson was the eighth president) or one (if Jackson was not the eighth president). Non-trivial frequentist probabilities, it seems, are reserved for phenomena that are standardized and repeatable (e.g., the standards of introductory statistics such as coin tossing and cards, or, perhaps, random sampling in survey research). Even greater difficulties for the frequentist position arise when considering events that have not yet occurred, e.g.,

- What is the probability that the Democrats win a majority of seats in the House of Representatives at the next Congressional elections?
- What is the probability of a terrorist attack in the next five years?
- What is the probability that someone I know will be incarcerated at some stage in their life?

All of these are perfectly legitimate and interesting social-scientific questions, but for which the objectivist/frequentist position apparently offers no clear answer. As Leamer (1978, 26)
notes, one frequentist response is to assert that probabilities are not defined for individual
events, but rather, that probabilities are objective properties of classes of events. If so,
then it would seem that either (1) frequentist probability statements are undefined for many
events of interest to social scientists; (2) we need to define the class of event to which our
particular event belongs, which sounds quite subjective. We will revisit this issues when
we consider statistical inference for non-repeatable social-science data in Chapter 3.

1.3 BAYES’ THEOREM, DISCRETE CASE

Bayes’ Theorem itself is uncontroversial: it is merely an accounting identity that follows
from the axioms of probability discussed above, plus the following additional definition:

**Definition 1.1** (Conditional Probability). Let A and B be events with $P(B) > 0$. Then the
conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}.$$  

Although conditional probability is presented here (and in most sources) merely as a
definition, it need not be. de Finetti (1980a) shows how coherence requires that conditional
probabilities behave as given in definition 1.1, first published in 1937. The thought experi-
ment is as follows: consider selling a bet at price $P(A) \cdot S$, that pays $S$ if event $A$ occurs, but
is annulled if event $B$ does not occur, with $A \subseteq B$. Then unless your conditional probability
$P(A|B)$ conforms to the definition above, someone could collect arbitrarily large winnings
from you via their choice of the stakes $S$; Leamer (1978, 39-40) provides a simple re-telling
of de Finetti’s argument.

Conditional probability is derived from more elementary axioms (rather than presented
as a definition) in the work of Bernardo and Smith (1994, ch 2). Some authors work with a
set of probability axioms that are explicitly conditional, consistent with the notion that there
are no such things as unconditional beliefs over parameters; e.g., Press (2003, ch2) adopts
the conditional axiomization of probability due to Rényi (1970) and see also the treatment
in Lee (2004, ch1).

The following two useful results are also implied by the probability axioms, plus the
definition of conditional probability:

**Proposition 1.1** (Multiplication Rule).

$$P(A \cap B) = P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$
**Proposition 1.2** (Law of Total Probability).

\[
P(B) = P(A \cap B) + P(\sim A \cap B)
\]

\[
= P(B|A)P(A) + P(B\sim)P(\sim A)
\]

Bayes’ Theorem can now be stated, following immediately from the definition of conditional probability:

**Proposition 1.3** (Bayes’ Theorem). If \(A\) and \(B\) are events with \(P(B) > 0\), then

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

*Proof.* By proposition 1.1 \(P(A, B) = P(B|A)P(A)\). Substitute into the definition of the conditional probability of \(P(A|B)\) given in definition 1.1. \(\square\)

Bayes’ Theorem is much more than an interesting result from probability theory, as the following re-statement makes clear. Let \(H\) denote a hypothesis and \(E\) evidence (data), then we have

\[
Pr(H|E) = \frac{Pr(E \cap H)}{Pr(E)} = \frac{Pr(E|H)Pr(H)}{Pr(E)}
\]

provided \(Pr(E) > 0\). In this version of Bayes’ Theorem, \(Pr(H|E)\) is the probability of belief in \(H\) after obtaining \(E\), and \(Pr(H)\) is the prior probability of \(H\) before considering \(E\). The conditional probability on the left-hand side of the theorem, \(Pr(H|E)\), is usually referred to as the posterior probability of \(H\). Bayes’ Theorem thus supplies a solution to the general problem of inference or induction (e.g., Hacking 2001), providing a mechanism for learning about the plausibility of a hypothesis \(H\) from data \(E\).

In this vein, Bayes’ Theorem is sometimes referred to as the rule of inverse probability, since it shows how a conditional probability \(B\) given \(A\) can be “inverted” to yield the conditional probability \(A\) given \(B\). This usage dates back to Laplace (e.g., see Stigler 1986b), and remained current up until the popularization of frequentist methods in the early twentieth century — and, importantly, criticism of the Bayesian approach by R. A. Fisher (Zabell 1989a).

I now state another version of Bayes’ Theorem, that is actually more typical of the way the result is applied in social-science settings.

**Proposition 1.4** (Bayes’ Theorem, Multiple Discrete Events). Let \(H_1, H_2, \ldots, H_k\) be mutually exclusive and exhaustive hypotheses, with \(P(H_j) > 0 \forall j = 1, \ldots, k\), and let \(E\) be evidence with \(P(E) > 0\). Then, for \(i = 1, \ldots, k\),

\[
P(H_i|E) = \frac{P(H_i)P(E|H_i)}{\sum_{j=1}^{k} P(H_j)P(E|H_j)}.
\]
Proof. Using the definition of conditional probability, \( P(H_i|E) = P(H_i, E)/P(E) \). But, again using the definition of conditional probability, \( P(H_i, E) = P(H_i)P(E|H_i) \). Similarly, \( P(E) = \sum_{j=1}^{k} P(H_j)P(E|H_j) \), by the law of total probability (proposition 1.2).

\[ \Box \]

**EXAMPLE 1.1**

**Drug Testing.** Elite athletes are routinely tested for the presence of banned performance-enhancing drugs. Suppose one such test has a false negative rate of .05 and a false positive rate of .10. Prior work suggests that about 3% of the subject pool uses a particular prohibited drug. Let \( H_U \) denote the hypothesis “the subject uses the prohibited substance”; let \( H_{\sim U} \) denote the contrary hypothesis. Suppose a subject is drawn randomly from the subject pool for testing, and returns a positive test, and denote this event as \( E \). What is the posterior probability that the subject uses the substance? Via Bayes’ Theorem in Proposition 1.4,

\[
P(H_U|E) = \frac{P(H_U)P(E|H_U)}{\sum_{i \in \{U, \sim U\}} P(H_i)P(E|H_i)} = \frac{.03 \times .95}{(.03 \times .95) + (.97 \times .10)} = \frac{.0285}{.23} = .23
\]

That is, in light of (1) the positive test result \( E \), (2) what is known about the sensitivity of the test, \( P(E|H_U) \), and (3) the specificity of the test, \( 1 - P(E|H_{\sim U}) \), we revise our beliefs about the probability that the subject is using the prohibited substance from the baseline or prior belief of \( P(H_U) = .03 \) to \( P(H_U|E) = .23 \). Note that this posterior probability is still substantially below .5, the point at which we would say it is more likely than not that the subject is using the prohibited substance.

**EXAMPLE 1.2**

**Classifying Congressional Districts.** The United States House of Representatives consists of 435 congressional districts, which are often classified as follows: safe Republican seats \( (T_i = 1) \), competitive seats \( (T_i = 2) \), and safe Democratic seats \( (T_i = 3) \). Let \( y_i \) be the proportion of the two-party vote won by the Democratic candidate in district \( i \), and \( \alpha_j \) be the proportion of districts in class \( j \). For purposes of exposition, assume that the distribution of the \( y_i \) within each of the \( j \in \{1, 2, 3\} \)
classes is well approximated by a normal distribution, i.e., $y_i | (T_i = j) \sim N(\mu_j, \sigma_j^2)$. Analysis of data from the 2000 U.S. Congressional elections ($n = 371$ contested districts) suggests the following values for $\mu_j$ and $\sigma_j$:

<table>
<thead>
<tr>
<th>Class</th>
<th>$\mu_j$</th>
<th>$\sigma_j$</th>
<th>$\alpha_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Safe Republican</td>
<td>.34</td>
<td>.07</td>
<td>.45</td>
</tr>
<tr>
<td>2. Competitive Seats</td>
<td>.50</td>
<td>.03</td>
<td>.02</td>
</tr>
<tr>
<td>3. Safe Democratic</td>
<td>.67</td>
<td>.13</td>
<td>.52</td>
</tr>
</tbody>
</table>

By Bayes’ Theorem (as stated in Proposition 1.4), the probability that district $i$ belongs to class $j$ is

$$P(T_i = j | y_i) = \frac{P(T_i = j) P(y_i | T_i = j)}{\sum_{k=1}^{K} [P(T_i = k) P(y_i | T_i = k)]} = \frac{\alpha_j \cdot \phi(y_i; \mu_j, \sigma_j)}{\sum_{k=1}^{K} [\alpha_k \cdot \phi(y_i; \mu_k, \sigma_k)]}$$

(1.1)

where $\phi(y; \mu, \sigma)$ is the normal pdf, i.e.,

$$\phi(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - \mu)^2}{2\sigma^2}\right).$$

In 2000, California’s 15th congressional district was largely comprised of Silicon Valley suburbs, at the southern end of the San Francisco Bay Area, and some of the wealthy, neighboring suburban communities running up into the Santa Cruz mountains. The incumbent, Republican Tom Campbell, had been re-elected in 1998 with over 61% of the two-party vote, but vacated the seat in order to run for the U.S. Senate: according to the *Almanac of American Politics* (Barone, Cohen and Ujifusa 2002, 198),

the authorities at Stanford Law School had told him [Campbell] he would lose tenure if he stayed in Congress, so instead of winning another term in the House as he could easily have done, he decided to gamble and win either the Senate or Stanford. Predictably, Stanford won.

In the parlance of American politics, CA 15 was an “open seat” in 2000. An interesting question is the extent to which Campbell’s incumbency advantage had been depressing Democratic vote share. With no incumbent contesting the seat in 2000, it
is arguable that the 2000 election would provide a better gauge of the district’s type. The Democratic candidate, Mike Honda, won with 56% of the two-party vote. So, given that \( y_i = .56 \), to which class of congressional district should we assign CA 15? An answer is given by substituting the estimates given in the above table into the version of Bayes’ Theorem given in equation 1.1:

\[
P(T_i = 1|y_i = .56) = \frac{.45 \times .042}{.45 \times .042 + .02 \times .90 + .52 \times 2.10} = .019 \\
P(T_i = 2|y_i = .56) = \frac{.02 \times .90}{1.15} = .017 \\
P(T_i = 3|y_i = .56) = \frac{.52 \times 2.10}{1.15} = .97
\]

i.e., the most probable outcome is that CA 15 is in the safe Democratic class of Congressional district, and in fact, this event is overwhelmingly more likely than the other two outcomes.

This calculation can be repeated for any plausible value of \( y_i \), and hence over any range of plausible values for \( y_i \), showing how posterior classification probabilities change as a function of \( y_i \). Figure 1.1 presents a graph of the posterior probability of membership in each of three classes of congressional district, as Democratic congressional vote share ranges over the values observed in the 2000 election. The graph makes clear that the two large classes of safe Republican and safe Democratic districts dominate the small class of competitive seats. Even with \( y_i = \mu_2 = .5 \), the probability that district \( i \) belongs to the competitive class is only .28. We will return to this example in Chapter XXXX.

1.4 BAYES’ THEOREM, CONTINUOUS PARAMETER

In most analyses in the social-sciences, we want to learn about a continuous parameter, rather than the discrete parameters considered in the discussion thus far. Examples include the mean of a continuous variable, a proportion (a continuous parameter on the unit interval), a correlation, or a regression coefficient. In general, let the unknown parameter be \( \theta \) and denote the data available for analysis as \( y = (y_1, \ldots, y_n)' \). In the case of continuous parameters, beliefs about the parameter are represented as probability density functions or pdfs (see definition ??); we denote the prior pdf as \( p(\theta) \) and the posterior pdf we denote as \( p(\theta|y) \).
Figure 1.1 Posterior Probability of Class Membership, Congressional Districts. Class 1 = safe Republican; Class 2 = competitive districts; Class 3 = safe Democratic.
Then, Bayes’ Theorem for a continuous parameter is as follows:

**Proposition 1.5** (Bayes’ Theorem, Continuous Parameter),

\[
p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}
\]

**Proof.** By the multiplication rule of probability (Proposition 1.1),

\[
p(\theta, y) = p(\theta|y)p(y) = p(y|\theta)p(\theta),
\]

where all these densities are assumed to exist and have the properties \( p(z) > 0 \) and \( \int p(z)dz = 1 \). The result follows by re-arranging the quantities in equation \( 1.2 \) and noting that \( p(y) = \int p(y, \theta)d\theta = \int p(y|\theta)p(\theta)d\theta \).

Bayes’ Theorem for continuous parameters is more commonly expressed as

\[
p(\theta|y) \propto p(y; \theta)p(\theta)
\]

where the constant of proportionality is

\[
\left[ \int p(y|\theta)p(\theta)d\theta \right]^{-1}
\]

i.e., ensuring that the posterior pdf integrates to one, and so is a proper pdf (again, see definition ??). The first term on the right hand side of equation \( 1.3 \) is the likelihood function (see Definition A.18). Thus, we can state this version of Bayes’ Theorem in words, providing the “Bayesian mantra”,

"the posterior is proportional to the prior times the likelihood."

This formulation highlights a particularly elegant feature of the Bayesian approach, showing how the likelihood function can be used to generate a probability statement about \( \theta \), given data \( y \).

Figure 1.2 shows the Bayesian mantra at work for a simple, single-parameter problem: the success probability, \( \theta \in [0, 1] \), underlying a binomial process, an example which we will return to in detail in Chapter 2. Each panel shows a combination of a prior, a likelihood, and a posterior distribution (with the likelihood re-normalized to be comparable to the prior and posterior densities).

The first two panels in the top row of Figure 1.2 have a uniform prior, \( \theta \sim \text{Unif}(0, 1) \), and so the prior is absorbed into the constant of proportionality, resulting in a posterior density over \( p \) that is proportional to the likelihood: given the normalization of the likelihood I use
in Figure 1.2, the posterior and the likelihood graphically coincide. In these cases, the mode of the posterior density is also that value of $p$ that maximizes the likelihood function. For the special case considered in Figure 1.2, the prior distribution $\theta \sim \text{Unif}(0, 1)$ corresponds to an uninformative prior over $\theta$, the kind of prior we might specify when we have no prior information about the value of $\theta$, and hence no way to a priori prefer one set of values for $\theta$ over any other. Of course, there is another way to interpret this result: from a Bayesian perspective, likelihood based analyses of data assume prior ignorance, although seldom is this assumption made explicit, even if it were plausible. In the examples we encounter in later chapters, we shall see circumstances in which prior ignorance is plausible, and cases in which it is not. We will also consider the priors that generate “the usual answer” for well-known problems (e.g., estimating a mean, a correlation, regression coefficients, etc).

The other panels in Figure 1.2 display how Bayesian inference works with more or less informative priors for $p$. In the top left of the Figure we see what happens when the prior and the likelihood more or less coincide. In this case, the likelihood is a little less diffuse than the prior, but the prior and the likelihood have the same mode. Application of Bayes Theorem in this instance yields a posterior distribution that has the same mode as the prior and the likelihood, but is more precise (less diffuse) than both the prior and the likelihood. In the other panels of Figure 1.2, this pattern is more or less repeated, except that the mode of the prior and the likelihood are not equal. In these cases, the mode of the posterior distribution lies between the mode of the prior distribution and the mode of the likelihood. Specifically, the mean of the posterior distribution is a precision-weighted average of the prior and the likelihood, a feature that we will see repeatedly in this book, a consequence of working with so-called conjugate priors in the exponential family. Many standard statistical models are in the exponential family (but not all), for which conjugate priors are convenient ways of mathematically representing prior beliefs over parameters. A definition of conjugacy will come in Chapter 2. For now, the important point is the one that appears in Figure 1.2: for a wide class of problems (i.e., when conjugacy holds), Bayesian statistical inference is equivalent to combining information, marrying the information in the prior with the information in the data, with the relative contributions of prior and data to the posterior being proportional to their respective precisions. When prior beliefs are “vague”, “diffuse”, or, in the limit, uninformative, the posterior distribution will be dominated by the likelihood (i.e., the data contains much more information than the prior about the parameters); e.g., the lower left panel of Figure 1.2. In the limiting case of an uninformative prior, the only information about the parameter is that in the data, and the
Figure 1.2 Priors, Likelihoods and Posterior Densities
When prior information is available, the posterior incorporates it, and rationally, in the sense of being consistent with the laws of probability via Bayes’ Theorem. In fact, when prior beliefs are quite precise relative to the data, it is possible that the likelihood is largely ignored, and the posterior distribution will look almost exactly like the prior, as it should in such a case; e.g., see the lower right panel of Figure 1.2. In the limiting case of a degenerate, infinitely-precise, “spike prior” (all prior probability concentrated on a point), the data are completely ignored, and the posterior is also a degenerate “spike” distribution. Should you hold such a dogmatic prior, no amount of data will ever result in you changing your mind about the issue.
Figure 1.3 displays a series of prior and posterior densities for less standard cases, where the prior distributions are not simple unimodal distributions. In each instance, Bayes rule applies as usual, with the posterior distribution being proportional to the prior distribution times the likelihood, and appropriately normalized such that the posterior distribution encloses an area equal to one. In the left-hand series of panels, the prior has a two modes, with the left mode more dominant than the right mode. The likelihood is substantially less dispersed than the prior, and attains a maximum at a point with low prior probability. The resulting posterior distribution clearly represents the merger of prior and likelihood: with a mode just to the left of the mode of the likelihood function, and a smaller mode just to the right of the mode of the likelihood function. The middle column of panels in Figure 1.3 shows a symmetric case: the prior is bimodal but symmetric around a trough corresponding to the mode of the likelihood function, resulting in a bimodal posterior distribution, but with modes shrunk towards the mode of the likelihood. In this case, the information in the data about $\theta$ combines with the prior information to reduce the depth of the trough in the prior density, and to give substantially less weight to the outlying values of $\theta$ that receive high prior probability. In the right-hand column of Figure 1.3 an extremely flamboyent prior distribution (but one that is nonetheless symmetric about its mean) combines with the skewed likelihood to produce the trimodal posterior distribution, with the posterior modes located in regions with relatively high likelihood. Although this prior (and posterior) are somewhat fanciful (in the sense that it is hard to imagine those distributions corresponding to beliefs over a parameter), the central idea remains the same: Bayes Rule governs the mapping from prior to posterior through the data. Implementing Bayes Rule may be difficult when the prior is not conjugate to the likelihood, but, as we shall see, this is where modern computational tools are particularly helpful.

1.4.1 Cromwell’s Rule

Note also that via Bayes Rule, if a particular region of the parameter space has zero prior probability, then it also has zero posterior probability. This feature of Bayesian updating has been dubbed “Cromwell’s Rule” by Lindley (1985). After the English deposed, tried and executed Charles I in 1649, the Scots invited Charles’ son, Charles II to become king. The English regarded this as a hostile act, and Oliver Cromwell led an army north. Prior to the outbreak of hostilities, Cromwell wrote to the synod of the Church of Scotland, “I beseech you, in the bowels of Christ, consider it possible that you are mistaken”. The relevance of Cromwell’s plea to the Scots for our purposes comes from noting that a prior that assigns
Figure 1.4 Discontinuous Priors and Posteriors
zero probability weight to a hypothesis can never be revised; likewise, a hypothesis with prior weight of 1.0 can never be refuted.

The operation of Cromwell’s Rule is particularly clear in the left-hand column of panels in Figure 1.4: the prior for $\theta$ is a uniform distribution over the left half of the support of the likelihood, and zero everywhere else. The resulting posterior assigns zero probability to values of $\theta$ assigned zero prior probability, and since the prior is uniform elsewhere, the posterior is a re-scaled version of the likelihood in this region of non-zero prior probability, where the re-scaling follows from the constraint that the area under the posterior distribution is one. The middle column of panels in Figure 1.4 shows a prior that has non-zero mass over all values of $\theta$ that has non-zero likelihood, and a discontinuity in the middle of the parameter space, with the left-half of the parameter space supporting having half as much probability mass as the right-half. The resulting posterior has a discontinuity at the point where the prior does, and has but, since the prior is otherwise uniform, the posterior inherits the shape of the likelihood on either side of the discontinuity, subject to the constraint (implied by the prior) that the posterior has twice as much probability mass in the right of the discontinuity than to the left, and integrates to one. The right-hand column of Figure 1.4 shows a more elaborate prior, a step function over the parameter space, decreasing to the right. The resulting posterior has discontinuities at the discontinuities in the prior, and some that are quite abrupt, depending on the conflict between the prior and likelihood in any particular segment of the prior.

The point here is that posterior distributions can sometimes look quite unusual, depending on the form of the prior and the likelihood for a particular problem. The fact that a posterior distribution may have a peculiar shape is of no great concern in a Bayesian analysis: provided one is updating prior beliefs via Bayes Rule, all is well. Unusual looking posterior distributions might suggest that one’s prior distribution was poorly specified, but, as a general rule, one should be extremely wary of engaging this kind of procedure. Bayes Rule is a procedure for generating posterior distributions over parameters in light of data. Although one can always re-run a Bayesian analysis with different priors (and indeed, this is usually a good idea), Bayesian procedures should not be used to hunt for priors that generate the most pleasing looking posterior distribution, given a particular data set and likelihood. Indeed, such a practice would amount to an inversion of the Bayesian approach: i.e., if the researcher has strong ideas as to what values of $\theta$ are more likely than others, aside from the information in the data, then that auxiliary information should be considered a prior,
with Bayes Rule providing a procedure for rationally combining that auxiliary information with the information in the data.

### 1.4.2 Bayesian Updating as Information Accumulation

Bayesian procedures are often equivalent to combining the information in one set of data with another set of data. In fact, if prior beliefs represent the result of a previous data analysis (or perhaps many previous data analyses), then Bayesian analysis is equivalent to pooling information. This is a particularly compelling feature of Bayesian analysis, and one that takes on special significance when working with conjugate priors. In these cases, Bayesian procedures accumulate information in the sense that the posterior distribution is more precise than either the prior distribution or the likelihood alone. Further, as the amount of data increases, say through repeated applications of the data generation process, the posterior precision will continue to increase, eventually overwhelming any non-degenerate prior; the upshot is that analysts with different (non-degenerate) prior beliefs over a parameter will eventually find their beliefs coinciding, provided they (1) see enough data and (2) update their beliefs using Bayes’ Theorem. In this way Bayesian analysis has been proclaimed as a model for scientific practice, acknowledging that while reasonable people may differ (at least prior to seeing data), our views will tend to converge as scientific knowledge accumulates, provided we update our views rationally, consistent with the laws of probability (i.e., via Bayes’ Theorem).

#### EXAMPLE 1.3

**Drug Testing, Example 1.1, continued.** Suppose that the randomly selected subject is someone you know personally, and you strongly suspect that she does not use the prohibited substance. Your prior over the hypothesis that she uses the prohibited substance is $P(H_U) = 1/1000$. I have no special knowledge regarding the athlete, and use the baseline prior $P(H_U) = .03$. After the positive test result, my posterior belief is $P(H_U|E) = .23$, while yours is

$$P(H_U|E) = \frac{P(H_U)P(E|H_U)}{\sum_{i \in \{U, \sim U\}} P(H_i)P(E|H_i)}$$

$$= \frac{.001 \times .95}{(.001 \times .95) + (.999 \times .10)}$$

$$= \frac{.00095}{.00095 + .0999}$$

$$= .009$$
A second test is performed. Now, our posteriors from the first test become the priors with respect to the second test. Again, the subject tests positive, which we denote as the event $E'$. My beliefs are revised as follows:

$$
P(H_U|E') = \frac{.23 \times .95}{(.23 \times .95) + (.77 \times .10)} = \frac{.2185}{.2185 + .077} = .74,
$$

while your beliefs are updated to

$$
P(H_U|E') = \frac{.009 \times .95}{(.009 \times .95) + (.991 \times .10)} = \frac{.00855}{.00855 + .0991} = .079.
$$

At this point, I am reasonably confident that the subject is using the prohibited substance, while you still attach reasonably low probability to that hypothesis. After a 3rd positive test your beliefs update to .45, and mine to .96. After a 4th positive test your beliefs update to .88 and mine to .996, and after a 5th test, your beliefs update to .99 and mine to .9996. That is, given this stream of evidence, common knowledge as to the properties of the test, and the fact that we are both rationally updating our beliefs via Bayes’ Theorem, our beliefs are converging.

In this case, given the stream of positive test results, our posterior probabilities regarding the truth of $H_U$ are asymptotically approaching 1.0, albeit mine more quickly than yours, given the low a priori probability you attached to $H_U$. Note that with my prior, I required just two consecutive positive test results to revise my beliefs to the point where I considered it more likely than not that the subject is using the prohibited substance, whereas you, with a much more skeptical prior, required four consecutive positive tests.

It should also be noted that the specific pattern of results obtained in this case depend on the properties of the test. Tests with higher sensitivity and specificity would see our beliefs be revised more dramatically given the sequence of positive test results. Indeed, this is the objective of the design of diagnostic tests of various sorts: given a prior $P(H_U)$, what levels of sensitivity and specificity are required such that after just one or two positive tests, $P(H_U|E)$ exceeds a critical threshold where an action is justified. See Exercise XXXX.
1.5 PARAMETERS AS RANDOM VARIABLES, BELIEFS AS DISTRIBUTIONS

One of the critical ways in which Bayesian statistical inference differs from frequentist inference is immediately apparent from equation 1.3 and the examples shown in Figure 1.2: the result of a Bayesian analysis, the posterior density \( p(\theta | y) \) is just that, a probability density. Given a subjectivist interpretation of probability that most Bayesians adopt, the “randomness” summarized by the posterior density is a reflection of the researcher’s uncertainty over \( \theta \), conditional on having observed data \( y \).

Contrast the frequentist approach, in which \( \theta \) is not random, but a fixed (but unknown) property of a population from which we randomly sample data \( y \). Repeated applications of the sampling process, if undertaken, would yield different \( y \), and different sample based estimates of \( \theta \), denoted \( \hat{\theta} = \hat{\theta}(y) \), this notation reminding us that estimates of parameters are functions of data. In the frequentist scheme, the \( \hat{\theta}(y) \) vary randomly across data sets (or would, if repeated sampling was undertaken), while the parameter \( \theta \) is a constant feature of the population from which data sets are drawn. The distribution of values of \( \hat{\theta} \) that would result from repeated application of the sampling process is called the sampling distribution, and is the basis of inference in the frequentist approach; the standard deviation of the sampling distribution of \( \hat{\theta} \) is the standard error of \( \hat{\theta} \), which plays a key role in frequentist inference.

The Bayesian approach does not rely on how \( \hat{\theta} \) might vary over repeated applications of random sampling. Instead, Bayesian procedures center on a simple question: what should I believe about \( \theta \) in light of the data available for analysis, \( y \). The quantity \( \hat{\theta}(y) \) has no special, intrinsic status in the Bayesian approach: as we shall see with specific examples in Chapter 2, a least squares or maximum likelihood estimate of \( \theta \) is a feature of the data that is usually helpful in computing the posterior distribution for \( \theta \). And, under some special circumstances, a least squares or maximum likelihood estimate of \( \theta \) \( \hat{\theta}(y) \) will correspond to a Bayes estimate of \( \theta \) (see section 1.6.1). We return to these points of difference between Bayesian and frequentist inference in Chapter 3. For now, the critical point to grasp is that in the Bayesian approach, the roles of \( \theta \) and \( \hat{\theta} \) are reversed relative to their roles in classical, frequentist inference: \( \theta \) is random, in the sense that the researcher is uncertain about its value, while \( \hat{\theta} \) is fixed, a feature of the data at hand.
1.6 COMMUNICATING THE RESULTS OF A BAYESIAN ANALYSIS

Having conducted a Bayesian analysis, all relevant information about \( \theta \) after having analyzed the data is represented by the posterior density, \( p(\theta | y) \). An important and interesting decision for the Bayesian researcher is how to communicate posterior beliefs about \( \theta \).

In a world where journal space was less scarce than it is, researchers could simply provide pictures of posterior distributions: e.g., density plots or histograms, as in Figure 1.2. Graphs are an extremely efficient way of presenting information, and, in the specific case of probability distributions, let the researcher and readers see the location, dispersion and shape of the distribution, immediately gauging what regions of the parameter space are more plausible than others, if any. This visualization strategy works well when \( \theta \) is a scalar, but quickly becomes more problematic when working with multiple parameters, and so the posterior density is a multivariate distribution: i.e., we have

\[
p(\theta | y) = p(\theta_1, \ldots, \theta_k | y) \propto p(\theta)p(y | \theta)
\]  

(1.4)

Direct visualization is longer feasible once \( k > 2 \): density plots or histograms have two-dimensional counterparts (e.g., contour or image plots, perspective plots), but all attempts to visualize higher dimensional surfaces fail. As the dimension of the parameter vector \( (k) \) increases, we can graphically present one or two dimensional slices of the posterior density. For problems with lots of parameters, this means that we may have lots of pictures to present, consuming more journal space than and even the most sympathetic editor may be able to provide.

Thus, for models with lots of parameters, graphical presentation of the posterior density may not be feasible, at least not for all parameters. In these cases, numerical summaries of the posterior distribution may be more feasible. Moreover, for most standard models, and if the researcher’s prior beliefs have been expressed with conjugate priors, the analytic form of the posterior is known (indeed, as we shall see, this is precisely the attraction of conjugate priors!). This means that for these standard cases, almost any interesting feature of the posterior can be computed directly: e.g., the mean, the mode, the standard deviation, or particular quantiles. For non-standard models, and/or for models where the priors are not conjugate, modern computational power lets us deploy Monte Carlo methods to compute these features of posterior densities; see Chapter 4. In this section I review proposals for summarizing posterior densities.
1.6.1 Bayesian Point Estimation

If a Bayesian point estimate is required — reducing the information in the posterior
distribution to a single number — this can be done, although some Bayesian statisticians
regard the attempt to reduce a posterior distribution to a single number as misguided and
ad hoc. For instance,

While it [is] easy to demonstrate examples for which there can be no satisfactory point
estimate, yet the idea is very strong among people in general and some statisticians in
particular that there is a need for such a quantity. To the idea that people like to have
a single number we answer that usually they shouldn’t get it. Most people know they
live in a statistical world and common parlance is full of words implying uncertainty.
As in the case of weather forecasts, statements about uncertain quantities ought to be
made in terms which reflect that uncertainty as nearly as possible (Box and Tiao 1973,
309-10).

This said, it is convenient to report a point estimate when communicating the results of a
Bayesian analysis, and, so long as information summarizing the dispersion of the posterior
distribution is also provided (see section 1.6.2, below), a Bayesian point estimate is quite a
useful quantity to report.

The choice of which point summary of the posterior distribution to report can be rational-
ized by drawing on (Bayesian) decision theory. Although we are interested in the specific
problem of choosing a single-number summary of a posterior distribution, the question of
how to make rational choices under conditions of uncertainty is quite general, and we begin
with a definition of loss:

**Definition 1.2** (Loss Function). Let $\Theta$ be a set of possible states of nature $\theta$, and let $a \in A$
be actions available to the researcher. Then define $l(\theta, a)$ as the loss to the researcher from
taking action $a$ when the state of nature is $\theta$.

Recall that in the Bayesian approach, the researcher’s beliefs about plausible values for $\theta$
are represented with a probability density function (or a probability mass function, if $\theta$
takes on discrete values), and, in particular, after looking at data $y$, beliefs about $\theta$ are represented
by the posterior density $p(\theta|y)$. Generically, let $\pi^*(\theta)$ be a probability distribution over
$\theta$. Such a distribution induces a distribution over losses. Averaging the losses over beliefs
about $\theta$ yields the Bayesian expected loss (Berger 1985, 8):
Definition 1.3 (Bayesian Expected Loss). If \( \pi^*(\theta) \) is the probability distribution for \( \theta \in \Theta \) at the time of decision making, the Bayesian expected loss of an action \( a \) is

\[
\varrho(\pi^*(\theta), a) = E[l(\theta, a)] = \int_{\Theta} l(\theta, a) \pi^*(\theta) d\theta
\]

A special case is where the distribution \( \pi^* \) in Definition 1.3 is a posterior distribution:

Definition 1.4 (Posterior Expected Loss). Given a posterior distribution for \( \theta \), \( p(\theta | y) \), the posterior expected loss of an action \( a \) is

\[
\varrho(p(\theta | y), a) = \int_{\Theta} l(\theta, a) p(\theta | y) d\theta
\]

A Bayesian rule for choosing actions \( a \) is to select \( a \) so to minimize posterior expected loss. In the specific context of point estimation, the decision problem is to choose a Bayes estimate, \( \tilde{\theta} \), and so actions \( a \in A \) now index feasible values for \( \tilde{\theta} \in \Theta \). The problem now is that since there are plausibly many different loss functions one might adopt, there are plausibly many Bayesian point estimates one might choose to report. If the chosen loss function is convex, then the Bayes estimate is unique (DeGroot and Rao 1963), so the choice of what Bayes estimate to report usually amounts to what (convex) loss function to adopt. Here I consider some well-studied cases.

Definition 1.5 (Quadratic Loss). If \( \theta \in \Theta \) is a parameter of interest, and \( \tilde{\theta} \) is an estimate of \( \theta \), then

\[
l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2
\]

is the quadratic loss arising from the use of the estimate \( \tilde{\theta} \) instead of \( \theta \).

With quadratic loss, we obtain the following useful result:

Proposition 1.6 (Posterior Mean as a Bayes Estimate). Under quadratic loss the Bayes estimate of \( \theta \) is the mean of the posterior distribution, i.e.,

\[
\tilde{\theta} = E(\theta | y) = \int_{\Theta} p(\theta | y) \theta d\theta
\]

Proof. Quadratic loss (Definition 1.5) implies that the posterior expected loss is

\[
\varrho(\theta, \tilde{\theta}) = \int_{\Theta} (\theta - \tilde{\theta})^2 p(\theta | y) d\theta.
\]

and we seek to minimize this expression with respect to \( \tilde{\theta} \). Expanding the quadratic yields

\[
\varrho(\theta, \tilde{\theta}) = \int_{\Theta} \theta^2 p(\theta | y) d\theta + \tilde{\theta}^2 \int_{\Theta} p(\theta | y) d\theta - 2\tilde{\theta} \int_{\Theta} \theta p(\theta | y) d\theta
\]

\[
= \int_{\Theta} \theta^2 p(\theta | y) d\theta + \tilde{\theta}^2 - 2\tilde{\theta} E(\theta | y),
\]

\[
= \int_{\Theta} (\theta - \tilde{\theta})^2 p(\theta | y) d\theta.
\]
Differentiate with respect to $\tilde{\theta}$, noting that the first term does not involve $\tilde{\theta}$, and set the result to zero to establish the result.

This result also holds for the case of performing inference with respect to a parameter vector $\theta = (\theta_1, \ldots, \theta_K)'$. In this more general case, we define a multidimensional quadratic loss function as follows:

**Definition 1.6 (Multidimensional Quadratic Loss).** If $\theta \in \mathbb{R}^K$ is a parameter, and $\tilde{\theta}$ is an estimate of $\theta$, then the (multidimensional) quadratic loss is

$$l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})'Q(\theta - \tilde{\theta})$$

where $Q$ is a positive definite matrix.

**Proposition 1.7 (Multidimensional Posterior Mean as Bayes Estimate).** Under quadratic loss (Definition 1.6), the posterior mean $E(\theta|y) = \int_\Theta p(\theta|y) \theta d\theta$ is the Bayes estimate of $\theta$.

**Proof.** The posterior expected loss is

$$\varrho(\theta, \tilde{\theta}) = \int_\Theta (\theta - \tilde{\theta})'Q(\theta - \tilde{\theta}) p(\theta|y)d\theta.$$ 

Differentiating with respect to $\tilde{\theta}$ yields

$$2Q\int_\Theta (\theta - \tilde{\theta}) p(\theta|y)d\theta$$

and setting to zero and re-arranging yields

$$\int_\Theta (\theta - \tilde{\theta}) p(\theta|y)d\theta = 0,$$

or

$$\int_\Theta \theta p(\theta|y)d\theta = \int_\Theta \tilde{\theta} p(\theta|y)d\theta.$$

The left-hand side of this expression is just the mean of the posterior density, $E(\theta|y)$, and so

$$E(\theta|y) = \int_\Theta \tilde{\theta} p(\theta|y)d\theta = \tilde{\theta} \int_\Theta p(\theta|y)d\theta = \tilde{\theta}.$$ 

**Remark.** This result holds irrespective of the specific weighting matrix $Q$, provided $Q$ is positive definite.
The mean of the posterior distribution is a popular choice among researchers seeking to quickly communicate features of the posterior distribution that results from a Bayesian data analysis; we now understand the conditions under which this is a rational point summary of one’s beliefs over $\theta$. Specifically, Proposition 1.6 rationalizes the choice of the mean of the posterior distribution as a Bayes estimate.

Of course, other loss functions rationalize other point summaries. Consider linear loss, possibly asymmetric around $\theta$:

**Definition 1.7 (Linear Loss).** If $\theta \in \Theta$ is a parameter, and $\tilde{\theta}$ is a point estimate of $\theta$, then the linear loss function is

$$l(\theta, \tilde{\theta}) = \begin{cases} k_0(\theta - \tilde{\theta}) & \text{if } \tilde{\theta} < \theta \\ k_1(\tilde{\theta} - \theta) & \text{if } \theta \leq \tilde{\theta} \end{cases}$$

Loss in absolute value results when $k_0 = k_1 = 1$, a special case of a class of symmetric, linear loss functions (i.e., $k_0 = k_1$). Asymmetric loss results when $k_0 \neq k_1$.

**Proposition 1.8 (Bayes Estimates under Linear Loss).** Under linear loss (definition 1.7), the Bayes estimate of $\theta$ is the $k_1/(k_0 + k_1)$ quantile of $p(\theta|y)$, the $\tilde{\theta}$ such that $P(\theta \leq \tilde{\theta}) = k_0/(k_0 + k_1)$.

**Proof.** Following Bernardo and Smith (1994, 256), we seek the $\tilde{\theta}$ that minimizes

$$\varrho(\theta, \tilde{\theta}) = \int_{\Theta} l(\theta, \tilde{\theta})p(\theta|y)d\theta = k_0 \int_{\{\theta < \tilde{\theta}\}} (\theta - \tilde{\theta})p(\theta|y)d\theta + k_1 \int_{\{\theta \leq \tilde{\theta}\}} (\tilde{\theta} - \theta)p(\theta|y)d\theta.$$

Differentiating this expression with respect to $\tilde{\theta}$ and setting the result to zero yields

$$k_0 \int_{\{\theta < \tilde{\theta}\}} p(\theta|y)d\theta = k_1 \int_{\{\theta \leq \tilde{\theta}\}} p(\theta|y)d\theta$$

Adding $k_0 \int_{\{\theta \leq \tilde{\theta}\}} p(\theta|y)d\theta$ to both sides yields

$$k_0 = (k_0 + k_1) \int_{\{\theta \leq \tilde{\theta}\}} p(\theta|y)d\theta$$

and so re-arranging yields

$$\int_{\{\theta < \tilde{\theta}\}} p(\theta|y)d\theta = k_0/(k_0 + k_1).$$

Note that with symmetric linear loss, we obtain the median of the posterior distribution as the Bayes estimate. Asymmetric loss functions imply using quantiles other than the median.
EXAMPLE 1.4

Graduate Admissions. A committee reviews applications to a Ph.D. program. Each applicant $i \in \{1, \ldots, n\}$ possesses ability $\theta_i$. After reviewing the applicants’ files (data, or $y$), the committees’ beliefs regarding $\theta_i$ can be represented as a distribution $p(\theta_i|y)$. The committee’s loss function is asymmetric, since the committee has determined that it is 2.5 times as costly to overestimate an applicant’s ability than it is to underestimate ability: i.e.,

\[ q(\theta, \tilde{\theta}) = \begin{cases} 
\theta - \tilde{\theta} & \text{if } \theta > \tilde{\theta} \\
2.5(\tilde{\theta} - \theta) & \text{if } \theta \leq \tilde{\theta}
\end{cases} \]

Ability is measured on an arbitrary scale, normalized to have mean zero and standard deviation one across the applicant pool. For applicant $i$, $p(\theta_i|y) \approx N(1.8, 0.4^2)$. Given the committee’s loss function, their Bayes estimate of $\theta_i$ is the $1/(1 + 2.5) = .286$ quantile of their $N(1.8, 0.4^2)$ posterior distribution, or 1.57.

1.6.2 Credible Regions

Bayes estimates are an attempt to summarize beliefs over $\theta$ with a single number, providing a rational, best guess as to the value of $\theta$. But Bayes estimates do not convey information as to the researcher’s uncertainty over $\theta$, and indeed, this is why many Bayesian statisticians find Bayes estimates fundamentally unsatisfactory. To communicate a summary of prior or posterior uncertainty over $\theta$, it is necessary to somehow summarize information about the location and shape of the prior or posterior distribution, $\pi^*(\theta)$. In particular, what is the set or region of more plausible values for $\theta$? More formally, what is the region $C \subseteq \Omega$ that supports proportion $\alpha$ of the probability under $\pi^*(\theta)$? Such a region is called a credible region:

Definition 1.8 (Credible Region). A region $C \subseteq \Omega$ such that

\[ \int_C \pi^*(\theta) d\theta = 1 - \alpha, \quad 0 \leq \alpha \leq 1 \]

is a $100(1 - \alpha)\%$ credible region for $\theta$.

For single-parameter problems (i.e., $\Omega \subseteq \mathbb{R}$), if $C$ is not a set of disjoint intervals, then $C$ is a credible interval.

If $\pi^*(\theta)$ is a (prior/posterior) density, then $C$ is a (prior/posterior) credible region.

There is trivially only one 100% credible region, the entire support of $\pi^*(\theta)$. But non-trivial credible regions may not be unique. For example, suppose $\theta \sim N(0, 1)$: it is obvious
that there is no unique 100(1 − α)% credible region for any α ∈ (0, 1): any interval spanning 100(1 − α) percentiles will be such an interval. A solution to this problem comes from restricting attention to credible regions that have certain desirable properties, including minimum volume (or, for a one dimensional parameter problem, minimum length) in the set of credible regions induced by a given choice of α, for a specific π∗(θ). This kind of optimal credible region is called a highest probability density region, sometimes referred to as a HPD region or a “HDR”. The following definition of a HPD region is standard and appears in many places in the literature, e.g., Box and Tiao (1973, 123) or Bernardo and Smith (1994, 260):

**Definition 1.9** (Highest Probability Density Interval). A region $C \subseteq \Omega$ is a 100(1 − α)% highest probability density region for θ under $\pi^*(\theta)$ if

1. $P(\theta \in C) = 1 − α$

2. $P(\theta_1) \geq P(\theta_2), \forall \theta_1 \in C, \theta_2 \notin C$

A 100(1 − α)% HPD region for a symmetric, unimodal density is obviously unique and symmetric around the mode. In fact, if $\pi^*(\theta)$ is a univariate normal distribution, a HPD is the same as a confidence interval around the mean:

**EXAMPLE 1.5**

Suppose $\pi^*(\theta) \equiv N(a, b^2)$. Then a 100(1 − α)% HPD region is the interval $(a - |z_\alpha|b, a + |z_\alpha|b)$ where $z_\alpha$ is the α quantile of the standard normal distribution. With $\alpha = .05$, $|z_\alpha| \approx 1.96$, and a 95% HPD corresponds to a 95% confidence interval; see Figure 1.5.

Note that the correspondence between confidence intervals and HPD intervals does not hold for non-symmetric distributions:

**EXAMPLE 1.6**

Figure 1.6 shows a $\chi^2$ distribution with 4 degrees of freedom, and its 50% HPD interval. Notice that the 50% HPD interval is more concentrated around the mode of the distribution, and has shorter length than the interval based on the 25th to 75th percentiles of the distribution.
Figure 1.5 Standard normal distribution and 95% HPD interval
Figure 1.6 $\chi^2_4$ distribution, with 50% HPD interval
As the next two examples demonstrate, (1) the HPD need not be connected set, but a collection of disjoint intervals (say, if \( \pi^*(\theta) \) is not unimodal), and (2) the HPD need not to be unique, if \( \pi^*(\theta) \) has a region of uniform probability sufficiently large relative to \( \alpha \).

**EXAMPLE 1.7**

Consider the data in Table 1.1., where two variables \((y_1, y_2)\) are observed subject to a pattern of severe missingness, but are otherwise assumed to be distributed bivariate normal each with mean zero, and an unknown covariance matrix. These manufactured data have been repeatedly analyzed to investigate the properties of algorithms for handling missing data (e.g., Murray 1977; Tanner and Wong 1987), and we will return to this example in more detail in Chapter XXXX.

Given the missing data pattern, what should we can we conclude about the correlation \( \rho \) between \( y_1 \) and \( y_2 \)? For this particular example, with an uninformative prior for the covariance matrix of \( Y = (y_1, y_2) \), the posterior distribution for \( \rho \) is bimodal, as shown in Figure 1.7. The shaded areas represents half the probability mass under the posterior distribution for \( \rho \); the intervals supporting the shaded areas together constitute a 50% HPD region for \( \rho \), and are the disjoint intervals \((-0.914, -0.602)\) and \((0.602, 0.914)\). We return to this example in more detail in Chapter XXXX.

\[
\begin{align*}
y_1: & \ 1 \ 1 \ -1 \ -1 \ 2 \ 2 \ -2 \ -2 \ NA \ NA \ NA \ NA \\
y_2: & \ 1 \ -1 \ 1 \ -1 \ NA \ NA \ NA \ NA \ 2 \ 2 \ -2 \ -2
\end{align*}
\]

**Table 1.1.** Twelve Observations from a Bivariate Normal Distribution.

**EXAMPLE 1.8**

Suppose \( \theta \sim \text{Uniform}(0, 1) \). Then any HPD region of content \( \alpha \) is not unique, \( \forall 0 < \alpha < 1 \). See Figure 1.8. The shaded regions are both supported by 25% HPDs, as are any other intervals of width .25 we might care to draw.

For higher dimensional problems, the HPD is a region in a parameter space and numerical approximations and/or simulation may be required to compute it; see Chapter 4. For some simple cases, such as multiple regression analysis with conjugate priors, although the posterior distribution is multivariate, it has a well known form for which it is straightforward to compute HPDs; see Chapter 5.
Figure 1.7  Bimodal Posterior Density for a Correlation Coefficient, and 50% HPD.
Figure 1.8 Uniform Distribution and 25% HPDs
1.7 ASYMPTOTIC PROPERTIES OF POSTERIOR DISTRIBUTIONS

As we have seen, Bayes Rule tells us how to rationally update our prior beliefs in light of data. In section 1.4 we saw that as the precision of one’s prior beliefs tends to zero, posterior beliefs are increasingly dominated by the data (through the likelihood). This also occurs as the data set “gets larger”: subject to an exception to be noted below, for a given prior, as the size of the data sets being analyzed grow without bound, the usual result is that the resulting sequence of posterior distributions collapses to a spike on the true values of the parameters in the model under consideration.

Of course, many Bayesians find such thinking odd: in a Bayesian analysis, we condition on the data at hand, updating beliefs via Bayes Rule. Unlike frequentist inference, Bayesian inference does not rest on the repeated sampling and/or asymptotic properties of the statistical procedures being used. Many Bayesians consider asking what would happen as one’s data set got infinitely large as an interesting mathematical exercise, but not particularly relevant to the statistical task at hand. Given a Bayesian approach, the analysis of imaginary data sets — of whatever size, finite or infinite — is simply unnecessary. Provided I update my beliefs via Bayes Rule in light of this data set, I’m behaving rationally, and the repeated sampling or asymptotic properties of my inferences are second order concerns.

This said, it is worthwhile to consider these “second order” concerns. The underlying idea is that subject to some regularity conditions, as the data set grows without bound, the posterior density is increasingly dominated by the contribution from the likelihood function, and the “usual” asymptotic properties of maximum likelihood estimators apply to the posterior. These properties include

- consistency, at least in the Bayesian sense of the posterior density being increasingly concentrated around the true parameter value as \( n \to \infty \); or, in the additional sense of Bayes point estimators of \( \theta \) (section 1.6.1) being consistent;

- asymptotic normality, i.e., \( p(\theta|y) \) tends to a normal distribution as \( n \to \infty \).

There is a large literature establishing the conditions under which frequentist and Bayesian procedures coincide, at least asymptotically. These results are too technical to be reviewed in any detail in this text; see, for instance, Bernardo and Smith (1994, ch5) for statements of necessary regularity conditions and proofs of the main results and references to the literature. Diaconis and Freedman (1986b, 1986a) provide some counter-examples to the consistency results; the “incidental parameters” problem is one such counter-example. I
Figure 1.9  Sequence of Posterior Densities (1). The prior remains fixed across the sequence, as sample size increases and \( \theta^* \) is held constant. In this example, \( n = 6, 30, 90, 450 \) across the four columns in the figure.

provide a brief illustration of “Bayesian consistency” two examples, below, and sketch a proof of a “Bayesian central limit theorem” in the Appendix.

Suppose the true value of \( \theta \) is \( \theta^* \). Then provided the prior distribution \( p(\theta) \neq 0 \) does not place zero probability mass on \( \theta^* \) (say, for a discrete parameter), or in a neighborhood of \( \theta^* \) (say, for a continuous parameter), then as \( n \to \infty \), the posterior will be increasingly dominated by the contribution from the likelihood, which, under suitable regularity conditions, tends to a spike on \( \theta^* \).
Figure 1.10  Sequence of Posterior Distributions (2). The prior remains fixed across the sequence, as sample size increases and $\theta^*$ is held constant. In this example, $n = 6, 30, 150, 1500$ across the four columns in the figure.
Figures 1.9 and 1.10 graphically demonstrate the Bayesian version of consistency as described above. In each case, the prior is held constant as the sample size increases, leading to a progressively tighter correspondence between the posterior and the likelihood. Even with modest amounts of data, the multimodality of the priors are being overwhelmed by the information in the data, and the likelihood and posterior distributions are collapsing to a spike on $\theta^*$. 

Although the scale used in Figures 1.9 and 1.10 doesn’t make it clear, the likelihoods and posterior distributions in the Figures are also tending to normal distributions: re-scaling by the usual $\sqrt{n}$ would make this clear. The fact that posterior densities start to take on a normal shape as $n \to \infty$ is particularly helpful. The normal is an extremely well-studied distribution, and completely characterized by its first two moments. This can drastically simplify the Bayesian computation of the posterior density and features of the posterior density, such as quantiles and highest posterior density estimates, especially when $\theta$ has many components.

## 1.8 BAYESIAN HYPOTHESIS TESTING

The posterior distribution also provides the information necessary to test hypotheses about $\theta$. Suppose we have a continuous parameter $\theta \in \mathbb{R}$ and two hypotheses $H_0 : \theta < c$ and the alternative hypothesis $H_1 : \theta \geq c$. Then it is straightforward to see that

$$P(H_0|y) = P(\theta < c|y) = \int_{-\infty}^{c} p(\theta|y)d\theta$$

and

$$P(H_1|y) = P(\theta \geq c|y) = \int_{c}^{\infty} p(\theta|y)d\theta.$$ 

Again, for standard models, where conjugate priors have been deployed, these probabilities are straightforward to compute; in other cases, modern computing power means Monte Carlo methods can be deployed to assess these probabilities, as we will see in Chapter 4.

### EXAMPLE 1.9

*Attitudes Towards Abortion.* Agresti and Finlay (1997, 133) report that in the 1994 General Social Survey, 1,934 respondents were asked

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason."
Of the 1,934 respondents, 895 reported “yes” and 1,039 said “no”. Let \( \theta \) be the unknown population proportion of respondents who agree with the proposition in the survey item, that a pregnant woman should be able to obtain an abortion if the woman wants it for any reason. The question of interest is whether a majority of the population supports the proposition in the survey item.

The survey estimate of \( \theta \) is 
\[
\hat{\theta} = \frac{895}{1934} \approx .46,
\]
the approximation coming via rounding to two significant digits. Although the underlying data are binomial (independent Bernoulli trials), with this large sample, the normal distribution provides an excellent approximation to the frequentist sampling distribution of \( \hat{\theta} \); binomial data are considered in detail in Chapter 2. The standard deviation of the normal sampling distribution (the standard error of \( \hat{\theta} \)) is given by the usual estimate
\[
se(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{.46 \times (1-.46)}{1934}} \approx .011.
\]

Suppose interest focuses on whether the unknown population proportion \( \theta = .5 \). A typical frequentist approach to this question is to test the null hypothesis \( H_0 : \theta = .5 \) against all other alternatives \( H_A : \theta \neq .5 \), or a one-sided alternative \( H_B : \theta > .5 \). We would then ask how unlikely is that one would see the value of \( \hat{\theta} \) actually obtained, or an even more extreme value if \( H_0 \) were true, by centering the sampling distribution of \( \hat{\theta} \) at the hypothesized value. Then, we would exploit the facts that (1) the sampling distribution of \( \hat{\theta} \) is well-approximated by a normal distribution; (2) the sample data provide an estimate of the standard deviation of that normal distribution, the standard error of \( \hat{\theta} \). In this case, we note that the actually realized value of \( \hat{\theta} \) is 
\[
(\frac{.5 - .46}{.011}) \approx 3.6 \text{ standard errors away from the hypothesized value. Under a normal distribution, this is a rare event. Over repeated applications of random sampling, only a small proportion of point estimates of } \theta \text{ will lie 3.6 or more standard errors away from the hypothesized mean of the sampling distribution. This proportion is}
\]
\[
2 \times \int_{3.6}^{\infty} \phi(z)dz = 2 \times [1 - \Phi(3.6)] \approx .0003,
\]
where \( \phi(\cdot) \) and \( \Phi(\cdot) \) are the normal pdf and cdfs, respectively. Given this result, most (frequentist) analysts would reject the null hypothesis in favor of either alternative hypothesis, reporting the \( p \)-values for \( H_0 \) against \( H_A \) as .0003 and for \( H_0 \) against \( H_A \) as .00016.
The Bayesian approach is as follows. The unknown parameter is $\theta \in [0, 1]$ and suppose we bring no prior information to the analysis, adopting the uninformative prior $\theta \sim \text{Uniform}(0, 1)$. In such a case, we know that the results of the Bayesian analysis and the frequentist analysis will coincide, save for the important distinction of what is random and what is fixed. The details of the Bayesian computation await Chapter 2, but the application of Bayes Rule here results in a posterior distribution for $\theta$ that is well approximated by a normal distribution centered on the maximum likelihood estimate of .46 with standard deviation .011; i.e., $p(\theta|y) \approx N(.46, .011^2)$ and any inferences about $\theta$ are based on this distribution. We note immediately that most of the posterior probability mass lies below .5, suggesting that the hypothesis $\theta > .5$ is not well-supported by the data. In fact, the posterior probability of this hypothesis is
\[
\Pr(\theta > .5|y) = \int_{.5}^{\infty} p(\theta|y)d\theta = \int_{.5}^{\infty} \phi\left(\frac{\theta - .46}{.011}\right) d\theta = \int_{3.6}^{\infty} \phi(z)dz = .00016.
\]
That is, there is a superficial symmetry between the frequentist and Bayesian answers: in both instances, the “answer” involved computing the same probability mass in the tail of a normal distribution, with probability of $H_0$ under the Bayesian posterior distribution corresponding with the $p$-value in the frequentist test of $H_0$ against the one-sided alternative $H_B$; see Figure 1.11. But this similarity really is only superficial. The Bayesian probability is a statement about the researcher’s beliefs about $\theta$, obtained via application of Bayes Rule, and is $P(H_0|y)$, obtained by computing the appropriate integral of the posterior distribution $p(\theta|y)$. The frequentist probability is obtained via a slightly more complex route, and has a quite different interpretation than the Bayesian posterior probability, since it conditions on the null hypothesis; i.e., the sampling distribution is $f(\hat{\theta}|H_0)$ and the $p$-value for $H_1$ against the one-sided alternative, the proportion of $\hat{\theta} < .46$ we would see under repeated sampling, with a sampling distribution given by the null hypothesis. When this $p$-value gets sufficiently small (below a “critical value” pre-set by the researcher), the null hypothesis is rejected in favor of the specified alternative. That is, there is another stage in the frequentist chain of reasoning to convert the $p$-value into a statement or decision about the plausibility of $H_0$. These and other contrasts between frequentist and Bayesian hypothesis testing are considered in greater detail in Chapter 3.

It it worth stressing that in the Bayesian framework, “point null hypotheses” are almost always uninteresting. If a continuous parameter $\theta \in \Omega \subseteq \mathbb{R}$ has the posterior distribution
Figure 1.11  Posterior Distribution and Sampling Distribution, for Example 1.9.
then a point null hypothesis such as \( H_0 : \theta = c \) has zero probability, since \( c \) is a 1-point set and has measure zero (i.e., \( \int_{\mathcal{C}} p(\theta | y) d\theta = 0, \forall c \in \Omega \)), or, put simply, the probability of a point in a continuous parameter space is zero. We return to the distinction between hypothesis testing in frequentist and Bayesian frameworks in Chapter 3.

### 1.8.1 Model Choice

In the Bayesian approach, decisions about the truth or falsity of hypotheses do not result directly from Bayes’ Rule. Recall that application of Bayes Rule produces a posterior distribution, \( f(\theta | y) \), not a point estimate or a binary decision about a hypothesis. Nonetheless, in many settings the goal of statistical analysis is to inform a discrete decision problem, such as choosing the “best model” from a class of models for a given data set. We now consider Bayesian procedures for making such a choice.

Let \( M_i \) index models under consideration for data \( y \). What may distinguish the models are parameter restrictions of various kinds. A typical example in the social sciences is when sets of predictors are entered or dropped from different regression-type models for \( y \); if \( j \) indexes candidate predictors, then dropping \( x_j \) from a regression corresponds to imposing the parameter restriction \( \beta_j = 0 \). Alternatively, the models under consideration may not nest or overlap. For example, consider situations where different theories suggest disjoint sets of predictors for some outcome \( y \). In this case two candidate models \( M_1 \) and \( M_2 \) may have no predictors in common.

Consider a closed set of models, \( \mathcal{M} = \{ M_1, \ldots, M_I \} \); i.e., the researcher is interested in choosing among a distinct number of models, rather than the (harder) problem of choosing a model from an infinite set of possible models. In the Bayesian approach, the researcher has prior beliefs as to which model is correct, which are formulated as prior probabilities, denoted \( P(M_i) \) with \( i \) indexing the set of models \( \mathcal{M} \). The goal of a Bayesian analysis is to produce posterior probabilities for each model, \( P(M_i | y) \), and to inform the choice of a particular model. This posterior probability comes via application of Bayes Rule for multiple discrete events, which we encountered earlier as Proposition 1.4. In the specific context of model choice, we have

\[
P(M_i | y) = \frac{P(M_i) p(y | M_i)}{\sum_{j=1}^{I} P(M_j) p(y | M_j)}.
\]  

(1.5)

The expression \( p(y | M_i) \) is the marginal likelihood, given by the identity

\[
P(y | M_i) = \int_{\theta_i} p(y | \theta_i, M_i) p(\theta_i) d\theta_i
\]

(1.6)
i.e., averaging the likelihood for $y$ under $M_i$ over the prior for the parameters $\theta_i$ of $M_i$.

As we have seen in the discussion of Bayes estimates, the mapping from a researcher’s posterior distribution to a particular decision depends on the researcher’s loss function. To simplify the model choice problem, suppose that one of the models in $M$ is the “best model”, $M^*$, and the researcher possesses the following simple loss function

$$l(M_i, M^*) = \begin{cases} 0 & \text{if } M_i = M^* \\ 1 & \text{if } M_i \neq M^* \end{cases}$$

For each model, $P([M_i \neq M^*] | y) = 1 - P(M_i | y)$, and so the expected posterior loss of choosing model $i$ is $1 - P(M_i | y)$. Thus, the loss minimizing choice is to choose the model with highest posterior probability.

**Example 1.10**

*Attitudes Towards Abortion, Example 1.9, continued.* The likelihood for these data is approximated by a normal distribution with mean $.46$ and standard deviation $.011$. We consider the following hypotheses:

$$H_0 : 0.5 \leq \theta \leq 1$$
$$H_1 : 0 \leq \theta < 0.5$$

which generate priors

$$p_0(\theta_0) \equiv \text{Uniform}(0.5, 1) = \begin{cases} 2 & \text{if } 0.5 \leq \theta_0 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_1(\theta_1) \equiv \text{Uniform}(0, 0.5) = \begin{cases} 2 & \text{if } 0 \leq \theta_1 < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

respectively. We are *a priori* neutral between the two hypotheses, setting $P(H_0) = P(H_1)$ to $1/2$. Now, under $H_0$, the marginal likelihood is

$$p(y|H_0) = \int_{0.5}^{1} p(y|H_0, \theta_0) p_1(\theta_0) d\theta_0$$

$$= 2 \int_{0.5}^{1} p(y|H_0, \theta_0) d\theta_0$$

$$= 2 \left( \Phi \left( \frac{1 - .46}{.011} \right) - \Phi \left( \frac{.5 - .46}{.011} \right) \right) = 0.00028$$

Under $H_1$, the marginal likelihood is

$$p(y|H_1) = \int_{0}^{0.5} p(y|H_1, \theta_1) p_1(\theta_1) d\theta_1$$

$$= 2 \int_{0}^{0.5} p(y|H_1, \theta_1) d\theta_1$$

$$= 2 \left( \Phi \left( \frac{.5 - .46}{.011} \right) - \Phi \left( \frac{- .46}{.011} \right) \right) = 2.$$
Thus, via equation 1.5:

\[
P(H_0|y) = \frac{\frac{1}{2} \times .00028}{(\frac{1}{2} \times .00028) + (\frac{1}{2} \times 2)} = \frac{.00014}{.00014 + 1} = .00014
\]

\[
P(H_1|y) = \frac{1}{.00014 + 1} = .99986
\]

indicating that \( H_1 \) is much more plausible than \( H_0 \).

### 1.8.2 Model Averaging

XXXXX

### 1.8.3 Bayes Factors

For any pairwise comparison of models or hypotheses, we can also rely on a quantity known as the Bayes factor. Before seeing the data, the prior odds of \( M_1 \) over \( M_0 \) are \( p(M_1)/p(M_0) \), and after seeing the data we have the posterior odds \( p(M_1|y)/p(M_0|y) \). The ratio of these two sets of odds is the Bayes factor:

**Definition 1.10 (Bayes Factor).** Given data \( y \) and two models \( M_0 \) and \( M_1 \), the Bayes factor

\[
B_{10} = \frac{p(y|M_1)}{p(y|M_0)} = \left\{ \frac{p(M_1|y)}{p(M_0|y)} \right\} \left/ \left\{ \frac{p(M_1)}{p(M_0)} \right\} \right.
\]

is a summary of the evidence for \( M_1 \) against \( M_0 \) provided by the data.

The Bayes’ factor provides a measure of whether the data have altered the odds on \( M_1 \) relative to \( M_0 \). For instance, \( B_{10} > 1 \) indicates that \( M_1 \) is now more plausible relative to \( M_0 \) than it was a priori.

The Bayes factor plays something of an analogous role to a likelihood ratio. In fact, twice the logarithm of \( B_{10} \) is on the same scale as the deviance and likelihood ratio test statistics for model comparisons. For cases where the models are labelled by point restrictions on \( \theta \), the Bayes factor is a likelihood ratio. However, unlike the likelihood ratio test statistic, in the Bayesian context there is no reference to a sampling distribution with which to assess the particular statistic obtained in the present sample. In the Bayesian approach, all inferences are made conditional on the data at hand (not with reference to what might happen over repeated applications of random sampling). Thus, the Bayes factor has to be interpreted as a summary measure of the information in the data about the relative plausibility of models.
or hypotheses, rather than offering a formulaic way to choose between those model or hypotheses. Jeffreys (1961) suggests the following scale for interpreting the Bayes factor:

<table>
<thead>
<tr>
<th>$B_{10}$</th>
<th>$2 \log B_{10}$</th>
<th>evidence for $M_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1</td>
<td>&lt; 0</td>
<td>negative (support $M_0$)</td>
</tr>
<tr>
<td>1 to 3</td>
<td>0 to 2</td>
<td>barely worth mentioning</td>
</tr>
<tr>
<td>3 to 12</td>
<td>2 to 5</td>
<td>positive</td>
</tr>
<tr>
<td>12 to 150</td>
<td>5 to 10</td>
<td>strong</td>
</tr>
<tr>
<td>&gt; 150</td>
<td>&gt; 10</td>
<td>very strong</td>
</tr>
</tbody>
</table>

Good (1988) summarizes the history of the Bayes factor, which long predates likelihood ratios as a model comparison tool.

**EXAMPLE 1.11**

*Attitudes Towards Abortion, Example 1.9, continued.* We computed the marginal likelihoods under the two hypotheses in Example 1.10, which we now use to compute the Bayes factor,

$$B_{10} = \frac{p(y|H_1)}{p(y|H_0)} = \frac{2}{.00028} = 7142,$$

again indicating that the data strongly favor $H_1$ over $H_0$.

### 1.9 FROM SUBJECTIVE BELIEFS TO PARAMETERS AND MODELS

In this chapter I introduced Bayes Theorem with some simple examples. But in so doing I have brushed over some important details. In particular, the examples are all parametric (as are almost all statistical models in the social sciences), in that the sense that the probability distribution of the data is written as a function of an unknown parameter $\theta$ (a scalar or vector). This approach to statistical inference – expressing the joint density of the data as a function of a relatively small number of unknown parameters – will be familiar to many readers, and may not warrant justification or elaboration. But, given the subjectivist approach adopted here, the question of how and why parameters and models enter the picture is not idle.

Recall that the in the subjectivist approach championed by de Finetti (and adopted here), the idea that probability is a property of a coin, a die, or any other object under study, is considered metaphysical nonsense. The only thing that is real is the data at hand and whatever information we may have about how the data were generated (e.g., a simulation exercise,
an experiment with random assignment to treatment and control groups, random sampling from a specific population, or complete enumeration of a population). But everything else is a more or less convenient fiction created in the mind of the researcher, including parameters and models.

To help grasp the issue a little more clearly, consider the following example. A coin is flipped $n$ times. The possible set of outcomes is $\mathcal{S} = \{\{H, T\}_1 \otimes \ldots \otimes \{H, T\}_n\}$, with cardinality $2^n$. Assigning probabilities over the elements of $\mathcal{S}$ is a difficult task, if only because for any moderate to large value of $n$, $2^n$ is a large number. Almost instinctively, we start reaching for familiar ways to simplify the problem. For example, reaching back to our introductory statistics classes, we would probably inquire “are the coin flips are independent?” If satisfied that the coin flips are independent, we would then fit a binomial model to the data, modeling the $r$ flips coming up heads as a function of a “heads” probability $\theta$, given the $n$ flips. In a Bayesian analysis we’d also have a prior density $p(\theta)$ as part of the model, and we would report the posterior density over $p(\theta | r, n)$ as the result of the analysis.

I now show that this procedure can be justified by recourse to a deeper principle called exchangeability. In particular, if data are infinitely exchangeable, then a Bayesian approach to modeling the data is not possible or desirable, but is actually implied by exchangeability. That is, prior distributions over parameters are not merely a “Bayesian addition” to an otherwise classical analysis, but necessarily arise when one believes that the data are exchangeable. This is the key insight of one of the most important theorems in Bayesian statistics – de Finetti’s Representation Theorem – which we will also encounter below.

### 1.9.1 Exchangeability

We begin with a definition:

**Definition 1.11 (Finite Exchangeability).** The random quantities $y_1, \ldots, y_n$ are finitely exchangeable if their joint probability density (or mass function, for discrete $y$),

$$p(y_1, \ldots, y_n) = p(y_{z(1)}, \ldots, y_{z(n)})$$

for all permutations $z$ of the indices of the $y_i$, $\{1, \ldots, n\}$.

**Remark.** An infinite sequence of random quantities $y_1, y_2, \ldots$ is infinitely exchangeable if every finite subsequence if finitely exchangeable.

Exchangeability is thus equivalent to the condition that the joint density of the data $y$ remains the same under any re-ordering or re-labeling of the indices of the data. Similarly,
exchangeability is often interpreted as the Bayesian version of the “iid assumption” that underlies much statistical modeling, where “iid” stands for “independently and identically distributed”. In fact, if data are exchangeable they are conditionally iid, where the conditioning is usually on a parameter, \( \theta \). Indeed, this is one of the critical implications of de Finetti’s Representation Theorem.

As we shall now see, de Finetti’s Theorem shows that beliefs about data being infinitely exchangeable imply a belief about the data having “something in common”, a “similiarity” or “equivalence” (de Finetti’s original term) such that I can swap \( y_i \) for \( y_j \) in the sequence without changing my beliefs that either \( y_i \) or \( y_j \) will be one or zero (i.e., there is nothing special about \( y_i \) having the label \( i \), or appearing in the \( i \)-th position in the sequence). That is, under exchangeability, two sequences, each with the same length \( n \), and the same proportion of ones, would be assigned the same probability. As Diaconis and Freedman (1980a) point out: “...only the number of ones in the...trials matters, not the location of the ones”.

de Finetti’s Theorem takes this implication a step further, showing that if I believe the data are infinitely exchangeable, then it as if there is a parameter \( \theta \) that drives a stochastic model generating the data, and a distribution over \( \theta \) that doesn’t depend on the data. This distribution is interpretable as a prior distribution, since it characterizes beliefs about \( \theta \) that are not conditioned on the data. That is, the existence of a prior distribution over a parameter is a result of de Finetti’s Representation Theorem, rather than an assumption.

We now state this remarkable theorem, referring interested readers elsewhere for a proof.

**Proposition 1.9** (de Finetti Representation Theorem, binary case). If \( y_1, y_2, \ldots \) is an infinitely exchangeable sequence, with \( y_i \in \{0, 1\}, \forall i = 1, 2, \ldots \), then there exists a probability density function \( P \) such that the joint probability mass function for the first \( n \) realizations of \( y_i \), \( p(y_1, \ldots, y_n) \) can be represented as follows,

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

where \( F(\theta) \) is the limiting distribution of \( \theta \), i.e.,

\[
F(\theta) = \lim_{n \to \infty} P(n^{-1} \sum_{i=1}^n y_i \leq \theta).
\]

**Proof.** See de Finetti (1931, 1937), Heath and Sudderth (1976).

**Remark.** Hewitt and Savage (1955) proved the uniqueness of the representation.
Since this theorem is so important to the subjectivist, Bayesian approach adopted here, we pause to examine it in some detail. First, consider the object on the left-hand side of the equality in the proposition. Given that \( y_i \in \{0, 1\} \), \( P(y_1, \ldots, y_n) \) is an assignment of probabilities to all \( 2^n \) possible realizations of \( y = (y_1, \ldots, y_n) \). It is daunting to consider allocating probabilities to all \( 2^n \) realizations, but an implication of de Finetti’s Representation Theorem is that we don’t have to. The proposition shows that probability assignments to \( y_1, \ldots, y_n \) (a finite subset of an infinitely exchangeable sequence) can be made in terms of a single parameter \( \theta \), interpretable as the limiting value of the proportion of ones in the infinite, exchangeable sequence \( y_1, y_2, \ldots \). This is extraordinarily convenient, since under exchangeability, the parameter \( \theta \) can become the object of statistical modeling, rather than much more cumbersome object \( P(y_1, \ldots, y_n) \). Thus, in the subjectivist approach, parameters feature in statistical modeling not necessarily because they are “real” features of the world, but because they are part of a convenient, mathematical representation of probability assignments over data.

Perhaps more surprisingly, di Finetti’s Representation Theorem also implies the existence of a prior probability distribution over \( \theta \), \( F(\theta) \), in the sense that it is a distribution over \( \theta \) that does not depend on the data. If \( F(\theta) \) in proposition 1.9 is absolutely continuous, then we obtain the probability density function for \( \theta \), \( p(\theta) = dF(\theta)/d\theta \). In this case, the identity in the proposition can be re-written as

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

We recognize the first term on the right-hand-side of equation 1.8 as the likelihood for a series of Bernoulli trials, distributed independently conditional on a parameter \( \theta \), i.e., under independence conditional on \( \theta \),

\[
L(\theta; y) \equiv f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)
\]

where

\[
f(y_i|\theta) = \begin{cases} 
  \theta^{y_i} & \text{if } y_i = 1 \\
  (1 - \theta)^{1-y_i} & \text{if } y_i = 0
\end{cases}
\]

The second term in equation 1.8, \( p(\theta) \), is a prior density for \( \theta \). The integration in equation 1.8 is how we obtain the marginal density for \( y \), as a weighted average of the likelihoods implied by different values of \( \theta \in [0, 1] \), where the prior density \( p(\theta) \) supplies the weights.
That is, a simple assumption such as (infinite) exchangeability implies the existence of a parameter \( \theta \) and a prior over \( \theta \), and hence a justification for adopting a Bayesian approach to inference:

This [de Finetti’s Representation Theorem] is one of the most beautiful and important results in modern statistics. Beautiful, because it is so general and yet so simple. Important, because exchangeable sequences arise so often in practice. If there are, and we are sure there will be, readers who find \( p(\theta) \) distasteful, remember it is only as distasteful as exchangeability; and is that unreasonable? (Lindley and Phillips 1976, 115)

1.9.2 Implications and Extensions of de Finetti’s Representation Theorem

The parameter \( \theta \) considered in proposition 1.9 is recognizable a success probability for Bernoulli trials. But other parameters and models can be considered. A simple example comes from switching our focus from the individual zeros and ones to \( S = \sum_{i=1}^{n} y_i \), the number of ones in the sequence \( y = (y_1, \ldots, y_n) \), with possible values \( s \in \{0, 1, \ldots, n\} \).

Since there are \( n! \) ways of obtaining \( S = s \) successes in \( n \) trials, de Finetti’s Representation Theorem implies that probability assignments for \( S \) represented as

\[
\Pr(S = s) = \binom{n}{s} \int_{0}^{1} \theta^s (1 - \theta)^{n-s} dF(\theta).
\]

where \( F(\theta) = \lim_{n \to \infty} \Pr(n^{-1} S \leq \theta) \) is the limiting probability distribution function for \( \theta \). Put differently, conditional on \( \theta \) and \( n \) (the number of trials), the number of successes \( S \) is distributed following the binomial probability point mass function.

A general form of de Finetti’s Representation Theorem exists, and here I re-state a relatively simple version of the general form, due to Smith (1984, 252) and reproduced in Bernardo and Smith (1994, 177):

**Proposition 1.10** (Representation Theorem for Real-Valued Random Quantities). If \( y_n = (y_1, \ldots, y_n) \) are realizations from an infinitely exchangeable sequence, with \( y_i \) real-valued and with probability measure \( P \), then there exists a probability measure \( \mu \) over \( F \), the space of all distribution functions on \( \mathbb{R} \) such that the joint distribution function of \( y_n \) has the representation

\[
P(y_1, \ldots, y_n) = \int_{F} \prod_{i=1}^{n} F(y_i) d\mu(F)
\]

where

\[
\mu(F) = \lim_{n \to \infty} P(F_n)
\]
and where \( F_n \) is the empirical distribution function for \( y \) i.e.,

\[
F_n(y) = n^{-1}[I(y_1 \leq y) + I(y_2 \leq y) + \ldots + I(y_n \leq y)]
\]

where \( I(\cdot) \) is an indicator function evaluating to one if its argument is true and zero otherwise.

**Proof.** See de Finetti (1937; 1938). The result is a special case of the more abstract situation considered by Hewitt and Savage (1955) and Diaconis and Freedman (1980b).  

Note for this general case that the Representation Theorem implies a nonparametric, or, equivalently, a infinitely-dimensional parametric model. That is, the \( F_n \) in Proposition 1.10 is the unknown distribution function for \( y \), a series of asymptotically-diminishing step functions over the range of \( y \). Conditional on this distribution, it is as if we have independent data. The distribution \( \mu \) is equivalent to a prior over what \( F_n \) would look like in a large sample.

Thus, the general version of de Finetti’s Representation Theorem is only so helpful, at least as a practical matter. What typically happens is that the infinite-dimensional \( F_n \) is approximated with a distribution indexed by a finite parameter vector \( \theta \); e.g., consider \( \theta = (\mu, \sigma^2) \), the mean and variance of a normal distribution, respectively. Note the parametric, modeling assumptions being made here. An assumption of exchangeability over real-valued normal quantities is not sufficient to justify a particular parametric model: extra assumptions are being made, at least tacitly. This said, the use of particular parametric models is not completely ad hoc. There is much work outlining the conditions under which exchangeability plus particular invariance assumptions imply particular parametric models (e.g., under what conditions does a belief of exchangeability over real-valued quantities justify a normal model, under what conditions does a belief of exchangeability over positive, integer-valued random quantities justify a geometric model). Bernardo and Smith (1994, Section 4.4) provide a summary of these results.

### 1.9.3 Finite Exchangeability

Note that both Propositions 1.9 and 1.10 rest on an assumption of infinite exchangeability: i.e., that the (finite) data at hand are part of a infinite, exchangeable sequence. de Finetti type theorems do not hold for finitely exchangeable data; see Diaconis (1977) for some simple but powerful examples. This seems problematic, especially in social science settings, where it is often not at all clear that data can be considered to be a subset of
an infinite, exchangeable sequence (we return to this point in Chapter XXXX). Happily, finitely exchangeable sequences can be shown to be approximations of infinitely exchangeable sequences, and so de Finetti type results hold approximately for finitely exchangeable sequences. Diaconis and Freedman (1980b) bound the error induced by this approximation for the general case, in the sense that the de Finetti type representation for \( P(y_n) \) under finite exchangeability differs from the representation obtained under an assumption of infinite exchangeability by a factor that is smaller than a constant times \( 1/n \). Thus, for large \( n \), the “distortion” induced by assuming infinite exchangeability is vanishingly small. A precise definition of this “distortion” and sharp bounds for specific cases are reported in Diaconis and Freedman (1981), Diaconis, Eaton and Lauritzen (1992), and Wood (1992).

1.9.4 Exchangeability and Prediction

Exchangeability also makes clear the close connections between prediction and Bayes Rule, and between parameters and observables.

Consider tossing a coin \( n \) times with the outcomes, \( y \), infinitely exchangeable. Arbitrarily, let \( y_i = 1 \) for a head. We observe \( r \) heads out of \( n \) tosses. Then we consider the next toss of the coin, with the outcome denoted \( y^* \), conditional on the observed sequence of \( r \) heads in \( n \) flips, \( y \).

\[
P(y^* = 1|y) = \frac{P(y^* = 1, y)}{P(y)}
\]

and, by exchangeability,

\[
= \int_0^1 \theta^y (1 - \theta)^n \theta^n (1 - \theta)^{r - y} \theta^r (1 - \theta)^{n - r} \theta^{y^*} (1 - \theta)^{n - y^*} p(\theta)\,d\theta
\]

\[
= \frac{\int_0^1 \theta^y (1 - \theta)^{n - y^*} \theta^{y^*} (1 - \theta)^{n - r} \theta^r (1 - \theta)^{n - y} \theta^{y^*} (1 - \theta)^{n - y^*} p(\theta)\,d\theta}{\int_0^1 \theta^r (1 - \theta)^{n - r} \theta^{y^*} (1 - \theta)^{n - y^*} \theta^{y^*} (1 - \theta)^{n - y^*} \theta^{y^*} (1 - \theta)^{n - y^*} p(\theta)\,d\theta}
\]

since up to a constant multiplicative factor

\[
\mathcal{L}(\theta; y) = \theta^r (1 - \theta)^{n - r}.
\]

But, by Bayes Rule (Proposition 1.5),

\[
f(\theta|y) = \frac{\mathcal{L}(\theta; y)p(\theta)}{\int_0^1 \mathcal{L}(\theta; y)p(\theta)\,d\theta}
\]
and so

\begin{align*}
P(y^* = 1 | y) &= \int_0^1 \theta^{y^*} (1 - \theta)^{1 - y^*} f(\theta | y) d\theta \\
&= \int_0^1 \theta f(\theta | y) d\theta \\
&= E(\theta | y)
\end{align*}

That is, under exchangeability (and via Bayes Rule), beliefs about the outcome of the next realization of the binary sequence corresponds to beliefs about the parameter \( \theta \). Whether \( \theta \) corresponds to anything in the physical world is besides the point. More generally, we rely on this property of modeling under exchangeability quite frequently, with parameters providing an especially useful way to summarize beliefs not only about the data at hand, but future realizations of the (exchangeable) data.

1.9.5 Conditional Exchangeability and Multiparameter Models

Consider again the simple case in Proposition 1.9, where \( y = (y_1, \ldots, y_n) \) is a sequence of zeros and ones. In this case, without any other information about the data, exchangeability seems quite plausible. That is, probability assignments over the data conform to the form given in Proposition 1.9, in which the data are considered independent Bernoulli trials, conditional on the parameter \( \theta \), and \( p(\theta) \) is a prior density over \( \theta \).

But consider a different situation. What if instead of (canonical) coin flips, we had asked survey respondents if they had ever engaged in political protest, for instance, a street demonstration. The data are coded \( y_i = 1 \) if survey respondent \( i \) responds “Yes” and 0 otherwise. But we also know that the data comes from \( J \) different countries: let \( j = 1, \ldots, J \) index the countries covered by the survey, and let \( C_i = j \) if respondent \( i \) is in country \( j \). Suppose for a moment that the country information is given to us only in the most minimal form: a set of integers, i.e., \( C_i \in \{1, \ldots, J\} \). That is, we know that the data come from different countries, but that is all.

Even with this little amount of extra information I suspect most social scientists would not consider the entire sequence of data \( y = (y_1, \ldots, y_n) \) as exchangeable, since there are good reasons to suspect levels of political protest vary considerably by country. We would want to condition any assignment of a zero or a one to \( y_i \) on the country label of case \( i \), \( C_i \).

Within any given country, and absent any other information, the data might be considered exchangeable. Data with this feature are referred to as partially exchangeable or conditionally exchangeable (e.g., Lindley and Novick 1981). In this example, exchangeability
within each country implies that each country’s data can be modeled via country-specific, Bernoulli models: i.e., for \( j = 1, \ldots, J \),

\[
\begin{align*}
y_i | C_i = j & \sim \text{Bernoulli}(\theta_j) \quad \text{(likelihoods)} \\
\theta_j & \sim \text{p}_j(\theta_j) \quad \text{(priors)}
\end{align*}
\]
or, equivalently, since the data are exchangeable within a country, we can model the number of respondents reporting engaging in political protest in a particular country \( r_j \) via a binomial model, conditional on \( \theta_j \) and the number of respondents in that country \( n_j \):

\[
\begin{align*}
r_j | \theta_j, n_j & \sim \text{Binomial}(\theta_j; n_j) \quad \text{(likelihood)} \\
\theta_j & \sim \text{p}_j(\theta_j) \quad \text{(priors)}
\end{align*}
\]

### 1.9.6 Exchangeability of Parameters: Hierarchical Modeling

The idea of exchangeability applies not just to data, but to parameters as well (recall the use of the quite general term “random quantities” rather than “data” in propositions 1.9 and 1.10). Consider the example just given. We know that data span \( J \) different countries. But that is all we know. Under these conditions, the \( \theta_j \) can be considered exchangeable: i.e., absent any information to distinguish the countries from one another, the probability assignment \( p(\theta_1, \ldots, \theta_J) \) is invariant to any change of the labels of the countries (see Definition 1.11). Put simply, the country labels \( j \) do not meaningfully distinguish the countries with respect to their corresponding \( \theta_j \). In this case, de Finetti’s Representation Theorem implies that the joint distribution of the \( \theta_j \) has the following representation

\[
p(\theta) = p(\theta_1, \ldots, \theta_J) = \int \prod_{j=1}^{m} p(\theta_j | \nu) p(\nu) d\nu \quad (1.9)
\]

where \( \nu \) is a hyperparameter. That is, under exchangeability at the level of countries, it is as if we have the following two-stage or hierarchical prior structure over the \( \theta_j \):

\[
\begin{align*}
\theta_j | \nu & \sim p(\theta_j | \nu) \quad \text{(hierarchical model for } \theta_j) \\
\nu & \sim p(\nu) \quad \text{(prior for hyperparameter } \nu)
\end{align*}
\]

For the example under consideration — modeling country-specific proportions — we might employ the following choices for the various distributions:

\[
\begin{align*}
r_j | \theta_j, n_j & \sim \text{Binomial}(\theta_j; n_j) \\
\theta_j | \nu & \sim \text{Beta}(\alpha, \beta) \\
\alpha & \sim \text{Exponential}(2) \\
\beta & \sim \text{Exponential}(2)
\end{align*}
\]
with \( \nu = (\alpha, \beta) \) the hyperparameters for this problem. Details on the specific distributions come later: e.g., in Chapter 2 we discuss models for proportions in some detail. At this stage, the key point is that exchangeability is a concept that applies not only to data, but to parameters as well.

We conclude this brief introduction to hierarchical modeling with an additional extension. If we possess more information about the countries other than case labels, then exchangeability might well be no longer plausible. Just as the information that survey respondents were located in different countries prompted us to revise a belief of exchangeability for them, so too does information allowing us to distinguish countries from one another lead us to revisit the exchangeability judgement over the \( \theta_j \) parameters. In particular, suppose we have variables at the country level, \( x_j \), measuring factors such as the extent to which the country’s constitution guarantees rights to assembly and freedom of expression, and the repressiveness of the current regime. In this case, exchangeability might hold conditional on a unique combination of those country-level predictors. A statistical model that exploits the information in \( x_j \) might be the following:

\[
\begin{align*}
  r_j | \theta_j, n_j & \sim \text{Binomial}(\theta_j; n_j) \\
  z_j & = \log \left( \frac{\theta_j}{1 - \theta_j} \right) \\
  z_j | x_j & \sim N(x_j \beta, \sigma^2) \\
  \beta & \sim N(b, B) \\
  \sigma^2 & \sim \text{Inverse-Gamma}(c, d).
\end{align*}
\]

Again, details on the specific models and distributions deployed here come in later chapters. The key idea is that the information in \( x_j \) enters as the independent variables in a regression model for \( z_j \), the log-odds of each country’s \( \theta_j \). In this way contextual information about country \( j \) is incorporated into a model for the survey responses. These types of exchangeability judgements will play an important role in the discussion of hierarchical models in Chapter 6.

1.10 HISTORICAL NOTE

Bayes’ Theorem is named for the Reverend Thomas Bayes, who died in 1761. The result that we now refer to as Bayes Theorem appeared in an essay attributed to Bayes and communicated to the Royal Society after Bayes’ death by Richard Price (Bayes 1763).
This famous essay has been republished many times since (e.g., Bayes 1958). The subject of Bayes’ *Essay towards solving a problem in the doctrine of chances* was a binomial problem: given $x$ successes in $n$ independent binary trials, what should we infer about $p$, the underlying probability of success?

Bayes himself studied the binomial problem a uniform prior. According to Stigler (1986b), in 1774 Laplace (apparently unaware of Bayes’ work) stated Bayes’ theorem in its more general form, and also considered non-uniform priors. Laplace’s article popularized what would later become known as “Bayesian” statistics. Bayes’ essay “...was ignored until after 1780 and played no important role in scientific debate until the twentieth century” (Stigler 1986b, 361). Additional historical detail can be found in Bernardo and Smith (1994, ch1), Lindley (2001), Seal (2003) and Stigler (1986a, ch3). We return to the relatively simple statistical problem considered by Bayes in Chapter 2.

1.11 SUMMARY

In this chapter...

Problems XXXX
Perhaps the first problem ever considered in Bayesian statistics — and probably the first problem you encountered in your first statistics class — is the problem of learning about a binomial proportion. Bayes opens his famous 1763 *Essay towards solving a problem in the doctrine of chances* with the following statement:

**PROBLEM:**

*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

In modern terms, we recognize this as the problem of trying to learn about the parameter \( \theta \) governing a binomial process, having observed \( r \) successes in \( n \) independent binary trials (today we call these Bernoulli trials). Estimation and inference for the binomial success probability is a fundamental problem in statistics (and a relatively easy one) and underlies many social-science problems, and so provides a convenient context in which to gain exposure to the mechanics of Bayesian inference.
2.1 LEARNING ABOUT PROBABILITIES, RATES AND PROPORTIONS

Binary data are ubiquitous in the social sciences. Examples include micro-level phenomena such as

- voter turnout ($y_i = 1$ if person $i$ turned out to vote, and 0 otherwise), purchasing a specific product, using the Internet, owning a computer;

- life events such as graduating from high school or college, marriage, divorce, having children, home ownership, employment, or going to jail, and even death itself;

- self-reports of attitudes (e.g., approval or disapproval of a policy or a politician; reporting that one is happy or satisfied with one's life, intending to vote for a specific candidate);

- correctly responding to a test item (e.g., $y_{ij} = 1$ if test-taker $i$ correctly answers test item $j$);

- legislative behavior (e.g., $y_{ij} = 1$ if legislator $i$ votes “Aye” on proposal $j$).

Aggregate-level binary data are also common in social-science settings, coding for

- the presence or absence of interstate or civil war,

- whether a country is a democracy or not

- if a congressional district is home to a military base,

- features of a country’s policy of social welfare (universalist or not) or industrial relations regime (centralized bargaining or not)

In general, the observed data consist of a string of binary data: $y_i \in \{0,1\}$, with $i = 1, \ldots, n$. Let $r$ be the number of times $y_i = 1$. In section 1.9 we saw that if binary data can be considered exchangeable, then it is as if the data were generated as $n$ independent Bernoulli trials with unknown parameter $\theta$, i.e.,

$$y_i \overset{iid}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \ldots, n,$$

(2.1)

and with a prior density $p(\theta)$ over $\theta$. Similarly, under exchangeability it is as if

$$r \sim \text{Binomial}(\theta; n)$$

(2.2)
and again \( p(\theta) \) is a prior density over \( \theta \). Note that in either model, \( \theta \) has the interpretation as the proportion of ones we would observed in arbitrarily long, exchangeable sequence of zeros and ones. In most applications involving binary data in the social sciences, a model such as equation 2.1 or 2.2 is adopted without recourse to exchangeability, the researcher presuming that \( \theta \) exists and that the data at hand were generated as independent Bernoulli trials, conditional on \( \theta \). A typical example is the analysis of survey data, where, via random sampling, respondents are considered to be providing binary responses independently of one another, and \( \theta \) is interpreted as a characteristic of the population, the proportion of the population with \( y_i = 1 \).

**EXAMPLE 2.1**

*Attitudes Towards Abortion, Example 1.9, continued.* 895 out of 1,934 survey respondents reported that they thought it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason. Under the assumptions of independence and random sampling (operationally equivalent to exchangeability), we can represent the data with the binomial model in equation 2.2, i.e.,

\[
895 \sim \text{Binomial}(\theta; n = 1, 934)
\]

The corresponding binomial likelihood function is

\[
L(\theta; r = 895, n = 1, 934) = \binom{1,934}{895} \theta^{895} (1 - \theta)^{1,934 - 895}
\]

and the maximum likelihood estimate is \( \hat{\theta}_{\text{MLE}} = 895/1,934 \approx .463 \) with standard error .011.

### 2.1.1 Priors for Probabilities, Rates and Proportions

Any Bayesian analysis requires a specification of the prior. In this case, the parameter \( \theta \) is a probability (or, under exchangeability, a population proportion), and is restricted to the \([0, 1]\) unit interval. Any prior density over \( \theta \), \( p(\theta) \) is thus a function with the following properties:

1. \( p(\theta) \geq 0, \theta \in [0, 1] \).
2. \( \int_0^1 p(\theta)d\theta = 1 \).

Obviously, a large class of functions have these properties: in fact, all the prior densities in Figures 1.2, 1.3 and 1.4 satisfy these conditions.
I initially consider **conjugate priors**: priors with the property that when we apply Bayes Rule — multiplying the prior by the likelihood — the resulting posterior density is of the same parametric type as the prior. Thus, whether a prior is conjugate or not is a property of the prior density with respect to the likelihood function. In this way a conjugate prior is sometimes said to be "closed under sampling", in the sense that after sampling data and modifying the prior via Bayes Rule, our posterior beliefs over $\theta$ can still be characterized with a probability density function in the same parametric class as we employed to characterize our prior beliefs (i.e., the class of probability densities representing beliefs is not expanded by considering the information about $\theta$ in the likelihood).

It is worth stressing that conjugacy is not as restrictive as it might sound. Conjugate priors are capable of characterizing quite a wide array of beliefs. This is especially true if one considers mixtures of conjugate priors (Diaconis and Ylvisaker 1979). In fact, mixtures of conjugate priors were used to generate the odd looking prior distributions in Figures 1.3 and 1.10.

As we now see, Beta densities are conjugate prior densities with respect to a binomial likelihood. The Beta density is quite flexible probability density, taking two shape parameters as its arguments, conventionally denoted with $\alpha$ and $\beta$, and with its support confined to the unit interval. A uniform density on $[0, 1]$ is a special case of the Beta density, arising when $\alpha = \beta = 1$. Symmetric densities with a mode/mean/median at .5 are generated when $\alpha = \beta$ for $\alpha, \beta > 1$. Unimodal densities with positive skew are generated by $\alpha > \beta > 1$; negative skew arises with $\beta > \alpha > 1$. Other features of the Beta density are simple functions of $\alpha$ and $\beta$:

- the mean: \[
\frac{\alpha}{\alpha + \beta}
\]
- the mode: \[
\frac{\alpha - 1}{\alpha + \beta - 2}
\]
- the variance: \[
\frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}
\]

Further details on the Beta density appear in Appendix A. The Beta density is

\[
p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},
\]  

(2.3)

where $\theta \in [0, 1]$, $\alpha, \beta > 0$ and $\Gamma(\cdot)$ is the Gamma function (again, see Appendix A). Note that the leading terms involving the Gamma functions do not involve $\theta$, and so equation 2.3 can be rewritten as

\[
p(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.
\]  

(2.4)
This prior is conjugate for $\theta$ with respect to the binomial likelihood in equation 2.2:

**Proposition 2.1 (Conjugacy of Beta Prior, Binomial Data).** Given a binomial likelihood over $r$ successes in $n$ Bernoulli trials, each independent conditional on an unknown success parameter $\theta \in [0, 1]$, i.e.,

$$L(\theta; r, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

then the prior density $p(\theta) = \text{Beta}(\alpha, \beta)$ is conjugate with respect to the binomial likelihood, generating the posterior density $p(\theta|r, n) = \text{Beta}(\alpha + r, \beta + n - r)$.

**Proof.** By Bayes Rule,

$$p(\theta|r, n) \propto \frac{L(\theta; r, n)p(\theta)}{\int_{0}^{1} L(\theta; r, n)p(\theta) d\theta} \propto L(\theta; r, n) p(\theta)$$

The binomial coefficient $\binom{n}{r}$ in the likelihood does not depend on $\theta$ and can be absorbed into the constant of proportionality. Similarly, we use the form of the Beta prior given in equation 2.4. Thus

$$p(\theta|r, n) \propto \frac{\theta^r (1 - \theta)^{n-r}}{\Gamma(\alpha + \beta)} \times \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$$

which we recognize as the kernel of a Beta distribution. That is, $p(\theta|r, n) = c\theta^{r+\alpha-1}(1 - \theta)^{n-r+\beta-1}$ where $c$ is the normalizing constant

$$c = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(r + \alpha)\Gamma(n - r + \beta)}$$

i.e., $\int_{0}^{1} \theta^{r+\alpha-1}(1 - \theta)^{n-r+\beta-1} = c^{-1}$, see Definition A.10. In other words,

$$\theta|r, n \sim \text{Beta}(\alpha + r, \beta + n - r) \quad (2.5)$$

As is the case for most conjugate priors, the form of the posterior density has a simple interpretation. It is as if our prior distribution represents the information in a sample of $\alpha + \beta - 2$ independent Bernoulli trials, in which we observed $\alpha - 1$ “successes” or $y_i = 1$. This interpretation of the parameters in the conjugate Beta prior in “data equivalent terms” makes it reasonably straightforward to specify conjugate priors for probabilities, rates and proportions.
EXAMPLE 2.2

Attitudes towards Abortion, Example 1.9, continued. A researcher has no prior information regarding \( \theta \), the proportion of the population agreeing with the proposition that it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason. Any value of \( \theta \in [0, 1] \) is \textit{a priori} as likely as any other, implying the researcher’s prior is a uniform distribution. The uniform distribution on the unit interval is a special case of the Beta distribution with \( \alpha = \beta = 1 \) (i.e., corresponding to possessing a prior set of observations of zero successes in zero trials). Recall that in these data \( r = 895 \) and \( n = 1,934 \). Then, applying the result in Proposition 2.1, the posterior density for \( \theta \) is

\[
p(\theta | r = 895, n = 1, 934) = \text{Beta}(\alpha^*, \beta^*),
\]

where

\[
\alpha^* = \alpha + r = 1 + 895 = 896
\]
\[
\beta^* = \beta + n - r = 1 + 1,934 - 895 = 1,040
\]

Given the “flat” or “uninformative” prior, the location of the the mode of the posterior density corresponds \textit{exactly} with the maximum likelihood estimate of \( \theta \): i.e.,

\[
\text{posterior mode of } p(\theta | r, n) = \frac{\alpha^* - 1}{\alpha^* + \beta^* - 2} = \frac{896 - 1}{896 + 1,040 - 2} = \frac{895}{1,934} = \frac{r}{n} = .463 = \hat{\theta}_{\text{MLE}}
\]

Moreover, when \( \alpha^* \) and \( \beta^* \) are large (as they in this example), the corresponding Beta density is almost exactly symmetric and so the mode is almost exactly the same as the mean and the median. That is, because of the near-perfect symmetry of the posterior density, Bayes estimates of \( \theta \) such as the posterior mean, median or mode (considered in section 1.6.1) almost exactly coincide. In addition, with the “flat”/uniform prior employed in this example, these Bayes estimate of \( \theta \) all almost exactly coincide with the maximum likelihood estimate. For instance, the mean of the posterior density is at \( 896/(896 + 1040) \approx .463 \), which corresponds to the maximum likelihood estimate to three decimal places.
The posterior variance of $\theta$ is

$$V(\theta| r, n) = \frac{\alpha^* \beta^*}{(\alpha^* + \beta^*)^2 (\alpha^* + \beta^* + 1)}$$

$$= \frac{896 \times 1, 040}{(896 + 1, 040)^2 \times (896 + 1, 040 + 1)}$$

$$= \frac{931, 840}{3, 748, 096 \times 1, 937} \approx 0.000128,$$

and so the posterior standard deviation of $\theta$ is $\sqrt{0.000128} = 0.0113$, corresponding to the standard error of the maximum likelihood estimate of $\theta$.

Finally, note that with $\alpha^*$ and $\beta^*$ both large (as they are in this example), the Beta density is almost exactly symmetric, and so a very accurate approximation to a 95% highest posterior density region for $\theta$ can be made by simply computing the 2.5 and 97.5 percentiles of the posterior density, the Beta($\alpha^*$, $\beta^*$) density. With $\alpha^* = 896$ and $\beta^* = 1, 040$, these percentiles bound the interval [.441, .485]. Computing the 95% HPD directly shows this approximation to be very good in this case: the exact 95% HPD corresponds to the approximation based on the percentiles, at least out to 4 significant digits.

Of course, the results of a Bayesian analysis will not coincide with a frequentist analysis if the prior density is informative about $\theta$, relative to the information in the data about $\theta$. This can happen even priors over $\theta$ are “flat” or nearly so, if the data set is small and/or relatively uninformative about $\theta$. Alternatively, even with a large data set, if prior information about $\theta$ is plentiful, then the posterior density will reflect this, and will not be proportional to the likelihood alone.

EXAMPLE 2.3

Combining Information for Improved Election Forecasting. In early March of 2000, Mason-Dixon Polling and Research conducted a poll of voting intentions in Florida for the November presidential election. The poll considered George W. Bush and Al Gore as the presumptive nominees of their respective political parties. The poll had a sample size of 621, and the following breakdown of reported vote intentions: Bush 45% ($n = 279$), Gore 37% (230), Buchanan 3% (19) and undecided 15% (93). For simplicity, we ignore the undecided and Buchanan vote share, leaving Bush with 55% of the two-party vote intentions, and Gore with 45%, and $n = 509$ respondents expressing a preference for the two major party candidates. We assume that the survey responses are independent, and (perhaps unrealistically) that the sample is a random
sample of Floridian voters. Then, if $\theta$ is the proportion of Floridian voters expressing a preference for Bush, the likelihood for $\theta$ given the data is (ignoring constants that do not depend on $\theta$),

$$
L(\theta; r = 279, n = 509) \propto \theta^{279} (1 - \theta)^{509-279}.
$$

The maximum likelihood estimate of $\theta$ is $\hat{\theta}_{\text{MLE}} = r/n = 279/509 = .548$ with a standard error of $\sqrt{(.548 \times (1-.548))/509} = .022$. Put differently, this poll provides very strong evidence to suggest that Bush was leading Gore in Florida at that relatively early stage of the 2000 presidential race.

But how realistic is this early poll result? Is there other information available that bears on the election result? Previous presidential elections are an obvious source of information. Even a casual glance at state-level presidential election returns suggests considerable variation between states, or, put differently, one can do a fairly job at predicting state-level election outcomes by looking at previous elections in that state, say, by taking a weighted average of previous election outcomes via regression analysis. This procedure is then used to generate a forecast for Republican presidential vote share in Florida in 2000. This prediction is 49.1%, with a standard error of 2.2 percentage points.

We can combine the information yielded by this analysis of previous elections with the survey via Bayes’ Theorem. We can consider the prediction from the analysis as supplying a prior, and the survey as “data”, although mathematically, it doesn’t matter what label we give to each piece of information.

To apply Bayes’ Theorem in this case, and to retain conjugacy, I first characterize the information from the regression forecast as a Beta distribution. That is, we seek values for $\alpha$ and $\beta$ such that

$$
E(\theta; \alpha, \beta) = \frac{\alpha}{\alpha + \beta} = .491
$$

$$
V(\theta; \alpha, \beta) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = .022^2
$$

which yields a system of equations in two unknowns. Leonard and Hsu (1999, 100) suggest the following strategy for solving for $\alpha$ and $\beta$. First define $\theta_0 = E(\theta; \alpha, \beta) = \alpha/(\alpha + \beta)$ and then note that $V(\theta; \alpha, \beta) = \theta_0 (1 - \theta_0)/(\gamma + 1)$, with $\gamma = \alpha + \beta$. 
interpretable as the size of a hypothetical, prior sample. Then,

\[ \gamma = \frac{\theta_0(1 - \theta_0)}{V(\theta; \alpha, \beta)} - 1 \]

\[ = \frac{.491 \times (1 - .491)}{.022^2} - 1 = 515.36, \]

\[ \alpha = \gamma \theta_0 = 515.36 \times .491 = 253.04, \]

\[ \beta = \gamma (1 - \theta_0) = 515.36 \times (1 - .491) = 262.32. \]

That is, the information in the previous elections is equivalent to having ran another poll with \( n \approx 515 \) in which \( r \approx 253 \) respondents said they would vote for the Republican presidential candidate.

Now, with a conjugate prior for \( \theta \) it is straightforward to apply Bayes Rule. The posterior density for \( \theta \) is a Beta density with parameters \( \alpha^* = 279 + 253.04 = 532.04 \) and \( \beta^* = 509 - 279 + 262.32 = 492.32 \). With \( \alpha^* \) and \( \beta^* \) both large, this Beta density is almost exactly symmetric, with the mean, mode and median all equal to .519 (to three significant digits). A ninety-five percent HPD is extremely well approximated by the 2.5 and 97.5 percentiles of the Beta density, which bound the interval [.489, .550]. Figure 2.1 displays the prior, the posterior and the likelihood (re-scaled so as to be comparable to the prior and posterior densities). The posterior density clearly represents a compromise between the two sources of information brought to bear in this instance, with most of the posterior probability mass lying between the mode of the prior density and the mode of the likelihood function. It is also clear that at least in this case, the variance of the posterior density is smaller than either the variance of the prior or the dispersion of the likelihood function. This reflects the fact that in this example the prior and the likelihood are not in wild disagreement with one another, and the posterior density “borrows strength” from each source of information.

Other features of the posterior density are worth reporting in this case. For instance, given the combination of the survey data with the information supplied by the historical election data, the probability that Bush would defeat Gore in Florida is
Figure 2.1  Florida Election Polling. The shaded area under the posterior density represents the posterior probability that Bush leads Gore, and is .893.
simply the posterior probability that $\theta > 1/2$, i.e.,

$$P(\theta > 1 - \theta | r, n, \alpha, \beta) = P(\theta > 0.5 | r, n, \alpha, \beta)$$

$$= \int_{0.5}^{1} p(\theta | r, n, \alpha, \beta) d\theta$$

$$= \int_{0.5}^{1} p(\theta; \alpha^* = 532.04, \beta^* = 492.32) d\theta$$

$$= 1 - \int_{0}^{0.5} p(\theta; \alpha^* = 532.04, \beta^* = 492.32) d\theta.$$

This probability is .893, suggesting that the probability of a Bush victory was still quite high, but not as high as near-certainty implied by the survey estimate.

**EXAMPLE 2.4**

*Will the Sun Rise Tomorrow?* In his *Essai philosopohique sur les probabilités*, Laplace considered the following problem: “If we place the dawn of history at 5,000 years before the present date, then we have 1,826,213 days on which the sun has constantly risen in each 24 hour period” (Laplace 1825, 11). Given this evidence, what is the probability that the sun will rise tomorrow? Bayes also considered this problem in his famous posthumous *Essay* (1763). Bayes and Laplace computed their solution to this problem supposing (in Bayes’ words), “a previous total ignorance of nature”. Bayes (or at least Richard Price, who edited and communicated Bayes’ essay to the *Royal Society*) imagined “a person just brought forth into this world”, who, after witnessing sunrise after sunrise is gradually revising their assessment that the sun will rise tomorrow:

this odds [of the sun rising] would increase, ... with the number of returns to which he was witness. But no finite number of returns would be sufficient to produce absolute or physical certainty (Bayes 1958, 313).

Laplace’s solution to this problem is known as Laplace’s “rule of succession”, and is one of the first, correct Bayesian results to be widely published.

In a more modern statistical language, both Bayes and Laplace considered the problem of what to believe about a binomial success probability $\theta$, given $n$ successes
in $n$ trials, and a uniform prior on $\theta$: i.e.,

$$L(\theta; n) = \binom{n}{\theta} \theta^n (1 - \theta)^{n-n} = \theta^n \quad \text{(likelihood)}$$

$$p(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{(prior)}$$

By Bayes Rule, the posterior distribution for $\theta$ is

$$p(\theta|n) \propto L(\theta; n) p(\theta) \propto \theta^n$$

which we recognize as the kernel of a Beta($\alpha^*$, $\beta^*$) distribution with $\alpha^* = n + 1$ and $\beta = 1$; i.e., $\theta|n \sim \text{Beta}(n + 1, 1)$. This posterior distribution has a mode at $\theta = 1$, but the posterior mean is

$$E(\theta|n \text{ success in } n \text{ trials}) = \frac{n + 1}{n + 2} = 1 - \frac{1}{n + 2}$$

As we saw in section 1.9.4, exchangeability and Bayes Rule imply that

$$P(y_{n+1} = 1 | \sum_{i=1}^{n} y_i = r) = E(\theta| \sum_{i=1}^{n} y_i = r),$$

for any $r = 0, 1, \ldots, n$. Here $r = n$ and Laplace considered $n = 1, 826, 213$. Thus, the probability of the sun rising tomorrow, conditional on a uniform prior over this probability and 5,000 years of exchangeable, daily sunrises, is $1 - 1.826215 \times 10^{-6}$ or $0.999998173785$. That is, it is not an absolute certainty that the sun will rise tomorrow, nor should it be.

Remark. This example highlights some interesting differences between the frequentist and Bayesian approaches to inference. I paraphrase Diaconis (2005) on this result:

if this seems wild to you, compare it with the frequentist estimate of $\theta$ given $n$ successes in $n$ trials. It is $\hat{\theta} = 1$ (even if $n = 2$).

That is, suppose we flip a coin twice. It comes up heads both times. The maximum likelihood estimate of $\theta$ is 1.0, which is also the frequentist estimate of the probability of a head on the next toss. With a uniform prior over $\theta$, the Bayesian estimate of the probability of a head on the next toss is $1 - 1/4 = .75$. It is an interesting exercise to see if in these circumstances a frequentist would be prepared to make a wager based on the estimate $\hat{\theta} = 1$. 
Remark. Laplace fully understood his example was contrived, ignoring relevant knowledge such as orbital mechanics in the solar system, the life-cycles of stars, and so on. Immediately after concluding that under the stipulated conditions of a uniform prior over $\theta$ and 5,000 years of sun rises “we may lay odds of 1,826,214 to 1 that it [the sun] will rise again tomorrow”, Laplace writes

But this number would be incomparably greater for one who, perceiving in the coherence of phenomena the principle regulating days and seasons, sees that nothing at the present moment can check the sun’s course (Laplace 1825, 11)

Further background on the historical and scientific background surrounding Laplace’s use of this example and his solution is provided in two essays by Zabell (1988; 1989b). The latter essay makes clear the links between Laplace’s rule of succession and exchangeability (see Section 1.9).

2.1.2 Bayes Estimates as Convex Combinations of Priors and Data

Proposition 2.1 shows that with a Beta($\alpha, \beta$) prior for a success probability $\theta$ and a binomial likelihood over the $r$ successes out of $n$ trials, the posterior density for $\theta$ is a Beta density with parameters $\alpha^* = \alpha + r$ and $\beta^* = \beta + n - r$. We have also seen that under these conditions, the posterior mean of $\theta$ is

$$E(\theta| r, n) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\frac{\alpha + r}{\alpha + \beta + n}}{\frac{n_0 \theta_0 + n \hat{\theta}}{n_0 + n}}$$

(2.6)

where $\hat{\theta} = r/n$ is the maximum likelihood estimate of $\theta$, $n_0 = \alpha + \beta$ and $\theta_0 = \alpha/(\alpha + \beta) = E(\theta)$ is the mean of the prior density for $\theta$. Equation 2.6 can be re-written as

$$E(\theta| r, n) = \gamma \theta_0 + (1 - \gamma) \hat{\theta}$$

where $\gamma = n_0/(n_0 + n)$, and since $n_0, n > 0$, $\gamma \in [0, 1]$. Alternatively,

$$E(\theta| r, n) = \hat{\theta} + \gamma (\theta_0 - \hat{\theta})$$

These simple, linear form of these three expressions reflect the fact that with a conjugate Beta prior, the posterior mean — a Bayes estimate of $\theta$ — is a weighted average of the prior mean $\theta_0$ and the maximum likelihood estimate $\hat{\theta}$. This means that with conjugate priors, the posterior mean for $\theta$ will be a convex combination of the prior mean and the maximum likelihood estimate, and so lie somewhere between the prior mean and the value supported by the data. In this way, the Bayes estimate is sometimes said to shrink the posterior towards...
the prior, with $0 < \gamma < 1$ being the shrinkage factor. With a increasingly precise set of prior beliefs — a prior based on an increasing number of hypothetical prior observations — the posterior mean will be shifted increasingly closer to the prior mean. On the other hand, as the precision of the prior information degrades, and the number of hypothetical number of prior observations, $n$ is large relative to $n_0$, and $\hat{\theta}$ dominates the weighted average in equation 2.6.

**EXAMPLE 2.5**

*Combining Information for Improved Election Forecasting, Example 2.3 continued.* Previous election results suggested a prior density over $\theta$, George W. Bush’s share of the two-party vote for president in Florida in 2000, represented by a Beta(253.04, 262.32) density. But it is reasonable to consider other priors. In particular, one might think that while the previous election results supply a reasonable “best guess” as to $\theta$, there is actually quite a lot of uncertainty as to the relevance of previous elections for the problem of making a forecast for the current election. Put differently, it might be reasonable to doubt that election outcomes in Florida are exchangeable. In fact, it could well be that precisely because Florida was expected to be a pivotal state in the 2000 election (if not the pivotal state), the campaigns devote a disproportionate amount of resources there, changing the nature of the election there, such that the 2000 election is “special” and no longer exchangeable with past elections. In this case we might consider a range of priors over $\theta$, reflecting more or less confidence in the prior over $\theta$ derived from previous elections.

The top panel of Figure 2.2 shows how the location of the posterior mean, $E(\theta|r, n)$ changes over more or less stringent versions of the prior derived from previous elections. The parameter $\lambda$ governs the precision of the prior, playing the role of a user-adjustable hyperparameter. That is, the prior is $\theta \sim \text{Beta}(253.04/\lambda, 262.32/\lambda)$ with $\lambda \in [0, \infty)$. For all values of $\lambda$, the prior mean remains constant, i.e., $E(\theta; \lambda) = (\alpha/\lambda)/(\alpha/\lambda + \beta/\lambda) = \alpha/(\alpha + \beta)$. But as $\lambda \to 0$, the precision of the prior is increasing, since $n_0 = (\alpha + \beta)/\lambda$ is growing large; conversely, as $\lambda \to \infty$, $n_0 \to 0$ and the prior precision decreases. In turn, since equation 2.6 shows that the mean of the posterior density for $\theta$ has the weighted average form

$$E(\theta|r, n) = \frac{n\hat{\theta} + n_0\theta_0}{n + n_0},$$
it follows that

\begin{align*}
\lim_{\lambda \to \infty} E(\theta | r, n) &= \hat{\theta} \\
\lim_{\lambda \to 0} E(\theta | r, n) &= \theta_0
\end{align*}

In the top panel of Figure 2.2 \( \lambda \) increases from left to right, on the log scale. On the left, as \( \lambda \to 0 \), the posterior is completely dominated by the prior; on the right of the graph, as \( \lambda \to \infty \), the prior receives almost zero weight, and the posterior mean is almost exactly to the maximum likelihood estimate \( \hat{\theta} \).

The top panel of Figure 2.2 also displays a 95% HPD around the locus of posterior means. As \( \lambda \to 0 \), the 95% HPD collapses around the prior mean, since the prior is becoming infinitely precise, and the posterior tends towards a degenerate density with point mass on \( \theta_0 \). On the other hand, as \( \lambda \to \infty \), the prior has virtually no impact on the posterior, and the posterior 95% HPD widens, to correspond to a classical 95% confidence interval around \( \hat{\theta} \).

The lower panel of Figure 2.2 shows the posterior probability that Bush wins Florida in 2000 (the posterior probability that \( \theta > .5 \)). Clearly, as the posterior density is shrunk towards \( \theta_0 \) as \( \lambda \to 0 \), this posterior probability can be expected to vary. The survey data alone suggest is quite high: the 95% HPD for \( \theta \) with \( \lambda \to \infty \) does not overlap .5 (top panel), and the posterior probability that \( \theta > .5 \) is almost 1.0 (lower panel). With the prior suggested by the historical election data (\( \lambda = 1 \)), the posterior probability of Bush winning Florida has fallen below the conventional 95% level of confidence, but Bush winning Florida is over 4 times as likely as the alternative outcome. With \( \lambda \approx .19 \), \( E(\theta | r, n) \approx .5 \) and the posterior probability that \( \theta > .5 \) falls to around .5. For values of \( \lambda < .19 \), the prior is sufficiently stringent that the posterior probability that Bush carries Florida is below .5. With \( \lambda \approx .19 \), \( n_0 \approx 2,712 \), or equivalently, it is as if the information in the historical election data is as precise as that obtained from a survey of around 2,700 respondents. That is, we would have to place considerable weight in the historical election data in order to contradict the implication of the survey data that Bush would defeat Gore in Florida in 2000.

As it turned out, Bush won 50% of the two-party vote in Florida, in famously controversial circumstances, and Florida was pivotal in deciding the 2000 presidential election. Bush’s 50% performance in Florida in 2000 is only one percentage point above the level implied by the prior, well within a prior 95% HPD and the posterior...
95% HPD generated using the historical data as a prior (i.e., with $\lambda = 1$), and outside the 95% HPD implied by the February survey data. With the benefit of hindsight (!), it would seem that combining the February survey data with the prior information supplied by historical election data was prudent. Bayes Rule tell us how to perform that combination of information.

This relatively simple, weighted average form for the mean of the posterior density is not unique to the use of conjugate Beta prior and a binomial likelihood. As we shall see below, we obtain similar weighted average forms for the mean of the posterior density in other contexts, when employing prior densities that are conjugate with respect to a likelihood. In fact, the weighted average form of the mean of the posterior mean is a general feature of conjugacy, and indeed, as been proposed as a definition of conjugacy by Diaconis and Ylvisaker (1979).

### 2.1.3 Parameterizations and Priors

In the examples considered above, prior ignorance about $\theta$ has been expressed via a uniform prior over the unit interval, which is actually a conjugate Beta prior with respect to the binomial likelihood. That is, absent any knowledge to the contrary, any value of $\theta$ should be considered as likely as any other. Bayes and Laplace employed this prior. Bayes motivated his 1763 Essay with reference to the location of a ball thrown onto a square table or a bounded plane, for which “there shall be the same probability that it rests upon any one equal part of the plane as another...”. In this case, and in many others, Bayes thought this postulate of prior ignorance appropriate:

...in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration; viz. that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another.

The more general principle — in the absence of a priori information, give equal weight to all possible outcomes — is known as the principle of insufficient reason, and is usually attributed to Jakob Bernoulli in Ars conjectandi, but may well be due to Leibniz (Hacking 1975, ch14). Laplace embraced this postulate, apparently accepting it as “an intuitively obvious axiom” (Stigler 1986b, 359), whereas Bayes offered a quite elaborate justification of what we today would call a uniform prior; of course, Laplace also considered non-uniform priors.
Figure 2.2  Posterior Mean as a Convex Combination of Prior and Data for binomial success parameter, \( \theta \). In the top panel, the solid line traces out a series of posterior means, \( E(\theta|r, n) \), generated using the prior, \( p(\theta; \alpha, \beta, \lambda) = \text{Beta}(\alpha/\lambda, \beta/\lambda) \). The shaded area shows the width of a 95% HPD for \( \theta \), also varying as a function of \( \lambda \). The lower panel shows the predicted probability of a Bush victory in Florida, the posterior probability that \( \theta > .5 \), as a function of \( \lambda \).
For the simple case of a parameter taking on values in a closed parameter space, e.g., \( \theta \in [0, 1] \), the principle of insufficient reason and the implied uniform prior poses no substantive or technical difficulty. But for the case where \(-\infty < \theta < \infty\), a uniform prior is no longer a proper probability density, and the simple Bayesian mechanics encountered thus far get a little more complicated (as we shall see in section XXXX).

Even the simple binomial problem considered above can become complicated if we consider a reparameterization. That is, suppose we are interested not in the success probability \( \theta \), but a parameter that is a function of \( \theta \), say \( g(\theta) \). R. A. Fisher (1922) famously pointed to a shortcoming of the simple use of uniform priors over \( \theta \) in this case. With \( n \) trials and \( r \) successes, the maximum likelihood estimate is \( \hat{\theta} = r/n \). But, Fisher notes, “we might equally have measured probability upon an entirely different scale” (Fisher 1922, 325), and considers the re-parameterization \( \sin q = 2\theta - 1 \). By the invariance property of maximum likelihood (see definition XXXX), the MLE of \( q \) corresponds to the MLE of \( \theta \), once we map \( \hat{q}_{\text{MLE}} \) back to \( \theta \), as we now demonstrate. Assuming conditional independence of the observations, the likelihood function is (omitting constants that do not depend on \( q \))

\[
L(q; r, n) = (1 + \sin q)^r (1 - \sin q)^{n-r}
\]

with log-likelihood

\[
\log L(q; r, n) = r \log(1 + \sin q) + (n - r) \log(1 - \sin q),
\]

and, after differentiating with respect to \( q \), the MLE is given by solving

\[
x \cos q = \frac{(n - r) \cos q}{1 + \sin q} = \frac{(n - r) \cos q}{1 - \sin q},
\]

yielding \( \sin q = 2(r/n) - 1 \) as the MLE. This implies that even if we maximize the likelihood with respect to \( q \), we still find that \( \hat{\theta}_{\text{MLE}} = r/n \), the same result that we get from solving for the MLE of \( \theta \) directly.

Fisher noted that in a Bayesian analysis of this problem, a uniform prior over \( q \) did not imply a uniform prior over a parameter such as \( \theta \), and that two sets of Bayes estimates for \( \theta \) could be obtained, depending on which “uniform prior” one adopted. That is, the results of the Bayesian analysis are sensitive to choice of parameterization and prior in a way that a maximum likelihood analysis is not. In Example 1.1 I show that the model

\[
\begin{align*}
q & \sim \text{Uniform} \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \\
\theta &= (1 + \sin q)/2 \\
r & \sim \text{Binomial}(\theta; n)
\end{align*}
\]
implies that the posterior for $\theta$ is a Beta($r + \frac{1}{2}, n - r + \frac{1}{2}$) density, with a mode at $(r - \frac{1}{2})/(n - 1)$. This differs from the maximum likelihood estimate of $r/n$ by an amount that vanishes as $n \to \infty$. On the other hand, a uniform prior on $\theta$ yields a posterior density that has the same shape as the likelihood function.

Using the results in Proposition A.2, the uniform prior on $q \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ induces a Beta($\frac{1}{2}, \frac{1}{2}$) prior density on $\theta$, which does not integrate to one over the unit interval since the density is infinite at zero and one. As Figure 2.3 displays, the Beta($\frac{1}{2}, \frac{1}{2}$) density is somewhat odd, in that it assigns high probability to relatively extreme values of $\theta$: e.g., a 50% highest density region for $\theta \sim$ Beta($\frac{1}{2}, \frac{1}{2}$) is the set of disjoint intervals [{0, .15}, [.85, 1}]. Nonetheless, this is the prior density that arises when specifying a uniform prior over a (non-linear) transformation of $\theta$.

We shall return to this prior density in section XXXX, where we will see that this prior density has a special interpretation. For now, the general point is to remind ourselves that the prior is vitally important in a Bayesian analysis, and that expressing prior ignorance may not be as straightforward as one might think. A prior density that is uninformative with respect to one parameter or parameterization may not be uninformative with respect to some other parameter of interest. What to do? For one thing, if priors really are expressing prior beliefs, then analysts need to decide precisely what is the object under study. If the binomial success probability, $\theta$, is an object of interest and for which it is easy to mathematically express one’s prior beliefs, then one should go ahead and use that prior. But we should not be surprised, let alone dismayed, that, say, a uniform prior on $\theta$ implies an unusual or surprising distribution over some parameter $g(\theta)$, or vice-versa. Likewise, if interest focuses on $g(\theta)$, then the analyst should formulate a prior with respect to that parameter.

2.1.4 The Variance of the Posterior Density

With a Beta($\alpha, \beta$) prior over $\theta$, and a binomial likelihood for $r$ successes in $n$ Bernoulli trials (independent conditional on $\theta$), the variance of the posterior density for $\theta$ is

$$V(\theta|r, n) = \frac{(\alpha + r)(\beta + n - r)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

It is not always the case that the posterior variance is smaller than the prior variance

$$V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$
Figure 2.3  The Beta($\frac{1}{2}, \frac{1}{2}$) density.
starkly contradict one another, then it is possible for the posterior uncertainty to exceed
the prior uncertainty. On the other hand, when the prior and the likelihood are in relative
agreement, then the posterior density will be less dispersed than the prior, reflecting the
gain in information about $\theta$ from the data.

**EXAMPLE 2.6**

**Will I be Robbed at Stanford?** Laura has just arrived at Stanford University for
graduate study. She guesses the probability that she will be the victim of property theft
while on campus during the course of her studies to be 5%; i.e., $E(\theta) = 0.05$. Laura
has heard lots of good things about Stanford: e.g., Stanford is a private university
surrounded by relatively affluent neighborhoods, with a high proportion of graduate
students to undergraduates. These factors not only lead Laura to make a low *a priori*
estimate of $\theta$, but to make Laura reasonably confident in her guess. While Laura’s
*a priori* probability estimate is subject to uncertainty, she figures that the probability
of her falling victim to theft is no greater than 10%, but almost surely no lower
than 1%. These prior beliefs are well approximated by a Beta distribution over $\theta$,
with its mean at .05, and with $Pr(\theta > .10) \approx .01$ and $Pr(\theta < .01) \approx .01$. That
is, if $\theta \sim Beta(\alpha, \beta)$, then Laura’s belief that $E(\theta) = .05$ implies the constraint
$\alpha/(\alpha + \beta) = .05$ or $\beta = 19\alpha$. Further, the additional constraints that $Pr(\theta > .10) \approx
.01$ and $Pr(\theta < .01) \approx .01$ imply that $\alpha \approx 7.4$ and $\beta \approx 140.6$. See Section XXXX
for details on the calculation of $\alpha$ and $\beta$.

Over the course of conversations with twenty other students (more or less randomly
selected), Laura raises the topic of theft on campus. It transpires that twelve of
the twenty students she speaks to report having been the victim of theft while at
Stanford. Laura is reasonably satisfied that the experiences of the twenty students
are independent of one another (i.e., the twenty students constitute are part of an
exchangeable sequence of Stanford students). Accordingly, Laura summarizes the
new information with a binomial likelihood for the $r = 12$ “successes” in $n = 20$
trials, considered independent conditional on $\theta$. Via Bayes Rule, Laura updates her
beliefs about $\theta$, yielding the posterior density

$$\theta | r, n \sim Beta(\alpha + r, \beta + n - r) = Beta(19.4, 148.6)$$

Laura’s prior “best guess” (prior mean for $\theta$) of .05 is revised upwards in light of the
fact that 60% of the 20 students she spoke to reported being a victim of theft. Laura’s
*a posteriori* estimate of \(\theta\) (the mean of the posterior density) is \(19.4/(19.4 + 148.6) \approx .12\).

But what is also interesting is what has happened to the *dispersion* or *precision* of Laura’s beliefs. The variance of Laura’s prior beliefs was

\[
V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \approx \frac{7.4 \times 140.6}{(7.4 + 140.6)^2(7.4 + 140.6 + 1)} \approx .00032
\]

but the variance of the posterior density over \(\theta\) is

\[
V(\theta|r, n) = \frac{(\alpha + r)(\beta + n - r)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \approx \frac{(7.4 + 12) \times (140.6 + 8)}{(7.4 + 140.6 + 20)^2(7.4 + 140.6 + 20 + 1)} \approx .00060,
\]

or roughly twice the size of the prior variance. That is, while Laura’s has revised her estimate of \(\theta\) upwards, she is in some senses more uncertain about \(\theta\) than she was previously. The 95% highest density region for Laura’s prior density is \([.018, .086]\], spanning .068, while the corresponding region for the posterior density is \([.069, .164]\], spanning .095.

The previous example demonstrates that in some circumstances posterior uncertainty can be larger than prior uncertainty. But this result does not hold if we work with an alternative parameterization. In particular, for conjugate analysis of binomial problems, if the analysis is performed with respect to the log of the odds ratio

\[
\Lambda = \logit(\theta) = \ln\left(\frac{\theta}{1 - \theta}\right)
\]

then posterior density of \(\Lambda\) has the property that it always has variance no greater than the variance of the prior density for \(\Lambda\). If \(\theta \sim \text{Beta}(\alpha, \beta)\), the using the change of variables result in Proposition A.2,

\[
f(\Lambda) = f_\theta\left(\frac{\exp(\Lambda)}{1 + \exp(\Lambda)}; \alpha, \beta\right) \frac{\exp(\Lambda)}{(1 + \exp(\Lambda))^2}
\]

where \(f_\theta\) is a Beta\((\alpha, \beta)\) density. For even moderately-sized \(\alpha\) and \(\beta\) (say, \(\alpha, \beta \geq 5\)), this density is extremely well-approximated by a normal distribution with mean \(\log(\alpha/\beta)\) and variance \(\alpha^{-1} + \beta^{-1}\); see Lindley (1964, 1965) and Bloch and Watson (1967). Since the
variance of \( f(\Lambda) \) is approximately \( \alpha^{-1} + \beta^{-1} \), it is obvious that the posterior variance will be smaller than the prior variance; i.e., if \( \theta \sim \text{Beta}(\alpha, \beta) \) and \( \theta | r, n \sim \text{Beta}(\alpha + r, \beta + n - r) \), then \( V(\Lambda | r, n) \approx (\alpha + r)^{-1} + (\beta + n - r)^{-1} < V(\Lambda) \approx \alpha^{-1} + \beta^{-1} \), since for any data set \( r \geq 0 \) and \( n > 0 \).

**EXAMPLE 2.7**

*Will I be Robbed at Stanford? (Example 2.6, continued)* Figure 2.4 shows Laura’s prior and posterior densities, on both the unit probability interval, and on the log-odds scale. Visual inspection confirms that on the log-odds scale, the posterior variance is smaller than the prior variance. Using the Lindley (1965) normal approximation to \( f(\Lambda) \), we obtain \( V(\Lambda) \approx .14 \), and \( V(\Lambda | r = 12, n = 20) \approx .058 \). The normal approximation to the prior and posterior densities for \( \Lambda \) is overlaid on Figure 2.4 as a dotted line.

Again, this example highlights that re-parameterizations are not necessarily innocuous in a Bayesian setting. We saw earlier that a prior that may be “flat” or “uninformative” with respect to \( \theta \) may not be uninformative with respect to \( g(\theta) \); Example 2.6 highlights that application of Bayes Rule can properly lead to a decrease in precision with respect to \( \theta \), but an increase in precision with respect to \( \Lambda = \text{logit}(\theta) \). There is nothing “wrong” with these results, or the underlying Bayesian analyses. But analysts should be aware of the possibility that sometimes, rationally updating one’s beliefs via Bayes Rule can lead to a decrease in precision, with the effect of new information that conflicts with prior beliefs being the certainty of one’s beliefs, depending on the particular way one is parameterizing the uncertain quantity.

### 2.2 ASSOCIATIONS BETWEEN BINARY VARIABLES

Analysis of the two-by-two crosstabulation has long been been a staple of data analysis in the social sciences. Just as many outcomes of interest in social-scientific analysis are binary variables, so too are plausible predictors of those outcomes. Examples include analyzing a binary outcome such as voter turnout by race (white/non-white), or examining whether democratic regimes are more likely to go to war with one another than with autocratic regimes.

Bayesian approaches to testing for association between binary variables builds on the ideas presented in the previous section. There we assumed that the data for analysis consisted
Figure 2.4  Prior and Posterior Densities for Example 2.6, probability and log-odds scale. The dotted lines on the log-odds scale indicate the Lindley (1965) normal approximation.
associations between binary variables

Table 2.1. Two-by-Two Cross Table

<table>
<thead>
<tr>
<th></th>
<th>$x_i$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>$n_0 - r_0$</td>
<td>$n_1 - r_1$</td>
</tr>
<tr>
<td>1</td>
<td>$r_0$</td>
<td>$r_1$</td>
</tr>
<tr>
<td></td>
<td>$n_0$</td>
<td>$n_1$</td>
</tr>
</tbody>
</table>

of a sequence of $n$ binary outcomes, $y_i \in \{0, 1\}$ that could be considered independent conditional on a parameter $\theta \in (0, 1) = E(y_i) = \Pr(y_i = 1)$. Given independence, the number of “successes”, $r = \sum_{i=1}^{n} y_i$, can be modeled as binomial random variable, conditional on the unknown success probability $\theta$ and the number of outcomes $n$. We now generalize this model to consider two sequences of binary outcomes, $y_{i0}$ for which a predictor $x_i = 0$, and $y_{i1}$ for which a predictor $x_i = 1$. Each sequence of binary data is presumed to be exchangeable, with success probabilities $\theta_0$ and $\theta_1$, respectively. We assume that the binary realizations are independent within and across the two groups defined by $x_i \in \{0, 1\}$, and so the likelihood for the two streams of binary data is just

$$L(\theta_0, \theta_1; y) = \prod_{i: x_i = 0} \theta_0^{y_i} (1 - \theta_0)^{1-y_i} \prod_{i: x_i = 1} \theta_1^{y_i} (1 - \theta_1)^{1-y_i}$$

Given independence, we can also model these two data as two binomial processes, and ignoring constants that do not depend on $\theta_0$ or $\theta_1$ we have

$$L(\theta_0, \theta_1; y) \propto \theta_0^{r_0} (1 - \theta_0)^{n_0 - r_0} \theta_1^{r_1} (1 - \theta_1)^{n_1 - r_1}$$

(2.8)

where $r_0$ and $r_1$ are the numbers of “successes” in the two groups defined by $x_i = 0$ and $x_i = 1$ respectively; see Table 2.1. Since the likelihood for $\theta_0$ and $\theta_1$ in equation 2.8 is a function solely of the four quantities $r_0, r_1, n_0$ and $n_1$, the likelihood is in effect a model for the two-by-two table in Table 2.1.

As in any Bayesian analysis, we formalize our prior beliefs over unknown parameters with probability densities. In this case, the parameter space is the unit square, since both $0 \leq \theta_0, \theta_1 \leq 1$; more formally, $\Theta = (\theta_0, \theta_1)^{\prime} \in \Theta = [0, 1] \otimes [0, 1]$. Specifying a prior for $\theta$ is simplified when prior information about the parameters $\theta_0$ and $\theta_1$ is independent, since in that case $p(\theta) = p(\theta_0) p(\theta_1)$, i.e., for independent random variables, the joint probability density is equal to the product of the marginal densities. But this need not be the case. For
example, if a priori, a belief that \( \theta_0 \) is more likely to be large than small is also informative about \( \theta_1 \), then \( \theta_0 \) and \( \theta_1 \) are not independent. In this case, we can not simply factor the joint prior density over \( \theta \) as the product of the two marginal prior densities for \( \theta_0 \) and \( \theta_1 \). For now, I focus on the simple case, where prior information about \( \theta \) is represented with Beta densities, which are conjugate with respect to the binomial likelihood in equation 2.8: i.e., if \( \theta_0 \sim \text{Beta}(\alpha_0, \beta_0) \) and \( \theta_1 \sim \text{Beta}(\alpha_1, \beta_1) \) then under the assumption that \( \theta_0 \) and \( \theta_1 \) are independent a priori

\[
p(\theta) = p(\theta_0)p(\theta_1)
\]

\[
\propto \theta_0^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}.
\]

Via Bayes Rule, the posterior is proportional to the prior times a likelihood, and so

\[
p(\theta_0, \theta_1| r_0, r_1, n_0, n_1) \propto p(\theta_0)p(\theta_1)L(\theta_0, \theta_1; y)
\]

\[
\propto \theta_0^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}
\]

\[
\times \theta_0^{\alpha_0+\alpha_1-n-1}(1-\theta_0)^{\beta_0+n_0-\alpha_1-r_1}
\]

\[
\times \theta_1^{\beta_0+n_1-\alpha_1-r_1}(1-\theta_1)^{\beta_1+\beta_1-1}.
\]

The terms involving \( \theta_0 \) we recognize as the kernel of a Beta density with parameters \( \alpha_0^* = \alpha_0 + r_0 \) and \( \beta_0^* = \beta_0 + n_0 - r_0 \), and similarly for \( \theta_1 \): i.e., \( p(\theta| r_0, r_1, n_0, n_1) = p(\theta_0| r_0, r_1) p(\theta_1| r_1, n_1) \). Thus, for this problem not only are \( \theta_0 \) and \( \theta_1 \) independent a priori, they are also independent a posteriori, and in particular,

\[
\theta_0| (r_0, r_1, n_0, n_1) \sim \text{Beta}(\alpha_0^*, \beta_0^*)
\]

\[
\theta_1| (r_0, r_1, n_0, n_1) \sim \text{Beta}(\alpha_1^*, \beta_1^*)
\]

The question of social-scientific interest with data such as these is whether the outcome \( y_i = 1 \) is more or less likely depending on whether \( x_i = 0 \) versus \( x_i = 1 \). Let the quantity of interest be \( q = \theta_1 - \theta_0 \). Then, in a Bayesian approach, we need to compute the posterior density for \( q \), a function of the unknown parameters. In particular, interest centers on how much of the posterior density for \( q \) lies above zero, since this the posterior probability that \( \theta_1 > \theta_0 \). Since the posterior densities for \( \theta_0 \) and \( \theta_1 \) are independent Beta densities, the posterior density of \( q \) is the density of the difference of two independent Beta densities. The analytical form of such a density has been derived by Pham-Gia and Turkkan (1993), but is extremely cumbersome and I do not reproduce it here.
Other functions of $\theta_0$ and $\theta_1$ are also of interest in the analysis of two-by-two tables. For instance, if we define $w = \theta_1/\theta_0$, then the posterior probability that $\theta_1 > \theta_0$ is the posterior probability that $w > 1$. The distribution of $w$, a ratio of independent Beta densities, has been derived by Weisberg (1972), and again, this has a particularly cumbersome form that I will not reproduce here.

A popular measure of association in two-by-two tables is the odds ratio: if

$$OR_0 = \frac{\theta_0}{1 - \theta_0} \quad \text{and} \quad OR_1 = \frac{\theta_1}{1 - \theta_1},$$

then the ratio

$$OR = \frac{OR_1}{OR_0} = \frac{\theta_1 (1 - \theta_0)}{(1 - \theta_1) \theta_0}$$

is a measure of association between $y_i$ and $x_i$. In particular, if $\theta_1 > \theta_0$, then $OR > 1$. Under the assumptions that $y_i|x_i = 0$ and $y_i|x_i = 1$ are independent Bernoulli processes of length $n_0$ and $n_1$, and assuming conjugate priors for $\theta_0$ and $\theta_1$, Altham (1969) showed that the posterior probability that $OR < 1$ to be computable as a finite sum of hypergeometric probabilities. In fact, Altham (1969) demonstrated a close connection between the Bayesian posterior probability and the $p$-value produced by Fisher’s (1935) exact test, the standard frequentist tool for testing associations in cross-tabulations with small samples. As we shall see below, relative to the posterior probabilities generated from a Bayesian analysis with uniform priors over $\theta_0$ and $\theta_1$, Fisher’s exact test is too conservative, in the sense that it generates $p$-values that are larger (in favor of the null hypothesis of no association) than the corresponding Bayesian posterior probabilities.

In addition, many analysts use the log of the odds-ratio as a measure of association in two-by-two tables. If

$$\Lambda_0 = \log \left( \frac{\theta_0}{1 - \theta_0} \right) \quad \text{and} \quad \Lambda_1 = \log \left( \frac{\theta_1}{1 - \theta_1} \right),$$

then the log of the odds ratio

$$\Lambda = \Lambda_1 - \Lambda_0 = \log \left( \frac{\theta_1}{1 - \theta_1} \right) - \log \left( \frac{\theta_0}{1 - \theta_0} \right) = \log \left( \frac{\theta_1 (1 - \theta_0)}{(1 - \theta_1) \theta_0} \right).$$

is positive when $\theta_1 > \theta_0$ (since the log-odds transformation is a strictly increasing function). The posterior density for $\Lambda$ is well approximated by a normal density, with the approximation being quite good when all the entries in the two-by-two crosstabulation are at least
Table 2.2. Young Adults Sexual Identity, by Sexual Orientation of Mother

<table>
<thead>
<tr>
<th>Child’s Identity ($y_i$)</th>
<th>Heterosexual ($x_i = 0$)</th>
<th>Lesbian ($x_i = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterosexual ($y_i = 0$)</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>Bisexual/Lesbian/Gay ($y_i = 1$)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 2.2. Young Adults Sexual Identity, by Sexual Orientation of Mother

5. This result follows from the fact that the distributions of $\Lambda_0$ and $\Lambda_1$ are approximately normal, as noted in section 2.1.3, and the fact that the difference (or sum) of two normal distributions is a normal distribution. In particular,

$$\Lambda| (r_0, r_1, n_0, n_1) \sim N \left( \log \left( \frac{(\alpha_0^* - \frac{1}{2}) (\beta_0^* - \frac{1}{2})}{(\beta_1^* - \frac{1}{2}) (\alpha_0^* - \frac{1}{2})} \right), \frac{1}{\alpha_1^*} + \frac{1}{\beta_1^*} + \frac{1}{\alpha_0^*} + \frac{1}{\beta_0^*} \right)$$

Closed-form expressions for the densities of these quantities are either extremely cumbersome or even unknown. In addition to the articles cited earlier, see Nurminen and Mutanen (1987) or Pham-Gia and Turkkan (2002) for a taste of the difficulties involved in deriving posterior densities for various summaries of the relationships in two-by-two tables. See Problem XXXXX for some simple cases. The simulation methods to be discussed in section 4 provide a simple way to characterize posterior densities with non-standard or cumbersome analytical forms. A mixture of simulation and analytical methods were used to compute and graph the prior and posterior densities reported in the following example; details await section 4.

**EXAMPLE 2.8**

Do Parents Influence the Sexual Orientation of Their Children? Golombok and Tasker (1996) studied whether there is any link between the sexual orientation of parents and children. Twenty five children of lesbian mothers ($x_i = 1$) and a control group of 21 children of heterosexual single mothers ($x_i = 0$) were first seen at age 9.5 years (on average), and again at 23.5 years (on average). In the second interview, children were asked about their sexual identity, responding as either as bisexual/lesbian/gay ($y_i = 1$) or heterosexual ($y_i = 0$); one of the participants did not supply information on this variable.

Table 2.2. reproduces the cross-tabulation of mother’s sexual orientation by child’s sexual identity, reported in Table 2 of Golombok and Tasker (1996). All of the
children of the heterosexual mothers report a heterosexual identity \((r_0 = 0; n_0 = 20)\), while almost all of the children of the lesbian mothers report a heterosexual identity \((r_1 = 2; n_1 = 25)\). Thus, the maximum likelihood estimates of \(\theta_0\) and \(\theta_1\) are \(\hat{\theta}_0 = r_0/n_0 = 0\) and \(\hat{\theta}_1 = r_1/n_1 = 0.08\). Golombok and Tasker (1996) report a frequentist test of the null hypothesis that \(\theta_0 = \theta_1\), presumably against the two-sided alternative hypothesis \(\theta_0 \neq \theta_1\) with Fisher’s (1935) exact test, and report the \(p\)-value of the test as “ns” for “not significant”. Given the data in Table 2.2, the actual \(p\)-values for Fisher’s exact test can be easily computed. Let \(H_0 : \theta_1 = \theta_0 \iff \text{OR} = 1\), where OR is the odds-ratio defined in equation 2.9 (note that the sample estimate of the odds ratio is infinite). Then, the \(p\)-values yielded by Fisher’s exact test against the following alternatives are

<table>
<thead>
<tr>
<th>(H_A)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{OR} \neq 1 \iff \theta_1 \neq \theta_0)</td>
<td>.49</td>
</tr>
<tr>
<td>(\text{OR} &gt; 1 \iff \theta_1 &gt; \theta_0)</td>
<td>.30</td>
</tr>
<tr>
<td>(\text{OR} &lt; 1 \iff \theta_1 &lt; \theta_0)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

A Bayesian analysis proceeds as follows. The data are presumed to be two realizations from two independent binomial processes, of lengths \(n_0\) and \(n_1\) respectively, with success probabilities \(\theta_0\) and \(\theta_1\), respectively. I posit uniform priors for both \(\theta_0\) and \(\theta_1\), i.e., \(\theta_0 \sim \text{Beta}(1, 1)\) and \(\theta_1 \sim \text{Beta}(1, 1)\). Then the posterior densities are \(\theta_0 \sim \text{Beta}(1, 21)\) and \(\theta_1 \sim \text{Beta}(3, 24)\).

Figure 2.5 shows prior densities (dotted lines) and posterior densities for various parameters: \(\theta_0, \theta_1\), the difference \(\theta_1 - \theta_0\), the ratio \(\theta_1/\theta_0\), the odds ratio (equation 2.9) and the log of the odds ratio (equation 2.10). Recall that the question of scientific interest is whether \(\theta_1 > \theta_0\); or more generally, is the probability of reporting a L/B/G sexual identity associated with whether one’s mother is lesbian or heterosexual? The four posterior densities in the lower four panels of Figure 2.5 indicate that the sample information does suggest \(\theta_1 > \theta_0\), but the evidence is not strong. For the difference of \(\theta_1\) and \(\theta_0\), their ratio, their odds ratio, and the log of their odds ratio, the posterior probability that \(\theta_1 > \theta_0\) is the area under the respective posterior density to the right of the vertical line. In each instance, this probability is .84, not overwhelming evidence, and falling short of conventional standards of statistical significance. However, this posterior probability is substantially higher than the probability implied by Fisher’s exact test, which yields a \(p\)-value of .30, for the test of \(H_0 : \text{OR} = 1\) against the one-
Figure 2.5  Posterior Densities for Example 2.8. The top two panels show the posterior densities of $\theta_0$ and $\theta_1$, which are Beta densities; the remaining panels show the posterior densities for various functions of $\theta_0$ and $\theta_1$: the difference $\theta_1 - \theta_0$, the ratio $\theta_1/\theta_0$, the odds ratio and the log odds ratio. The vertical lines indicate the point where $\theta_1 = \theta_0$; the area under the various curves to the right of this line is the posterior probability that $\theta_1 > \theta_0$. The dotted line in each panel is the prior density for the particular parameter, implied by the uniform priors on $\theta_0$ and $\theta_1$. 
sided alternative $\text{OR} > 1$, suggesting that the alternative hypothesis is considerably much less likely \textit{a posteriori} than we obtained with the Bayesian procedure (with uniform priors on $\theta_0$ and $\theta_1$).

Finally, note the normal approximation to the posterior density of the log-odds ratio, the gray line in the lower right panel of Figure 2.5, corresponding to the density in equation 2.11. With the small cell counts observed for this example, it is well known that the normal approximation is not accurate, and is shifted to the right relative to the actual posterior for the log of the odds ratio, $\Lambda$. Under the normal approximation, the estimated posterior probability that $\Lambda > 0 \iff \theta_1 > \theta_0$ is .89, reasonably close to the actual result of .84, but a substantial overestimate nonetheless.

Fisher’s exact test is well known to produce $p$-values that are too conservative relative to the analogous quantities obtained from a Bayesian analysis of two binomial proportions that places uniform priors over each unknown parameters. Specifically, Altham (1969) showed that the $p$-values against the null hypothesis produced by Fisher’s test can be obtained from a Bayesian procedure — again, approaching the problem as the difference of two binomial proportions — that employs improper Beta(0,1) and Beta(1,0) improper priors over the two unknown binomial success parameters. This prior is somewhat odd, in that it corresponds to a strong prior belief in negative association between the two variables.

\textbf{EXAMPLE 2.9}

\textit{War and Revolution in Latin America} Sekhon (2005) presents a reanalysis of a cross-tabulation originally appearing in Geddes (1990), examining the relationship between foreign threat and social revolution in Latin America, in turn, testing a claim advanced by Skocpol (1979). Sekhon updates the data presented by Geddes (1990), generating the following 2-by-2 table:

<table>
<thead>
<tr>
<th>Revolution</th>
<th>No Revolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defeated and Invaded or Lost Territory</td>
<td>1</td>
</tr>
<tr>
<td>Not Defated with 20 years</td>
<td>2</td>
</tr>
</tbody>
</table>

Each observation is a 20 year period for each Latin American country, with any given Latin American country contributing multiple observations (this makes exchangeability seem questionable, but like the original authors we ignore this potential complication). The sole observation in the top left of the cross-tabulation is Bolivia:
it suffered a military defeat in 1935, and a social revolution in 1952. The two observations in the lower left of the table are Mexico (revolution in 1910) and Nicaragua (revolution in 1979). We are interested in the conditional, binomial probabilities

- $\theta_0 = \Pr(y_i = 1|x_i = 0)$, the probability of revolution conditional on no military defeat within 20 years;

- $\theta_1 = \Pr(y_i = 1|x_i = 1)$, the probability of revolution conditional on having suffered a military loss.

The MLEs are $\hat{\theta}_0 = 2/(2 + 74) = .026$ and $\hat{\theta}_1 = 1/(7 + 1) = .125$, suggesting that revolutions are much more likely conditional on military defeat than conditional on not having experienced a military defeat. However, Fisher’s test of the null hypothesis of no association against an unspecified alternative yields a $p$-value of .262, which is the same $p$-value obtained with the one-sided alternative $H_A: \theta_1 > \theta_0$.

With the uniform priors $\theta_j \sim \text{Beta}(1, 1), j = 0, 1$, the posterior densities are $\theta_0 \sim \text{Beta}(3, 75)$ and $\theta_1 \sim \text{Beta}(2, 8)$. Sekhon (2005) reports the results of a simulation-based analysis of the posterior density of the quantity $q = \theta_1 - \theta_0$, finding the posterior probability that $\theta_1 > \theta_0$ to be .9473, which is quite strong evidence in favor of the hypothesis. Thus, the “conservatism” of Fisher’s exact test relative to this Bayesian approach is readily apparent.

But, consider the improper priors $\theta_0 \sim \text{Beta}(1, 0)$ and $\theta_1 \sim \text{Beta}(0, 1)$. These improper priors have the properties $\lim_{\theta_0 \to 0} p(\theta_0) = \lim_{\theta_1 \to 1} p(\theta_1) = +\infty$ and have the effect of being favoring the null hypothesis that $\theta_1 = \theta_0$. With these priors the posterior densities are $\theta_0 \sim \text{Beta}(3, 74)$ and $\theta_1 \sim \text{Beta}(1, 8)$, and the posterior probability that $\theta_1 > \theta_0$ is .738, corresponding to result given by Fisher’s exact test. That is, Fisher’s exact test has a $p$-value of $1 - .738 = .262$, which frequentists interpret as the relative frequency (and hence, “probability”) of observing a result at least as extreme as the result actually observed, in repeated sampling of 2-by-2 tables with the same row and column marginals as the table actually observed, if the null hypothesis were true. Two conclusions follow: (1) the priors that “rationalize” Fisher’s exact test for a Bayesian are improper and (2) lead to the conclusions that are “too conservative” (in favor of the null hypothesis of no association), at least in two-by-two tables with small cell counts.
2.3 LEARNING ABOUT A NORMAL MEAN

Much quantitative work in the social sciences involves making inferences about the mean of a population given sample data. When the variable being analyzed is continuous (or approximately so), the normal distribution is often used as an underlying probability model for the sample observations: e.g., $y_i \sim N(\mu, \sigma^2)$, with interest centering on the unknown mean parameter $\mu$. Examples include analysis of surveys measuring incomes and working hours. Sometimes the normal distribution is used when the data being analyzed are discrete, but take on a relatively large number of categories and so are approximately continuous (e.g., 7 point scales in surveys measuring political attitudes). Regression analysis, one of the most widely used tools in the social sciences, is a special case of this approach to modeling, where the mean of continuous response $y_i$ is modeled as a function of predictors $x_i$; for instance, consider the normal, linear regression model $y_i|x_i \sim N(x_i\beta, \sigma^2)$, with $\beta$ a vector of unknown parameters. We return to the regression model in section 2.5. For data for which normal distributions are implausible, other distributions are required: examples include binary data, considered in the previous sections, or counts, for which the Poisson or negative binomial distributions may be more appropriate. This said, the normal distribution is enormously popular in statistical modeling, and many variables that do not appear to have a normal distribution can be transformed to better approximate normality; e.g., taking the logarithm of a variable that is highly skewed, such as income. In addition, as sample size increases, discrete distributions such as the binomial or the Poisson are well approximated by the normal distribution. Thus, the normal distribution is often used as a substitute for these discrete distributions in large samples; a common example is the analysis of a binary variable in a large sample (e.g., estimating vote intentions from surveys with large numbers of respondents).

2.3.1 Variance Known

Consider the case where we have $n$ observations on a variable $y$, yielding a vector of observations $y = (y_1, \ldots, y_n)'$. Conditional on the unknown mean parameter $\mu$, we consider the $y_i$ exchangeable, and we also assume that the variance $\sigma^2$ is known. Since the normal distribution is characterized by two parameters — the mean and the variance — no further parameters or assumptions are required to deploy a normal model. If each (exchangeable)
observation is modeled as \( y_i \sim N(\mu, \sigma^2) \), with \( \sigma^2 \) known, then the likelihood is

\[
L(\mu; y, \sigma^2) \equiv f(y|\mu) = \prod_{i=1}^{n} \exp \left[ -\frac{(y_i - \mu)^2}{2\sigma^2} \right].
\] (2.12)

The maximum likelihood estimate of \( \mu \) is simply the sample mean, i.e., \( \hat{\mu} = \bar{y} = n^{-1} \sum y_i \).

Bayesian inference for \( \mu \) in this case proceeds as usual: we multiply the likelihood for \( \mu \) by a prior density for \( \mu \), \( p(\mu) \), to obtain the posterior density for \( \mu \), \( p(\mu|y) \). For this likelihood, generated from the normal densities for each \( y_i \), a conjugate prior density is the normal density. That is, if prior beliefs about \( \mu \) are be represented with a normal density, then given a normal model for the data (modeled as exchangeable conditional on \( \mu \) in the likelihood in equation 2.12), the posterior density for \( \mu \) is also a normal density. Specifically,

**Proposition 2.2.** Let \( y_i \overset{iid}{\sim} N(\mu, \sigma^2) \), \( i = 1, \ldots, n \), with \( \sigma^2 \) known, and \( y = (y_1, \ldots, y_n)' \). If \( \mu \sim N(\mu_0, \sigma_0^2) \) is the prior density for \( \mu \), then \( \mu \) has posterior density

\[
\mu|y \sim N \left( \frac{\mu_0 \sigma_0^{-2} + \bar{y} n}{\sigma_0^{-2} + n/\sigma^2}, \left( \sigma_0^{-2} + n/\sigma^2 \right)^{-1} \right).
\]

**Proof.** See Proposition B.1 in the Appendix. \( \Box \)

This result draws on familiar result in conjugate, Bayesian analysis: the mean of the posterior density is a precision-weighted average of the mean of the prior density, and the maximum likelihood estimate. In this case, the mean of the prior density is \( \mu_0 \) and has precision \( \sigma_0^{-2} \), equal to the inverse of the variance of the prior density. The maximum likelihood estimate of \( \mu \) is \( \bar{y} \), which has precision equal to \( n/\sigma^2 \), the inverse of the variance of \( \bar{y} \). Inspection of the result in Proposition 2.2 makes it clear that as prior information about \( \mu \) becomes less precise (\( \sigma_0^{-2} \to 0, \sigma_0^2 \to \infty \)), the posterior density is dominated by the information about \( \mu \) in the likelihood. In this case popular Bayes estimates such as the posterior mean (or mode) tend to the maximum likelihood estimate, \( \bar{y} \), and the variance of the posterior density tends to \( \sigma^2/n \), the variance of the sampling distribution of \( \hat{\mu} \) that would arise via a frequentist approach to inference for this problem.

Note that just as we saw in the case of inference for a binomial success probability, a Bayes estimate can be expressed as the sample estimate, plus an offset that reflects the influence of the prior. In this case, the mean of the posterior density can be expressed as

\[
E(\mu|y) = \bar{y} + \lambda(\mu_0 - \bar{y})
\]
where
\[ \lambda = \frac{\sigma_0^{-2}}{\sigma_0^{-2} + \frac{n}{\sigma^2}} \]

is the precision of the prior relative to the total precision (the prior precision plus the precision of the data). As \( \lambda \to 1 \), \( E(\mu|\mathbf{y}) \to \mu_0 \), while as \( \lambda \to 0 \), say, when prior information about \( \mu \) is sparse relative to the information in the data, \( E(\mu|\mathbf{y}) \to \bar{y} \). And in general, the Bayes estimate \( E(\mu|\mathbf{y}) \) is a convex combination of the prior mean \( \mu_0 \) and the MLE \( \bar{y} \).

**EXAMPLE 2.10**

**Combining Information from Polls, via the Normal Approximation to the Beta Density.** In Example 2.3, information from a poll and from regression analysis of previous election returns was combined via Bayes Rule to generate a posterior density over George Bush’s likely share of the two-party vote in Florida in the 2000 presidential election. We now revisit this example, but representing both sources of information with normal densities, since the Beta can be well approximated by a normal as the parameters of the Beta get large (as is the case, say, when the Beta density is being used to represent uncertainty about a population parameter generated by a survey based on a large sample). That is, if \( \theta \sim \text{Beta}(\alpha, \beta) \), then as \( \alpha, \beta \to \infty \), \( \theta \sim N(\tilde{\theta}, \sigma^2) \), where \( \tilde{\theta} = E(\theta; \alpha, \beta) = \alpha / (\alpha + \beta) \) and \( \sigma^2 = V(\theta; \alpha, \beta) = \alpha \beta / [(\alpha + \beta)^2 (\alpha + \beta + 1)] \).

So, in Example 2.3, 279 respondents said they would for Bush in the November 2000 election, of 509 respondents expressing a preference for either Bush or Gore; i.e., \( \alpha = 279 \) and \( \alpha + \beta = 509 \), so \( \tilde{\theta} = 279/509 = .548 \) and \( \sigma^2 = [279 \times (509 - 279)]/[509^2 \times 510] = .000486 \). The regression analysis of historical election returns produces a predictive distribution for \( \theta \), specifically, \( \theta \sim N(.491, .000484) \). For the purposes of modeling, we treat the predictive distribution from the historical analysis as a prior over \( \theta \), and the information in the poll as the likelihood; that is, the poll result \( \tilde{\theta} \) is observed data, generated by a sampling process governed by \( \theta \) and the poll’s sample size (n.b., \( n = \alpha + \beta \)). Specifically, using the normal approximation, \( \tilde{\theta} \sim N(\theta, \sigma^2) \) or \( .548 \sim N(\theta, .000486) \).

We are now in a position to use the result in Proposition 2.2, as follows:

\[
\begin{align*}
\theta & \sim N(.491, .000484) \quad \text{(prior)} \\
.548 & \sim N(\theta, .000486) \quad \text{(data/likelihood)} \\
\theta|\text{data} & \sim N(\theta^*, \sigma^{*2}) \quad \text{(posterior)}
\end{align*}
\]
where
\[
\begin{align*}
\theta^* &= \frac{0.491}{0.000484} + \frac{0.548}{0.000486} = 0.519 \\
\sigma_{2*}^2 &= \left( \frac{1}{0.000484} + \frac{1}{0.000486} \right)^{-1} = 0.000243
\end{align*}
\]

Notice that the resulting posterior density has smaller variance than either the prior variance or the variance of the distribution representing the data; the situation is well-represented by Figure 2.1, which accompanies Example 2.3. This elegant feature is a property of conjugate analysis of normal-based likelihoods when the variance term is known, and is consistent with Bayesian analysis being equivalent to combining information from multiple sources. However, contrast the way that application of Bayes Rule led to a loss of precision in Example 2.6.

**Remark.** The quality of the normal approximation to the Beta density is excellent when \(\alpha\) and \(\beta\) are even of modest size. Figure 2.6 shows the error arising from computing various quantiles of a Beta(20,\(\beta\)) density with the normal approximation used above, defined as the difference between the normal quantile and the Beta quantile. For low values of \(\beta\), the Beta density is quite asymmetric, and the (symmetric) normal approximation is relatively poor, especially as the quantile of interest is further into the tail of the density. But as \(\beta \to \infty\), even with \(\alpha = 20\), the normal approximation is quite good, and produces quite small errors for frequently used quantiles (e.g., .025, .05). Note that \(\beta = 20\) corresponds to the case of a symmetric Beta density, for which the normal approximation is quite good, with the exception of the relatively extreme .001 quantile.

### 2.3.2 Mean and Variance Unknown

The more typical case encountered in practice is that both the mean and the variance are unknown parameters. That is, the typical situation is that we have data \(y = (y_1, \ldots, y_n)'\) with \(y_i \sim N(\mu, \sigma^2)\), and interest focuses on the vector of parameters \(\theta = (\mu, \sigma^2)'\). The prior density is now a bivariate, joint density, \(p(\theta) = p(\mu, \sigma^2)\), as is the posterior density \(p(\mu, \sigma^2|y)\). The likelihood is defined with respect to the two unknown parameters, i.e.,

\[
L(\mu, \sigma^2; y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_i - \mu)^2}{2\sigma^2} \right].
\]  \tag{2.13}
Figure 2.6  Error in Using a Normal Approximation to a Beta Density. The vertical axis shows the difference between a given quantile of a Beta(20,β) density, and the same quantile of normal density with mean α/(α + β) and variance αβ/[(α + β)²(α + β + 1)]. Different values of the β parameter are given on the horizontal axis (log-scale). Each line corresponds to a given quantile of each distribution: .001, .01, .025, .05, .10 and .25. Note that (a) at β = 20, the Beta distribution is symmetric (α = β = 20); (b) as β → ∞, the normal gives an increasingly better approximation to the Beta density.
The conjugate prior for $\theta = (\mu, \sigma^2)'$ with respect to the normal likelihood in equation 2.13 requires factoring the joint prior density $p(\mu, \sigma^2)$ as the product of the conditional density $p(\mu|\sigma^2)$ and a marginal prior density over $\sigma^2$, as the following proposition makes clear:

**Proposition 2.3** (Conjugate Priors for Mean and Variance, Normal Data). Let $y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, and let $y = (y_1, \ldots, y_n)'$. If $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0)$ is the (conditional) prior density for $\mu$, and $\sigma^2 \sim \text{Inverse-Gamma}(\nu_0, \sigma_0^2)$ is the prior density for $\sigma^2$ then

$$
\mu|\sigma^2, y \sim N \left( \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n}, \frac{\sigma^2}{n_0 + n} \right),
$$

$$
\sigma^2|y \sim \text{Inverse-Gamma} \left( \frac{\nu_0 + n}{2}, \frac{1}{2} \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{y})^2 \right),
$$

where $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$.

**Proof.** See Proposition B.2 in the Appendix.

It is a feature of a conjugate Bayesian analysis that the posterior densities are of the same type as the prior densities. Thus, since we used a normal prior density for $\mu$, conditional on $\sigma^2$, the posterior density for $\mu$ is also of the same form. That is, the conditional posterior density for $\mu$ is a normal density in which $\sigma^2$ appears in the expression for the variance of the conditional posterior density. But usually, interest focuses on $\mu$ itself, and we need to confront the fact that $\sigma^2$ is itself subject to posterior uncertainty, and that uncertainty should rightly propagate into inferences about $\mu$ itself. That is, we seek the marginal posterior density for $\mu$, obtained by integrating out (or “averaging over”) the posterior uncertainty with respect to $\sigma^2$; i.e.,

$$
p(\mu|y) = \int_0^\infty p(\mu|\sigma^2, y) p(\sigma^2|y) d\sigma^2
$$

where the limits of integration follow from the fact that variances are strictly positive. Evaluating this integral is not as hard as it might seem, and indeed, doing so generates a familiar result.

**Proposition 2.4.** Under the conditions of Proposition 2.3, the marginal posterior density of $\mu$ is a student-$t$ density, with location parameter

$$
E(\mu|\sigma^2, y) = \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n},
$$

scale parameter $\sigma_t^2/(n_0 + n)$, and $\nu_0 + n$ degrees of freedom, where $\sigma_t^2 = S_1/(\nu_0 + n)$ and

$$
S_1 = \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{n_0 n}{n_1} (\bar{y} - \mu_0)^2.
$$

**Proof.** See Proposition B.3.
EXAMPLE 2.11

Voter Fraud in Pennsylvania

PICTURES.

2.3.3 An Improper, Reference Prior

A Bayesian analysis can be made to produce a posterior density that corresponds with the results of a frequentist analysis. In the case of normal data with unknown mean and variance, considering adopting the improper prior \( p(\mu, \sigma^2) \propto 1/\sigma^2 \). This prior density is uniform with respect to \( \mu \), generating a so-called marginal reference prior \( p(\mu) \propto 1 \) and \( p(\sigma^2) \propto 1/\sigma^2 \). This prior is completely uninformative with respect to \( \mu \), but attaches greater prior density to smaller values of \( \sigma^2 \) than larger values; however, the prior is uniform with respect to \( \log \sigma^2 \). Note also that unlike the conjugate prior considered above, \( \mu \) and \( \sigma^2 \) are independent a priori — the improper, joint prior density is simply the product of the two marginal prior densities — consistent with the notion that if we are quite ignorant about \( \mu \) and \( \sigma^2 \), then gaining some information over one of these parameters should not affect beliefs about the other parameter (e.g., Lee 2004, 62-63).

With this prior, the posterior density is

\[
p(\mu, \sigma^2|y) \propto (\sigma^2)^{-n/2-1} \exp\left[\frac{-1}{2\sigma^2} \left( (n - 1)s^2 + n(\bar{y} - \mu)^2 \right) \right],
\]

where \( s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n - 1) \) is the conventional, unbiased sample estimate of the variance \( \sigma^2 \). After some simple algebra (e.g., see the proofs of Propositions B.2 and B.3 for the algebra with proper, conjugate priors), we obtain the following familiar looking results:

\[
\begin{align*}
\mu|\sigma^2, y & \sim N(\bar{y}, \sigma^2/n) \\
\sigma^2|y & \sim \text{Inverse-Gamma} \left( \frac{n-1}{2}, \frac{(n-1)s^2}{2} \right) \\
\frac{\mu - \bar{y}}{\sqrt{s^2/n}} & \sim \text{student-} t_{n-1}
\end{align*}
\]

The first of these results simply states that with the improper reference prior, the conditional posterior density for \( \mu \) is centered over the conventional estimate of \( \mu \), \( \bar{y} \). Moreover, the conditional posterior density for \( \mu \) has variance \( \sigma^2/n \), equal to the variance of frequentist sampling distribution. The second result provides the marginal posterior density of \( \sigma^2 \), an inverse-Gamma density with mean XXXX and mode XXXX. The third result provides the marginal posterior density of \( \mu \) in standardized form, a (standardized) student-\( t \) distribution.
with \( n - 1 \) degrees of freedom, identical to the density arising in frequentist analysis when testing hypotheses about \( \mu \).

2.4 LEARNING ABOUT A CORRELATION

2.5 LEARNING ABOUT PARAMETERS IN A REGRESSION MODEL

2.6 SUMMARY

In this chapter...

XXXX
Conjugacy summary.

Problems

2.1 If

\[
\text{OR} = \frac{\theta_1(1 - \theta_0)}{(1 - \theta_1)\theta_0}
\]

be the odds-ratio of \( \theta_1 \) to \( \theta_0 \), show that \( \text{OR} > 1 \iff \theta_1 > \theta_0 \).

2.2 Suppose \( \theta \sim \text{Unif}(0, 1) \).

1. Derive the density of the quantity

\[
\text{OR} = \frac{\theta}{1 - \theta}
\]

2. Derive the density of the quantity

\[
\log \left( \frac{\theta}{1 - \theta} \right)
\]

2.3 Suppose \( \theta_0 \sim \text{Unif}(0, 1) \) and \( \theta_1 \sim \text{Unif}(0, 1) \).

1. Derive the density of the quantity \( w = \theta_1/\theta_0 \).

2. Derive the density of the quantity \( \delta = \theta_1 - \theta_0 \).

2.4 Prove that \( p(\theta) \equiv \text{Beta}(0, 1) \) is improper.

2.5 Suppose \( y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \) with \( \mu \) and \( \sigma^2 \) unknown, and \( i = 1, \ldots, n \). If the prior density for \( (\mu, \sigma^2) \) is \( p(\mu, \sigma^2) \propto 1/\sigma^2 \), prove that
1. $\mu | \sigma^2, y \sim N(\bar{y}, \sigma^2/n)$

2. $\sigma^2 | y \sim \text{Inverse-Gamma} \left( \frac{n-1}{2}, \frac{(n-1)s^2}{2} \right)$
CHAPTER 3

WHY BE BAYESIAN?

diatribes to come here, perhaps as a later chapter, perhaps not at all.

XXXX
PART II

SIMULATION BASED BAYESIAN STATISTICS
CHAPTER 5

BAYESIAN ANALYSIS OF EXTENDED REGRESSION MODELS
CHAPTER 6

COMBINING INFORMATION: BAYESIAN HIERARCHICAL MODELING
PART III

ADVANCED APPLICATIONS IN THE SOCIAL SCIENCES
CHAPTER 7

BAYESIAN ANALYSIS OF CHOICE MAKING
CHAPTER 8

BAYESIAN APPROACHES TO MEASUREMENT
Definition A.1 (σ-algebra). Let \( A \) be a family of subsets of \( \Omega \). Then \( A \) is a σ-algebra over \( \Omega \) if and only if

1. \( \emptyset \in A \)
2. \( H \in A \Rightarrow H^c \in A \)
3. if \( H_1, H_2, \ldots \) is a sequence in \( A \) then their countable union \( \bigcup_{i=1}^{\infty} H_i \in A \)

Definition A.2 (Probability Measure). Let \( A \) be a σ-algebra over \( \Omega \). Then the function \( \mu \) is a probability measure on \( A \) if

1. \( \mu(\emptyset) = 0 \)
2. \( \mu(\bigcup_{i} H_i) = \sum_{i} \mu(H_i) \) for any sequence of pairwise disjoint sets, \( H_1, H_2, \ldots \) in \( A \)
3. \( \mu(\Omega) = 1 \)

Definition A.3 (Probability Space). Let \( \Omega \) be a set, and let \( A \) be a σ-algebra over \( \Omega \). Let \( \mu \) be a probability measure on \( A \). Then the tuple \((\Omega, A, \mu)\) is a probability space.
Definition A.4 (Real-Valued Random Variable). Let \((\Omega, A, \mu)\) be a probability space. Then a function \(X : \Omega \to \mathbb{R}\) is a real-valued random variable if for all subsets \(A_r = \{\omega : X(\omega) \leq r\}, r \in \mathbb{R}, A_r \in A\).

Definition A.5 (Cumulative Distribution Function). If \(X\) is a real-valued random variable, then the cumulative distribution function (CDF) of \(X\) is defined as

\[
F(x) = P(X \leq x), \quad -\infty < x < \infty
\]

Remark. A CDF of a real-valued random variable has the following properties

1. \(x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)\)
2. \(\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1\)
3. \(F(x) = F(x^+), \) where \(F(x^+) = \lim_{y \to x, y > x} F(x)\) (continuous from the right).

Definition A.6 (Probability Density Function).

Definition A.7 (Improper Density Function). Suppose \(x \in X \subseteq \mathbb{R}^p\) has the density function \(g(x)\). If \(\int_X g(x)dx \neq 1\) then the density function is an improper density.

Remark. A density function that is not improper is said to be proper.

Definition A.8 (Kernel of a Probability Density Function).

Definition A.9 (Jacobian of a vector-valued function). Let \(x \in X \subseteq \mathbb{R}^p, v \in V \subseteq \mathbb{R}^p\) and let \(\phi\) be a bijection from \(V\) onto \(X\) with continuous partial derivatives, i.e., \(\frac{\partial x_i}{\partial v_j}\) exists and is continuous on \(V, \forall i, j = 1, \ldots, p\). The quantity

\[
J \left( \begin{array}{c} x \\ v \end{array} \right) = \det \begin{bmatrix} \frac{\partial x_1}{\partial v_1} & \frac{\partial x_2}{\partial v_1} & \cdots & \frac{\partial x_p}{\partial v_1} \\ \frac{\partial x_1}{\partial v_2} & \frac{\partial x_2}{\partial v_2} & \cdots & \frac{\partial x_p}{\partial v_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial v_p} & \frac{\partial x_2}{\partial v_p} & \cdots & \frac{\partial x_p}{\partial v_p} \end{bmatrix}
\]

is the Jacobian determinant of \(\phi\), often simply referred to as the Jacobian of \(\phi\).

Proposition A.1 (Jacobian Theorem). Define \(x \in X, v \in V, \phi,\) and \(J\) as in Definition A.9. Then

\[
\int_X g(x)dx = \int_V g(\phi(v)) \left| J \left( \begin{array}{c} x \\ v \end{array} \right) \right| dv
\]
Proof. The proof is trivial for the special case of scalar quantities \((p = 1)\). Then the proposition is \(\int g(x)dx = \int g(\phi(v))dx/dv dv = \int g(x)dx\) via the chain rule of differential calculus and the facts that the determinant of a scalar is the scalar and \(x = \phi(v)\). See Aliprantis and Burkinshaw (1981, Theorem 30.7) for the proof in the general case. \(\triangleright\)

**Proposition A.2** (Probability Density Functions Under Transformations). Define \(x \in \mathcal{X}, v \in \mathcal{V}, \phi,\) and \(J\) as in Definition A.9, and let \(f_x\) be the probability density function of \(x\). Then

\[
f_v = f_x(\phi(v)) \left| J \left( \begin{array}{c} x \\ v \end{array} \right) \right|
\]

is a probability density function for \(v \in \mathcal{V}\).

Proof. By Proposition A.1,

\[
\int_{\mathcal{X}} f_x dx = \int_{\mathcal{V}} f_x(\phi(v)) \left| J \left( \begin{array}{c} x \\ v \end{array} \right) \right| dv.
\]

But since \(f_x\) is a pdf, \(\int f_x dx = 1\). Likewise, \(\int_{\mathcal{V}} f_v dv = 1\) and so \(\int_{\mathcal{X}} f_x dx = \int f_v dv\) and the result follows. \(\triangleright\)

**Example 1.1**

**Reparameterization of a Binomial Success Probability (Fisher 1922).** We observe \(r\) successes in \(n\) Bernoulli trials, independent conditional on an unknown success probability \(\theta\). With a uniform prior for \(\theta\), the posterior for \(\theta\) is a Beta\((r + 1, n - r + 1)\) density, with the same shape as the likelihood function (i.e., a mode at the maximum likelihood estimate \(\hat{\theta} = r/n\)). Consider inference for the parameter \(q\), where \(\sin q = 2\theta - 1\). Assume a uniform prior for \(q\),

\[
f(q) = \begin{cases} 
\frac{1}{\pi} & \frac{-\pi}{2} \leq q \leq \frac{\pi}{2} \\
0 & \text{otherwise}
\end{cases}
\]

Note that \(\theta = (1 + \sin q)/2\). Thus, as a function of \(q\), and ignoring constants that do not depend on \(q\), the likelihood is

\[
\mathcal{L}(q; r, n) = (1 + \sin q)^r(1 - \sin q)^{n-r}
\]

With a prior that is uniform over \(q\), the posterior density for \(q\) is proportional to this likelihood. But what is the posterior density for \(\theta\)? Note that that \(q(\theta) = \)
\[ \sin^{-1}(2\theta - 1) \text{ and that } \frac{\partial q}{\partial \theta} = (\theta(1 - \theta))^{-1/2} \]

Then, using the result in Proposition A.2,

\[
f(\theta; r, n) = f_q(q(\theta); r, n) \left| \frac{\partial q}{\partial \theta} \right| = \left(1 + \sin[\sin^{-1}(2\theta - 1)]\right)^r \left(1 - \sin[\sin^{-1}(2\theta - 1)]\right)^{n-r} (\theta(1 - \theta))^{-1/2}
\]

\[ \propto \theta^{r-\frac{1}{2}} (1 - \theta)^{n-r-\frac{1}{2}} \]

which we recognize as the kernel of a Beta\((r + \frac{1}{2}, n - r + \frac{1}{2})\) density. That is, under a uniform prior for \(q\), the implied posterior density for \(\theta\) has a mode at \((r - \frac{1}{2})/(n - 1)\). This differs from the maximum likelihood estimate of \(r/n\) by an amount that vanishes as \(n \to \infty\). Nonetheless, this example demonstrates that a uniform prior with respect to \(\theta\) does not necessarily generate the results one would expect via a uniform prior over \(g(\theta)\). See section 2.1.3 for additional discussion.

**Definition A.10 (Gamma Function).** The gamma function is defined as
\[
\Gamma(a) = \int_0^\infty t^{a-1} \exp^{-t} \, dt \tag{A.1}
\]

Note that for \(a = 0, 1, 2, \ldots, a! = \Gamma(a + 1)\).

**Definition A.11 (Beta Function).** The beta function is defined as
\[
\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \tag{A.2}
\]

This function appears as the normalizing constant in the beta distribution.

**Definition A.12 (Beta Density).**

**Definition A.13 (Normal Density).**

**Definition A.14 (Gamma Density).**

**Definition A.15 (Inverse-Gamma Density).**

**Definition A.16 (student-t Density).**

**Definition A.17 (Multivariate Normal Density).**

**Definition A.18 (Likelihood Function).** Consider data \(y = (y_1, \ldots, y_n)'\) with joint pdf \(f_y(y; \theta)\). When \(f_y\) is re-written as a function of \(\theta\) given \(y\) it is called the likelihood function: i.e., \(L(\theta; y)\).
Proposition A.3 (Invariance Property of Maximum Likelihood).

Definition A.19 (Identification).
Here I provide a informal re-statement of the fact that under a wide set of conditions, posterior densities tend to normal densities as the amount of data available for analysis becomes arbitrarily plentiful (i.e., as $n \to \infty$). This “heuristic proof” appears in numerous places in the literature: e.g., O’Hagan (2004, 73), Williams (2001, 204) and Gelman et al. (2004, 587); the following discussion is based on Bernardo and Smith (1994, 287).

To prove that $p(\theta|bfy)$ tends to a normal distribution as $n \to \infty$, the general strategy is to first take a Taylor series expansion of the posterior distribution. Then, after ignoring higher order terms in the expansion that disappear asymptotically (and this is what the regularity conditions ensure), we have something recognizable as a normal distribution. So, consider a parameter vector $\theta \in \mathbb{R}^k$. Bayes Theorem tells us that a posterior distribution is proportional to a prior times a likelihood, or $p(\theta|y) \propto p(\theta)p(y|\theta)$, which can be re-written as

$$p(\theta|y) \propto \exp \left( \log(p(\theta)) + \log p(y|\theta) \right)$$
At the maximum of the log prior and the log likelihood, denoted $\theta_0$ and $\hat{\theta}_n$, respectively, we have
\[
\frac{\partial \log p(\theta)}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial \log p(y|\theta)}{\partial \theta} = 0,
\]
respectively. Taylor series expansions around the respective log maxima yield
\[
\log p(\theta) = \log p(\theta_0) - \frac{1}{2}(\theta - \theta_0)'Q_0(\theta - \theta_0) + r_0
\]
\[
\log p(y|\theta) = \log p(y|\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)'Q_n(\theta - \hat{\theta}_n) + r_n
\]
where $r_0$ and $r_n$ are higher-order terms and
\[
Q_0 = -\frac{\partial^2 \log p(\theta_0)}{\partial \theta \partial \theta'} \bigg|_{\theta = \theta_0} \quad \text{and} \quad Q_n = -\frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta'} \bigg|_{\theta = \hat{\theta}_n}.
\]
Note that the leading terms in the respective Taylor series expansions are not functions of $\theta$. Thus,
\[
p(\theta|y) \propto \exp \left( -\frac{1}{2}(\theta - \theta_p)'Q_p(\theta - \theta_p) \right), \tag{B.1}
\]
where
\[
Q_p = Q_0 + Q_n \tag{B.2}
\]
and
\[
\theta_p = Q_p^{-1} \left( Q_0\theta_0 + Q_n\hat{\theta}_n \right). \tag{B.3}
\]
The right-hand side of equation B.1 is recognizable as the kernel of a multivariate normal distribution with mean vector $\theta_p$ and variance-covariance matrix $Q_p^{-1}$; see definition XXXX.

This result relies on an extremely useful identity that is used frequently in Bayesian statistics, called “completing the square”:

XXXX

**Proposition B.1** (Conjugate Prior for Mean, Normal Data, Variance Known). Let $y_i \sim iid N(\mu, \sigma^2)$, $i = 1, \ldots, n$, with $\sigma^2$ known, and $y = (y_1, \ldots, y_n)'$. If $\mu \sim N(\mu_0, \sigma_0^2)$ is the prior density for $\mu$, then $\mu$ has posterior density
\[
\mu|y \sim N \left( \frac{\mu_0\sigma_0^{-2} + \frac{\bar{y}}{\sigma^2}}{\sigma_0^{-2} + \frac{n}{\sigma^2}}, \left( \sigma_0^{-2} + \frac{n}{\sigma^2} \right)^{-1} \right).
\]

**Proof.** The prior for $\mu$ is
\[
p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left( \frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2 \right) \propto \exp \left( \frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2 \right).
\]
and the likelihood is
\[
\mathcal{L}(\mu; y, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right] 
\]
\[
\propto \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right].
\]

Via Bayes Rule, the posterior density is proportional to the prior density times the likelihood:
\[
p(\mu | y) \propto \exp \left[ -\frac{1}{2} \left( \frac{1}{\sigma_0^2} (\mu - \mu_0)^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \right].
\]

Note that
\[
\frac{1}{\sigma_0^2} (\mu - \mu_0)^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2
\]
\[
= \mu^2 + \frac{\mu_0^2}{\sigma_0^2} - 2\mu \mu_0 + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - n\mu^2 + 2\mu \sum_{i=1}^{n} y_i
\]
\[
= \mu^2 \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right) + \frac{\mu_0^2}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}
\]
\[
= \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) \left( \mu - \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right)^2 + \frac{\mu_0^2}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right) \left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right)
\]
\[
= \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) \left( \mu - \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right) + r
\]

where
\[
r = \frac{\mu_0^2}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right)^2
\]
is a set of terms that do not involve \( \mu \). Accordingly, the posterior density is
\[
p(\mu | y) \propto \exp \left[ -\frac{1}{2} \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) \left( \mu - \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \right)^2 \right]
\]
which is the kernel of a normal density for \( \mu \), with mean
\[
\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}
\]
\[
\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}
\]
and variance \( \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \). \( \triangleright \)
Proposition B.2 (Conjugate Priors for Mean and Variance, Normal Data). Let \( y_i \overset{iid}{\sim} N(\mu, \sigma^2) \), \( i = 1, \ldots, n \), and let \( \mathbf{y} = (y_1, \ldots, y_n)' \). If \( \mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0) \) is the (conditional) prior density for \( \mu \), and \( \sigma^2 \sim \text{Inverse-Gamma}(\nu_0, \sigma_0^2/2) \) is the prior density for \( \sigma^2 \) then

\[
\begin{align*}
\mu|\sigma^2, \mathbf{y} &\sim N \left( \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n}, \frac{\sigma^2}{n_0 + n} \right) \\
\sigma^2|\mathbf{y} &\sim \text{Inverse-Gamma} \left( \frac{\nu_0 + n}{2}, \frac{1}{2} \nu_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{y})^2 \right).
\end{align*}
\]

where \( s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n - 1) \).

Proof. Consider the prior \( p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2) \) where \( p(\mu|\sigma^2) \) is a \( N(\mu_0, \sigma^2/n_0) \) density and \( p(\sigma^2) \) is an Inverse-Gamma density with parameters \( \nu_0/2 \) and \( \nu_0 \sigma_0^2/2 \). That is,

\[
p(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi \frac{\sigma^2}{n_0}}} \exp \left[ -\frac{n_0}{2\sigma^2} (\mu - \mu_0)^2 \right]
\]


\[
\propto (\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} n_0 (\mu - \mu_0)^2 \right],
\]

and

\[
p(\sigma^2) = \left( \frac{\nu_0 \sigma_0^2}{2} \right)^{\nu_0/2} (\sigma^2)^{-\left( \frac{\nu_0}{2} + 1 \right)} \exp \left( -\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right)
\]


\[
\propto (\sigma^2)^{-\left( \frac{\nu_0}{2} + 1 \right)} \exp \left( -\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right).
\]

Thus the prior density is

\[
p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)
\]


\[
\propto (\sigma^2)^{-1/2}(\sigma^2)^{-\left( \frac{\nu_0}{2} + 1 \right)} \exp \left[ -\frac{1}{2\sigma^2} \left( \nu_0 \sigma_0^2 + n_0 (\mu - \mu_0)^2 \right) \right].
\]

The likelihood is

\[
\mathcal{L}(\mu, \sigma^2; \mathbf{y}) = p(\mathbf{y} | \mu, \sigma^2) = \prod_{i=1}^n p(y_i; \mu, \sigma^2)
\]


\[
= \prod_{i=1}^n \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right]
\]


\[
\propto (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n (y_i - \mu)^2 \right) \right]
\]


\[
= (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n y_i^2 + n\mu^2 - 2\mu \sum_{i=1}^n y_i \right) \right].
\]
Note that

\[ \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu n\bar{y} \]

\[ = \sum_{i=1}^{n} y_i^2 + n\mu^2 - 2\mu n\bar{y} + n\bar{y}^2 - 2n\bar{y}^2 \]

\[ = \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \bar{y}^2 - 2n\bar{y}^2 + n\mu^2 + n\bar{y}^2 - 2\mu n\bar{y} \]

\[ = \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \bar{y}^2 - \sum_{i=1}^{n} 2\bar{y} y_i + n\mu^2 + n\bar{y}^2 - 2\mu n\bar{y} \]

\[ = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2, \]

and so

\[ L(\mu, \sigma^2; y) \propto (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right) \right] \]

\[ = (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( (n-1)s^2 + n(\bar{y} - \mu)^2 \right) \right]. \]

where \( s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1) \) is the conventional, unbiased sample estimate of the variance \( \sigma^2 \). Then by Bayes Rule, the posterior density is proportional to the prior density times the likelihood:

\[ p(\mu, \sigma^2|y) \propto p(\mu|\sigma^2)p(\sigma^2)p(y|\mu, \sigma^2) \]

\[ \propto (\sigma^2)^{-1/2}(\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp \left[ -\frac{1}{2\sigma^2} \left( v_0\sigma_0^2 + n_0(\mu - \mu_0)^2 \right) \right] \]

\[ \times (\sigma^2)^{-n/2} \exp \left[ -\frac{1}{2\sigma^2} \left( (n-1)s^2 + n(\bar{y} - \mu)^2 \right) \right] \]

\[ = (\sigma^2)^{-1/2}(\sigma^2)^{-\left(\frac{\nu_0+n}{2}+1\right)} \exp \left[ -\frac{1}{2\sigma^2} \left( v_0\sigma_0^2 + (n-1)s^2 + n_0(\mu - \mu_0)^2 + n(\bar{y} - \mu)^2 \right) \right]. \]
Note that
\[ n_0(\mu - \mu_0)^2 + n(\bar{y} - \mu)^2 \]
\[ = n_0\mu^2 + n_0\mu_0^2 - 2n_0\mu\mu_0 + n\bar{y}^2 + n\mu^2 - 2n\bar{y}\mu \]
\[ = \mu^2(n_0 + n) - 2\mu(n_0\mu_0 + n\bar{y}) + n_0\mu_0^2 + n\bar{y}^2 \]
\[ = (n_0 + n) \left[ \mu^2 - \frac{2\mu(n_0\mu_0 + n\bar{y})}{n_0 + n} + \frac{n_0\mu_0^2}{n_0 + n} + \frac{n\bar{y}^2}{n_0 + n} \right] \]
\[ = (n_0 + n) \left[ \mu^2 - \frac{2\mu(n_0\mu_0 + n\bar{y})}{n_0 + n} + \left( \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 - \left( \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 + \frac{n_0\mu_0^2}{n_0 + n} + \frac{n\bar{y}^2}{n_0 + n} \right] \]
\[ = (n_0 + n) \left[ \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} + n_0\mu_0^2 + n\bar{y} \right] \]
\[ = (n_0 + n) \left[ \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} + n_0\mu_0^2 + n\bar{y} \right] \]
\[ = (n_0 + n) \left[ \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{y})^2 \right]. \]

Substituting into the expression for the posterior density,
\[ p(\mu, \sigma^2 | y) \]
\[ \propto (\sigma^2)^{-1/2} (\frac{\sigma^2}{\sigma^2})^{-\left(\frac{m}{2} + \frac{n}{2} + 1\right)} \]
\[ \times \exp \left[ \frac{-1}{2\sigma^2} \left( \frac{n_0}{\sigma^2} + (n - 1)s^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{y})^2 + (n_0 + n) \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 \right) \right] \]
\[ \propto (\sigma^2)^{-1/2} \exp \left[ \frac{-n_0 + n}{2\sigma^2} \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 \right] \]
\[ \times (\sigma^2)^{-\left(\frac{m}{2} + \frac{1}{2} + 1\right)} \exp \left[ \frac{-v_0\sigma^2}{2\sigma^2} - \frac{(n_0 n)(\mu_0 - \bar{y})^2}{2\sigma^2} \right], \tag{B.4} \]

The first term in equation (B.4) is the kernel of a normal density over \( \mu \), conditional on \( \sigma^2 \), i.e.,
\[ p(\mu | \sigma^2, y) \propto (\sigma^2)^{-1/2} \exp \left[ \frac{-n_0 + n}{2\sigma^2} \left( \mu - \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \right)^2 \right], \]

or, equivalently,
\[ \mu | \sigma^2, y \sim N \left( \frac{n_0\mu_0 + n\bar{y}}{n_0 + n}, \frac{\sigma^2}{n_0 + n} \right). \]

The second term in equation (B.4) is the kernel of an inverse-Gamma density, such that the marginal posterior density of \( \sigma^2 \) is
\[ p(\sigma^2 | y) \propto (\sigma^2)^{-\left(\frac{m}{2} + \frac{1}{2} + 1\right)} \exp \left[ \frac{-v_0\sigma^2}{2\sigma^2} - \frac{(n_0 n)(\mu_0 - \bar{y})^2}{2\sigma^2} \right]. \]
or equivalently,

\[ \sigma^2 | y \sim \text{Inverse-Gamma} \left( \frac{v_0 + n}{2}, \frac{1}{2} \left[ v_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{y})^2 \right] \right). \]

**Proposition B.3.** Under the conditions of Proposition B.2, the marginal posterior density of \( \mu \) is a student-t density, with location parameter

\[ E(\mu|\sigma^2, y) = \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n}, \]

scale parameter \( \sigma_1^2/(n_0 + n) \), and \( v_0 + n \) degrees of freedom, where \( \sigma_1^2 = S_1/(v_0 + n) \) and

\[ S_1 = v_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_1} (\bar{y} - \mu_0)^2. \]

**Proof.** The marginal posterior density of \( \mu \) is

\[ p(\mu|y) = \int_0^\infty p(\mu|\sigma^2, y) p(\sigma^2|y) d\sigma^2 \]

where \( p(\mu|\sigma^2, y) p(\sigma^2|y) \) is the joint posterior density given in equation B.4 of Proposition B.2. Letting

\[
\begin{align*}
n_1 &= n_0 + n, \\
\mu_1 &= \frac{n_0}{n_1} \mu_0 + \frac{n}{n_1} \bar{y}, \\
v_1 &= v_0 + n, \\
S_1 &= v_1 \sigma_1^2 \\
&= v_0 \sigma_0^2 + (n - 1)s^2 + \frac{n_0 n}{n_1} (\bar{y} - \mu_0)^2
\end{align*}
\]

the joint posterior can be rewritten as

\[ p(\mu, \sigma^2|y) \propto (\sigma^2)^{-1/2} (\sigma^2)^{-(\nu_1/2 + 1)} \exp \left[ -\frac{1}{2\sigma^2} \left( S_1 + n_1(\mu - \mu_1)^2 \right) \right], \]

and so the marginal posterior density for \( \mu \) is

\[ p(\mu|y) \propto \int_0^\infty (\sigma^2)^{-1/2} (\sigma^2)^{-(\nu_1/2 + 1)} \exp \left[ -\frac{1}{2\sigma^2} \left( S_1 + n_1(\mu - \mu_1)^2 \right) \right] d\sigma^2. \]

Let \( D = S_1 + n_1(\mu - \mu_1)^2 \) and re-express the integral via the change of variables to \( z = D/(2\sigma^2) \), noting that \( \sigma^2 = \frac{1}{2} Dz^{-1} \propto Dz^{-1} \) and \( d\sigma^2/dz \propto Dz^{-2} \).

\[
\begin{align*}
p(\mu|y) &\propto \int_0^\infty (\sigma^2)^{-(\nu_1/2 + 1)} \exp(-z) \frac{d\sigma^2}{dz} dz \\
&\propto \int_0^\infty (\sigma^2)^{-(\nu_1/2 + 1)} \exp(-z) Dz^{-2} dz \\
&\propto D^{-\nu_1/2} \int_0^\infty z^{(\nu_1/2 - 1)} \exp(-z) dz.
\end{align*}
\]
The integral is the Gamma function (see Definition A.10) and evaluates to $\Gamma(\nu_1/2)$, which is not a function of $\mu$. Thus,
\[
p(\mu|y) \propto D^{-\nu_1/2} = \left[ S_1 + n_1(\mu - \mu_1)^2 \right]^{-\nu_1/2}
\]
\[
\propto \left[ 1 + \frac{n_1(\mu - \mu_1)^2}{S_1} \right]^{-\nu_1/2}
\]
\[
\propto \left[ 1 + \frac{1}{\nu_1} \frac{(\mu - \mu_1)^2}{\sigma_1^2/n_1} \right]^{-\nu_1/2}
\]
which is recognizable as the kernel of a (unstandardized) student-$t$ density (see Definition A.16), with location parameter $\mu_n$, scale parameter $\sigma_1^2/n_1$ and $\nu_1$ degrees of freedom.

Remark. As Lee (2004, 63-64) points out, it is more usual to express the result in Proposition B.3 in terms of a standardized student-$t$ density. That is, let
\[
t = \frac{\mu - \mu_1}{\sqrt{\sigma_1^2/n_1}}.
\]
The change of variables from $\mu$ to $t$ is simple. Following Proposition A.2, we seek
\[
\frac{p(t|y)}{f_t} = \left| \frac{\mu}{J(t)} \right|
\]
where $f_t = p(t|y)$, $f_\mu \propto p(\mu|y)$ is derived above, and $|J(\cdot)| = |d\mu/dt|$ is the Jacobian of the transformation. Note that $\phi(t) = t\sqrt{\sigma_1^2/n_1} + \mu_1$ and the Jacobian is a constant with respect to $t$. Thus
\[
p(t|y) \propto \left[ 1 + \frac{1}{\nu_1} \frac{(t\sqrt{\sigma_1^2/n_1} + \mu_1 - \mu)^2}{\sigma_1^2/n_1} \right]^{-\nu_1/2}
\]
\[
\propto \left[ 1 + \frac{t^2}{\nu_1} \right]^{-\nu_1/2}
\]
which is the kernel of a standardized $t$ density with $\nu_1$ degrees of freedom. That is,
\[
\frac{\mu - \mu_1}{\sqrt{\sigma_1^2/n_1}} \sim t_{\nu_1}.
\]
References


REFERENCES


REFERENCES


Lindley, Dennis V. 1965. *Introduction to probability and statistics from a Bayesian viewpoint*. Cambridge: Cambridge University Press.


Sekhon, Jasjeet S. 2005. “Making Inference from $2 \times 2$ Tables: The Inadequacy of the Fisher Exact Test for Observational Data and a Bayesian Alternative.” Typescript. Survey Research Center, University of California, Berkeley.

Skocpol, Theda. 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge: Cambridge University Press.


