1

The foundations of Bayesian inference

In this chapter I elaborate on the overview of Bayesian statistical inference provided in the introduction. 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 distinguish the subjective notion of probability from the classical, objective or frequentist approach, before stating Bayes Theorem in the various forms it is used in statistical settings. I then review how Bayesian data analysis is actually done. 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 about unknowns. Of course, there is much to be said on the implementation of this procedure in any specific application, and these details are the subjects of later chapters. The discussion in this chapter deals 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? How does hypothesis testing work in the Bayesian approach? What kinds of considerations motivate model specification in the Bayesian approach?

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 defintion of probability, which I reproduce here (with measure-theoretic details omitted; see the Appendix for a more rigorous set of definitions). If Ω 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)

Bayesian Analysis for the Social Sciences S. Jackman

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 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 identify at least *four* interpretations of probability (e.g., Galavotti 2005). 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

In classical statistics probability is often understood 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. Thus, 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 probabilty of A, where m is the number of times we observe the event A and n is the number of repetitions. Given this definition of probability, we can understand why classicial 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.

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 (quoted in Barnett 1999, 76).

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 examples from games of chance, and the naïve view of the history of statistics is that interest in games of chance spurred the development of probability (e.g. Todhunter 1865), and, in particular, the frequentist interpretation of probability. That is, for simple games of chance it is feasible to enumerate the set of possible outcomes, and hence generate statements of the likelihood of particular outcomes in relative frequency terms (e.g. the 'probability' of throwing a seven with two dice, an important quantity in craps). 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' (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 neighborhourhood of certainty and demonstration, quite down to improbability and unlikeliness, even to the confines of impossibility'. 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 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 is not just *any* set of subjective beliefs, but beliefs that conform to the axioms of probability. The Ramsey-de Finetti Theorem states that if $p_1, p_2, \ldots$ are a set of betting quotients on hypotheses $h_1, h_2, \ldots$, then if the p_j do not satisfy the probability axioms, there exists 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 (e.g. 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 new information (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 suprise 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 infintisimal magnitude. Indeed, if this occured 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 coming up heads, which may change while observing a sequence of coin flips, via Bayes' Theorem.

Bayesian probability statements are thus about states of mind over states of the world, and not about states of the world *per se*. 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 a '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. What 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 not the eighth president) or one (if Jackson was the eighth president). Non-trivial frequentist probabilities, it seems, are reserved for phenomena that are standardized and repeatable (e.g. the exemplars 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 occured, 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 United States in the next five years?
- What is the probability that over the course of my life, someone I know will be incarcerated?

All of these are perfectly legitimate and interesting social-scientific questions, but for which the objectivist/frequentist position apparently offers no helpful answer.

With this distinction between objective and subjective probability firmly in mind, we now consider how Bayes Theorem tells us how we should rationally update subjective, probabilistic beliefs in light of evidence.

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, in work first published in 1937. The thought experiment 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 retelling 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, ch. 2) adopts the conditional axiomization of probability due to Rényi (1970) and see also the treatment in Lee (2004, ch. 1).

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)

$$\begin{aligned} P(B) &= P(A \cap B) + P(\sim A \cap B) \\ &= P(B|A)P(A) + P(B| \sim A)P(\sim A) \end{aligned}$$

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. ◁

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 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). ◁

■ Example 1.1

Drug testing. Elite athletes are routinely tested for the presence of banned performance-enhacing 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,

$$\begin{aligned} P(H_U|E) &= \frac{P(H_U)P(E|H_U)}{\sum_{i\in\{U,\sim U\}} P(H_i)P(E|H_i)} \\ &= \frac{.03 \times .95}{(.03 \times .95) + (.97 \times .10)} \\ &= \frac{.0285}{.0285 + .097} \\ &\approx .23 \end{aligned}$$

That is, in light of (1) the positive test result (the evidence, 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. Even a casual, visual inspection of district level election results suggests that there are $J = 3$ 'clumps' or classes of districts: Republican seats ($T_i = 1$), Democratic seats ($T_i = 2$), and a small cluster of extremely Democratic seats ($T_i = 3$); see Figure 6.6 in Example 6.8. Let y_i be the proportion of the two-party vote won by the Democratic candidate for Congress in district i, and λ_j be the proportion of districts in class j (i.e. $\sum_{j=1}^{J} \lambda_j = 1$). We will assume that the distribution of the y_i within each of the $J = 3$ 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 μ_j, σ_j and λ_j (to two decimal places, see Example 6.8 for details):

Class	μ_j	σ_j	λ_j
1. Republican	.35	.08	.49
2. Democratic	.66	.10	.46
3. Extremely Democratic	.90	.03	.05

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

$$\begin{aligned} P(T_i = j|y_i) &= \frac{P(T_i = j) \cdot P(y_i|T_i = j)}{\sum_{k=1}^{J} \left[P(T_i = k) \cdot P(y_i|T_i = k)\right]} \\ &= \frac{\lambda_j \cdot \phi([y_i - \mu_j]/\sigma_j)}{\sum_{k=1}^{J} \left[\lambda_k \cdot \phi([y_i - \mu_k]/\sigma_k)\right]} \end{aligned} \tag{1.1}$$

where $\phi(y; \mu, \sigma)$ is the normal probability density function (see Definition B.30).

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 US 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: to two decimal places we have

$$P(T_i = 1|y_i = .56) = \frac{.49 \times \phi([.56 - .35]/.07)}{.49 \times \phi([.56 - .35]/.08) + .46 \times \phi([.56 - .66]/.10) + .05 \times \phi([.56 - .90]/.03)}$$

$$= \frac{.49 \times .11}{(.49 \times .11) + (.46 \times 2.46) + (.05 \times 9.7 \times 10^{-27})}$$

$$= \frac{.05}{1.18} = .04$$

$$P(T_i = 2|y_i = .56) = \frac{.46 \times 2.46}{1.18} = .96$$

$$P(T_i = 3|y_i = .56) = \frac{.05 \times 9.7 \times 10^{-27}}{1.18} \approx 0$$

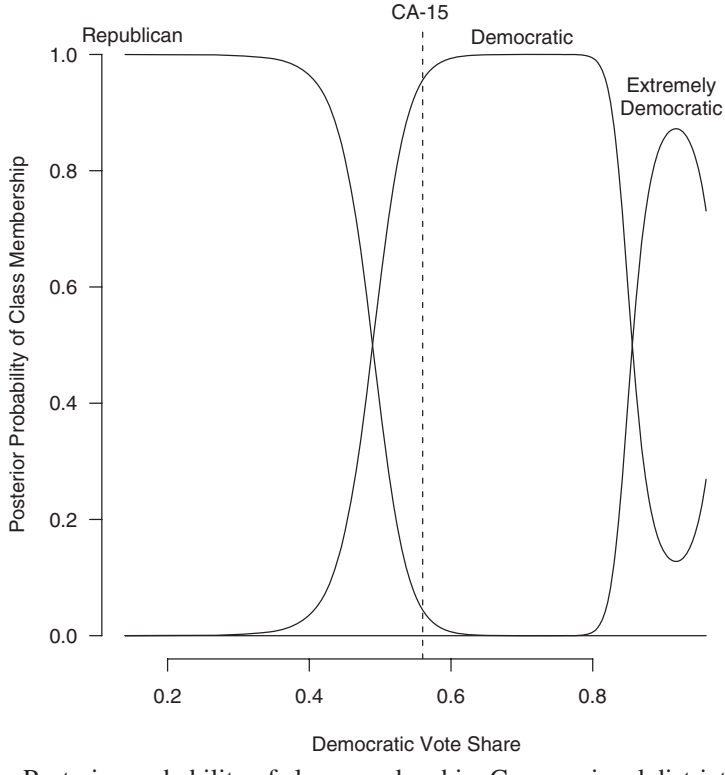

Figure 1.1 Posterior probability of class membership, Congressional districts. The probability that CA-15 ($y_i = .56$) belongs to the 'Democratic' class is .96.

That is, the posterior probability that CA-15 belongs to the 'Democratic' class is .96. Note that the result in CA-15, $y_i = .56$ lies a long way from the 'extremely Democratic' class ($\mu_3 = .90, \sigma_3 = .03$) and so the probability of assigning CA-15 to that class is virtually zero.

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. We will return to this example in Chapter 5.

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 θ and denote the data available for analysis as $\mathbf{y} = (y_1, \ldots, y_n)'$. In the case of continuous parameters, beliefs about the parameter are represented as *probability density functions* or pdfs (see Definition B.12); we denote the prior pdf as $p(\theta)$ and the posterior pdf as $p(\theta|\mathbf{y})$.

Then, Bayes Theorem for a continuous parameter is as follows:

Proposition 1.5 (Bayes Theorem, continuous parameter).

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}$$

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

$$p(\theta, y) = p(\theta|y)p(y) = p(y|\theta)p(\theta), \tag{1.2}$$

where all these densities are assumed to exist and have the properties $p(z) > 0$ and $\int p(z)dz = 1$ (i.e. are *proper* probability densities, see Definitions B.12 and B.13). 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 follows, perhaps the most important formula in this book:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta), \tag{1.3}$$

where the constant of proportionality is

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

i.e. ensuring that the posterior density integrates to one, as a proper probability density must (again, see Definitions B.12 and B.13).

The first term on the right hand side of Equation 1.3 is the *likelihood function* (see Definition B.16), the probability density of the data $\mathbf{y}$, considered as a function of θ. 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 of Bayes Rule highlights a particularly elegant feature of the Bayesian approach, showing how the likelihood function $p(\mathbf{y}|\theta)$ can be 'inverted' to generate a probability statement about θ, given data $\mathbf{y}$.

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 θ 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 θ 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 θ, the kind of prior we might specify when we have no prior information about the value of θ, and hence no way to *a priori* prefer one set of values for θ 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.).

Posterior densities as precision-weighted combination of prior information and likelihood

The other panels in Figure 1.2 display how Bayesian inference works with more or less *informative priors* for θ. In the top left of Figure 1.2 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, which

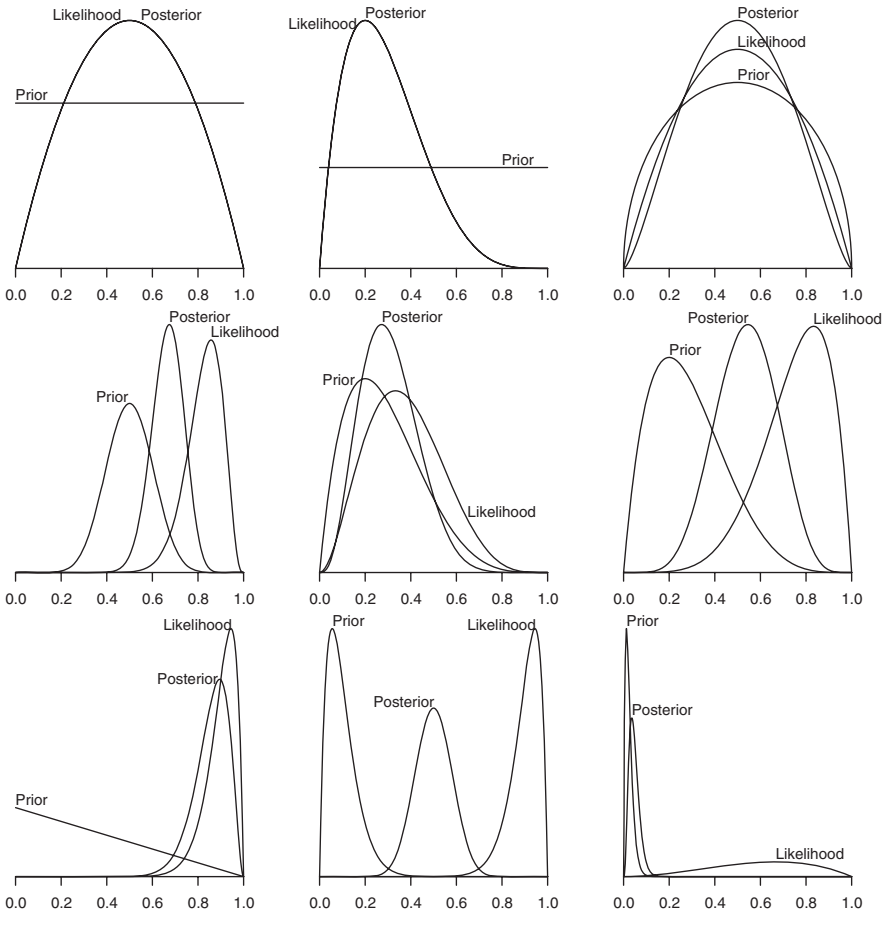

Figure 1.2 Priors, likelihoods and posterior densities. Each panel shows a prior density, a likelihood, and a posterior density over a parameter $\theta \in [0, 1]$. In the top two panels on the left the posterior and the likelihood coincide, since the prior is uniform over the parameter space.

we define in the next section. 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, and make Bayesian analysis mathematically and computationally quite simple.

1.4.1 Conjugate priors

Since conjugacy is such an important concept in Bayesian statistics, it is worth pausing to sketch a definition:

Definition 1.2 *Suppose a prior density $p(\theta)$ belongs to a class of parametric of densities, $\mathcal{F}$. Then the prior density is said to be conjugate with respect to a likelihood $p(\mathbf{y}|\theta)$ if the posterior density $p(\theta|\mathbf{y})$ is also in $\mathcal{F}$.*

Of course, this definition rests on the unstated definition of a 'class of parametric densities', and so is not as complete as one would prefer, but a thorough explanation involves more technical detail than is warranted for now. Examples are perhaps the best way to illustrate the simplicity that conjugacy brings to a Bayesian analysis. And to this end, all the examples in Chapter 2 use priors that are conjugate with respect to their respective likelihoods.

In particular, the examples in Figure 1.2 show the results of a Bayesian analysis of binomial data (n independent realizations of a binary process, also known as Bernoulli trials, such as coin flipping), for which the unknown parameter is $\theta \in [0, 1]$, the probability of a 'success' on any given trial. For the likelihood function formed with binomial data, any Beta density (see Definition B.28) over θ is a conjugate prior: that is, if prior beliefs about θ can be represented as a Beta density, then after those beliefs have been updated (via Bayes Rule) in light of the binomial data, posterior beliefs about θ are also characterized by a Beta density. In Section 2.1 we consider the Bayesian analysis of binomial data in considerable detail.

For now, one of the important features of conjugacy is the one that appears graphically 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. That is, Bayesian analysis with conjugate priors over a parameter θ is equivalent to taking a *precision-weighted average* of prior information about θ and the information in the data about θ.

Thus, when prior beliefs about θ are 'vague', 'diffuse', or, in the limit, uninformative, the posterior density 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 posterior has the same shape as the likelihood function. 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.

1.4.2 Bayesian updating with irregular priors

Figure 1.3 displays a series of prior and posterior densities for less standard cases, where the prior densities are not simple unimodal densities. In each instance, Bayes Rule applies as usual, with the posterior density being proportional to the prior density times the likelihood, and appropriately normalized such that the posterior density 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 density clearly represents the merger of prior and likelihood: with a mode just

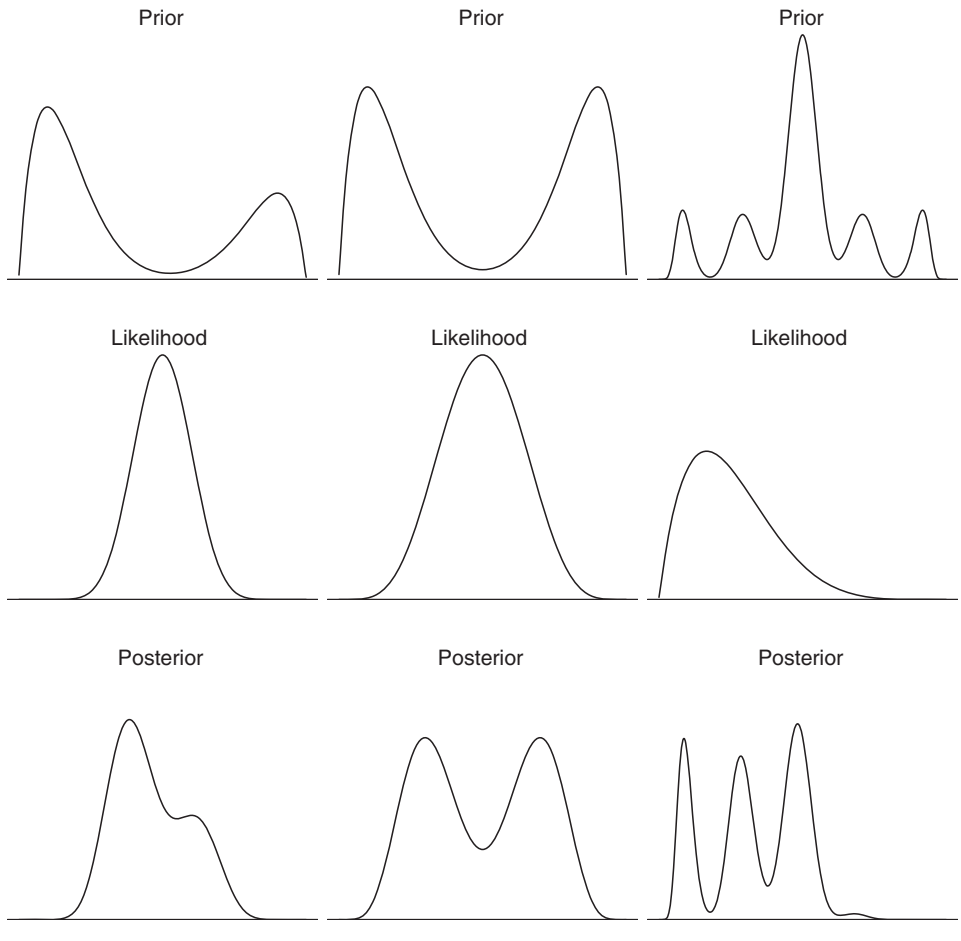

Figure 1.3 Priors, likelihoods and posterior densities for non-standard cases. Each column of panels shows the way Bayes Rule combines prior information (top) with information in the data (characterized by the likelihood, center) to yield a posterior density (lower panels).

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 θ 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 θ that receive high prior probability. In the right-hand column of Figure 1.3 an extremely flamboyant prior distribution (but one that is nonetheless symmetric about its mean) combines with the skewed likelihood to produce the trimodal posterior density, 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 densities 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 (see Chapter 3).

1.4.3 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 zero probability 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 θ 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 θ 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 positive probability over all values of θ 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, 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 to 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,

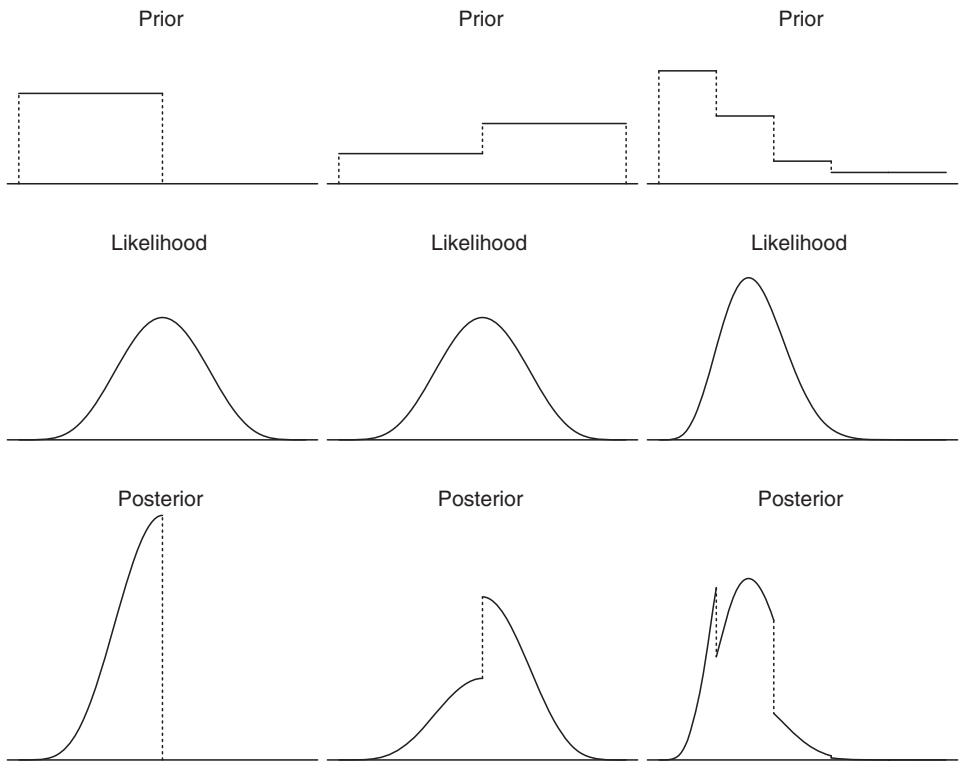

Figure 1.4 Discontinuous prior and posterior densities. Each column of panels shows the way Bayes Rule combines prior information (top) with information in the data (characterized by the likelihood, center) to yield a posterior density (lower panels). The dotted lines indicate discontinuities.

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 θ 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.

1.4.4 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 cojugate 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 (Blackwell and Dubins 1962). In this way Bayesian analysis has been proclaimed as a model for scientific practice (e.g. Howson and Urbach 1993; Press 2003) 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

$$\begin{aligned} P(H_U|E) &= \frac{P(H_U)P(E|H_U)}{\sum_{i\in\{U,\sim U\}} P(H_i)P(E|H_i)} \\ &= \frac{.001 \times .95}{(.001 \times .95) + (.999 \times .10)} \\ &= \frac{.00095}{.000095 + .0999} \\ &\approx .009 \end{aligned}$$

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:

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

while your beliefs are updated to

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

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 postitive 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 postive 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 1.2.

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 $p(\theta|\mathbf{y})$ is just that, a probability density. Given a subjectivist interpretation of probabilty that most Bayesians adopt, the 'randomness' summarized by the posterior density is a reflection of the researcher's uncertainty over θ, conditional on having observed data $\mathbf{y}$.

Contrast the frequentist approach, in which θ is not random, but a fixed (but unknown) property of a population from which we randomly sample data $\mathbf{y}$. Repeated applications of the sampling process, if undertaken, would yield different $\mathbf{y}$, and different sample based estimates of θ, denoted $\hat{\theta} = \hat{\theta}(\mathbf{y})$, this notation reminding us that estimates of parameters are functions of data. In the frequentist scheme, the $\hat{\theta}(\mathbf{y})$ vary randomly across data sets (or would, if repeated sampling was undertaken), while the parameter θ 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 θ in light of the data available for analysis, $\mathbf{y}$?" The quantity $\hat{\theta}(\mathbf{y})$ 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 θ is a feature

of the data that is usually helpful in *computing* the posterior distribution for θ. And, under some special circumstances, a least squares or maximum likelihood estimate of θ, $\hat{\theta}(\mathbf{y})$, will correspond to a Bayes estimate of θ (see Section 1.6.1). But the critical point to grasp is that in the Bayesian approach, the roles of θ and $\hat{\theta}$ are reversed relative to their roles in classical, frequentist inference: θ 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

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

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 θ 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

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\theta_1, \ldots, \theta_k|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}) \tag{1.4}$$

Direct visualization is no longer feasible once $k > 2$: density plots or histograms have two-dimensional counterparts (e.g. contour or image plots, used throughout this book, and perspective plots), but we simply run out of dimensions at this point. As the dimension of the parameter vector 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 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 density (or the marginal posterior densities specific to particular parameters) are 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 congujate, modern computational power lets us deploy Monte Carlo methods to compute these features of posterior densities; see Chapter 3. Finally, it should be noted that with large sample sizes, provided the prior is not degenerate, the posterior densities are usually well approximated by normal densities, for which it is straightforward to compute numerical summaries (see Section 1.7). 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 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 rationalized 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 1.3 (Loss Function). *Let Θ be a set of possible states of nature θ, and let $a \in \mathcal{A}$ be actions availble to the researcher. Then define $l(\theta, a)$ as the loss to the researcher from taking action a when the state of nature is θ.*

Recall that in the Bayesian approach, the researcher's beliefs about plausible values for θ are represented with a probability density function (or a probability mass function, if θ take discrete values), and, in particular, after looking at data $\mathbf{y}$, beliefs about θ are represented by the posterior density $p(\theta|\mathbf{y})$. Generically, let $p(\theta)$ be a probability density over θ, which in turn induces a density over losses. Averaging the losses over beliefs about θ yields the Bayesian expected loss (Berger 1985, 8):

Definition 1.4 (Bayesian expected loss). *If $p(\theta)$ is the probability density for $\theta \in \Theta$ at the time of decision making, the Bayesian expected loss of an action a is*

$$\varrho(p(\theta), a) = E[l(\theta, a)] = \int_{\Theta} l(\theta, a)\, p(\theta) d\theta.$$

A special case is where the density p in Definition 1.4 is a posterior density:

Definition 1.5 (Posterior expected loss). *Given a posterior density for θ, $p(\theta|\mathbf{y})$, the posterior expected loss of an action a is $\varrho(p(\theta|\mathbf{y}), a) = \int_{\Theta} l(\theta, a)\, p(\theta|\mathbf{y}) d\theta$.*

A Bayesian rule for choosing among actions $\mathcal{A}$ is to select $a \in \mathcal{A}$ 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 $a \in \mathcal{A}$ 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 corresponding 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. We briefly consider some well-studied cases.

Definition 1.6 (Quadratic loss). *If $\theta \in \Theta$ is a parameter of interest, and $\tilde{\theta}$ is an estimate of θ, then $l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$ is the quadratic loss arising from the use of the estimate $\tilde{\theta}$ instead of θ.*

With quadratic loss, we obtain the following useful result:

Proposition 1.6 (Posterior mean as a Bayes estimate under quadratic loss). *Under quadratic loss the Bayes estimate of θ is the mean of the posterior density, i.e. $\tilde{\theta} = E(\theta|\mathbf{y}) = \int_\Theta \theta p(\theta|\mathbf{y}) d\theta$.*

Proof. Quadratic loss (Definition 1.6) implies that the posterior expected loss is

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

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

$$\begin{aligned}\varrho(\theta, \tilde{\theta}) &= \int_\Theta \theta^2 p(\theta|\mathbf{y}) d\theta + \tilde{\theta}^2 \int_\Theta p(\theta|\mathbf{y}) d\theta - 2\tilde{\theta} \int_\Theta \theta p(\theta|\mathbf{y}) d\theta \\ &= \int_\Theta \theta^2 p(\theta|\mathbf{y}) d\theta + \tilde{\theta}^2 - 2\tilde{\theta} E(\theta|\mathbf{y}),\end{aligned}$$

Differentiate with respect to $\tilde{\theta}$, noting that the first term does not involve $\tilde{\theta}$. Then set the derivative to zero and solve for $\tilde{\theta}$ to establish the result. ◁

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

Definition 1.7 (Multidimensional quadratic loss). *If $\boldsymbol{\theta} \in \mathbb{R}^K$ is a parameter, and $\tilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, then the (multidimensional) quadratic loss is $l(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{Q} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$ where $\mathbf{Q}$ is a positive definite matrix.*

Proposition 1.7 (Multidimensional posterior mean as Bayes estimate). *Under quadratic loss (Definition 1.7), the posterior mean $E(\boldsymbol{\theta}|\mathbf{y}) = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ is the Bayes estimate of $\boldsymbol{\theta}$.*

Proof. The posterior expected loss is $\varrho(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \int_{\boldsymbol{\Theta}} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{Q} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. Differentiating with respect to $\tilde{\boldsymbol{\theta}}$ yields $2\mathbf{Q} \int_{\boldsymbol{\Theta}} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. Setting the derivative to zero and re-arranging yields $\int_{\boldsymbol{\Theta}} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = 0$ or $\int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \tilde{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. The left-hand side of this expression is just the mean of the posterior density, $E(\boldsymbol{\theta}|\mathbf{y})$, and so $E(\boldsymbol{\theta}|\mathbf{y}) = \int_{\boldsymbol{\Theta}} \tilde{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. ◁

Remark. This result holds irrespective of the specific weighting matrix $\mathbf{Q}$, provided $\mathbf{Q}$ 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 θ. Specifically, Proposition 1.6 rationalizes the choice of the mean of the posterior density as a Bayes estimate.

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

Definition 1.8 (Linear loss). *If $\theta \in \Theta$ is a parameter, and $\tilde{\theta}$ is a point estimate of θ, 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 linear loss results when $k_0 \neq k_1$.

Proposition 1.8 (Bayes estimates under linear loss). *Under linear loss (definition 1.8), the Bayes estimate of θ is the $k_1/(k_0 + k_1)$ quantile of $p(\theta|\mathbf{y})$, 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

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

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

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

Adding $k_0 \int_{\{\theta \leq \tilde{\theta}\}} p(\theta|\mathbf{y}) d\theta$ to both sides yields $k_0 = (k_0 + k_1) \int_{\{\theta \leq \tilde{\theta}\}} p(\theta|\mathbf{y}) d\theta$ and so re-arranging yields $\int_{\{\theta < \tilde{\theta}\}} p(\theta|\mathbf{y}) d\theta = k_0/(k_0 + k_1)$. ◁

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

■ **Example 1.4**

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

$$\varrho(\theta, \tilde{\theta}) = \begin{cases} \theta - \tilde{\theta} & \text{if } \theta > \tilde{\theta} \\ 2.5(\tilde{\theta} - \theta) & \text{if } \theta \leq \tilde{\theta} \end{cases}$$

Ability is measured on an arbitrary scale, normalized to have mean zero and standard deviation one across the applicant pool. Suppose that for applicant i, $p(\theta_i|\mathbf{y}) \approx N(1.8, 0.4^2)$, while for applicant j, $p(\theta_j|\mathbf{y}) \approx N(2.0, 1.0^2)$; i.e. there is considerably greater posterior uncertainty as to the ability of applicant j. Given the professors's loss function, the Bayes estimate of θ_i is the $1/(1 + 2.5) = .286$ quantile of the $N(1.8, 0.4^2)$ posterior density, or 1.57; for applicant j, the Bayes estimate is the .286 quantile of a $N(2.0, 1.0^2)$ density, or 1.43. Thus, although $E(\theta_j|\mathbf{y}) > E(\theta_i|\mathbf{y})$, the greater uncertainty associated with applicant j, when coupled with the asymmetric loss function, results in the professor assigning a higher Bayes estimate to applicant j than to applicant i (i.e. $\tilde{\theta}_i < \tilde{\theta}_j$).

1.6.2 Credible regions

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

Definition 1.9 (Credible region). *A region $C \subseteq \Omega$ such that $\int_C p(\theta)d\theta = 1 - \alpha$, $0 \leq \alpha \leq 1$ is a $100(1-\alpha)\%$ credible region for θ.*

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 $p(\theta)$ 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 $p(\theta)$. 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 $100(1-\alpha)\%$ credible region for any $\alpha \in (0, 1)$: any interval spanning $100(1-\alpha)$ 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 $p(\theta)$. 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 1.10 (Highest probability density interval). *A region* $C \subseteq \Omega$ *is a* $100(1-\alpha)\%$ *highest probability density region for* θ *under* $p(\theta)$ *if*

1. $P(\theta \in C) = 1 - \alpha$
2. $P(\theta_1) \geq P(\theta_2), \forall\, \theta_1 \in C,\ \theta_2 \notin C$

A $100(1-\alpha)\%$ HPD region for a symmetric, unimodal density is obviously unique and symmetric around the mode. In fact, if $p(\theta)$ is a univariate normal density, a HPD is the same as a interval around the mean:

■ **Example 1.5**

Suppose $p(\theta) \equiv N(a, b^2)$. Then a $100(1-\alpha)\%$ HPD region is the interval

$$(a - |z_\alpha|b,\ a + |z_\alpha|b)$$

where z_α is the α quantile of the standard normal density. With $\alpha = .05$, $|z_\alpha| \approx 1.96$, and a 95 % HPD corresponds to a 95 % interval; see Figure 1.5.

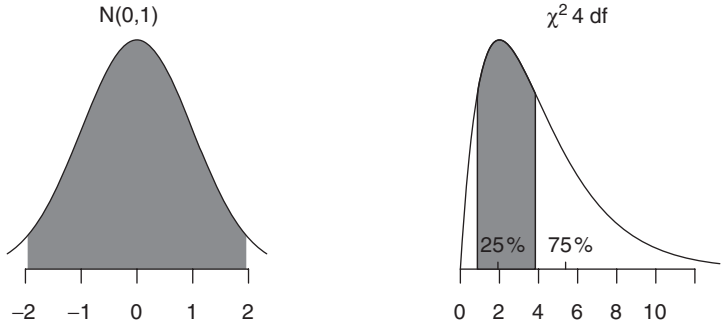

Figure 1.5 95 % HPD intervals for standard normal (left panel) and χ^2_4 densities.

Note that the correspondence between intervals and HPD intervals does not hold for non-symmetric densities, as we demonstrate with a simple example.

■ **Example 1.6**

The right panel of Figure 1.5 shows a χ^2 density with 4 degrees of freedom, and its 50 % HPD interval. Notice that the 50 % HPD interval is more concentrated around the

mode of the density, and has shorter length than the interval based on the 25th to 75th percentiles of the density.

As the next two examples demonstrate, (1) the HPD need not be a connected set, but a collection of disjoint intervals (say, if $p(\theta)$ is not unimodal), and (2) the HPD need not be unique.

■ **Example 1.7**

Extreme missingness in bivariate normal data. Consider the data in Table 1.1, where two variables (y_1 and 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).

Table 1.1 Twelve observations from a bivariate normal distribution.

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

Given the missing data pattern, what should we conclude about the correlation ρ between y_1 and y_2? For this particular example, with an uninformative prior for the covariance matrix of $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2)$, the posterior density for ρ is bimodal, as shown in Figure 1.6. The shaded areas represent half of the posterior density for ρ; the intervals supporting the shaded areas together constitute a 50 % HPD region for ρ, and are the disjoint intervals $(-.914, -.602)$ and $(.602, .914)$. We return to this example in more detail in Examples 5.11 and 6.6.

Figure 1.6 Bimodal posterior density for a correlation coefficient, and 50 % HPD.

■ **Example 1.8**

Non-unique HDRs. Suppose $\theta \sim \text{Uniform}(0, 1)$. Then any HPD region of content α is not unique, $\forall\ 0 < \alpha < 1$. See Figure 1.7. The shaded regions are both supported by 25 % HPDs, as are any other intervals of width .25 we might care to draw.

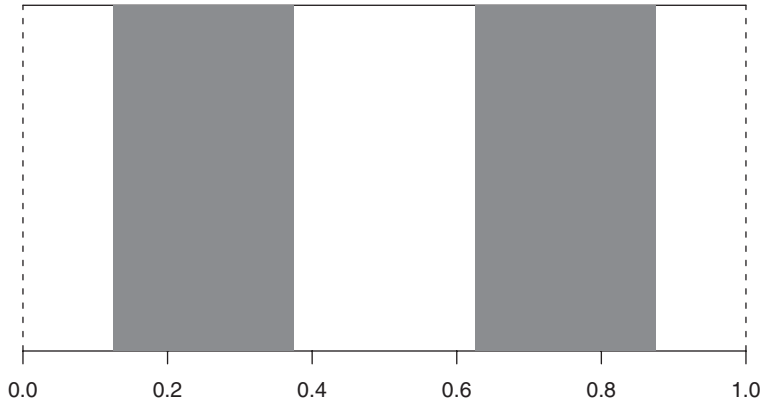

Figure 1.7 Uniform density and (non-unique) 25 % HPDs.

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. 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 Proposition 2.13.

1.7 Asymptotic properties of posterior distributions

As we have seen, Bayes Rule tells us how we ought to revise 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 set being analyzed grows without bound, the usual result is that the resulting sequence of posterior densities collapses to a spike on the true values of the parameters in the model under consideration.

Of course, some 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 gets infinitely large as an interesting mathematical exercise, but not particularly relevant to the inferential task at hand. This view holds that provided we update our beliefs via Bayes Rule in light of *this* data set, and with a model/likelihood appropriate to the data at hand (not a trivial matter), we are behaving rationally, and the

repeated sampling or asymptotic properties of our inferences are second order concerns. Some Bayesians even go further, arguing that models and parameters have no objective, exterior reality, but are mathematical fictions we conjure so as to help us make probability assignments over data (we explore this 'subjectivist' position further in §1.9), and so questions such as consistency are moot.

My own position – echoing that of Diaconis and Freedman (1986a, 11) – is that even subjectivist Bayesians ought to consider asymptotic properties of Bayes estimates, since if Bayesian inference is to be a model of scientific practice, we should be able to establish the convergence of (initially disparate) opinions as relevant evidence accumulates.

So what can we say about Bayesian inferences, asymptotically? The key idea here 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 data through the likelihood function, and the standard asymptotic properties of maximum likelihood estimators apply to the posterior density. These properties include

- consistency, at least in the sense that the posterior density is increasingly concentrated around the true parameter value as $n \to \infty$; or, in the additional sense of Bayes point estimators of θ (Section 1.6.1) being consistent;
- asymptotic normality, i.e. $p(\theta|\mathbf{y})$ 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, ch. 5) for statements of necessary regularity conditions and proofs of the main results and references to the literature. Diaconis and Freedman (1986a,b) provide some counter-examples to the consistency results; the 'incidental parameters' problem (Neyman and Scott 1948) is one such counter-example which we briefly return to in Section 9.1.2. I provide a brief illustration of 'Bayesian consistency' with two examples, below, and sketch a proof of a 'Bayesian central limit theorem' in the Appendix.

Bayesian consistency works as follows. Suppose the true value of θ is θ^*. Then provided the prior distribution $p(\boldsymbol{\theta})$ does not place zero probability mass on θ^* (say, for a discrete parameter), or on a neigborhood of θ^* (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 θ^*.

Figures 1.8 and 1.9 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 are collapsing to a spike on θ^*.

Although the scale used in Figures 1.8 and 1.9 doesn't make it clear, the likelihoods and posterior 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

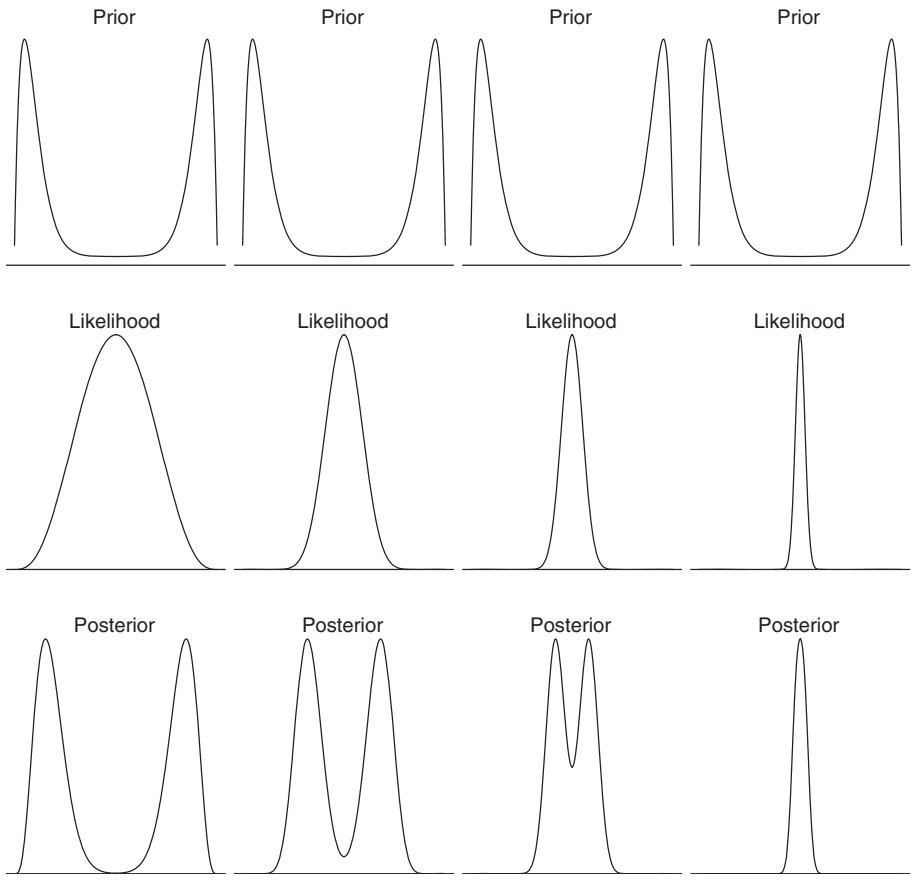

Figure 1.8 Sequence of posterior densities (1). The prior remains fixed across the sequence, as sample size increases and θ^* is held constant. In this example, $n = 6$, 30, 90, 450 across the four columns in the figure.

simplify the Bayesian computation of the posterior density and features of the posterior density, such as quantiles and highest posterior density estimates, especially when θ has many components.

1.8 Bayesian hypothesis testing

The posterior density of θ also provides the information necessary to test hypotheses about θ. At the outset, it is worth stressing that Bayesian hypothesis testing and frequentist hypothesis testing differ starkly. The most common hypothesis test of classical statistics, $H_0 : \theta = 0$, is untestable in the Bayesian approach if θ is a continuous parameter; to see this, note that if a continuous parameter $\theta \in \Omega \subseteq \mathbb{R}$ has the posterior distribution $p(\theta|\mathbf{y})$, then a 'point null' hypothesis such as $H_0 : \theta = c$ has zero probability, since c is a one-point set with measure zero (see Definition B.3). This difficulty also afflicts

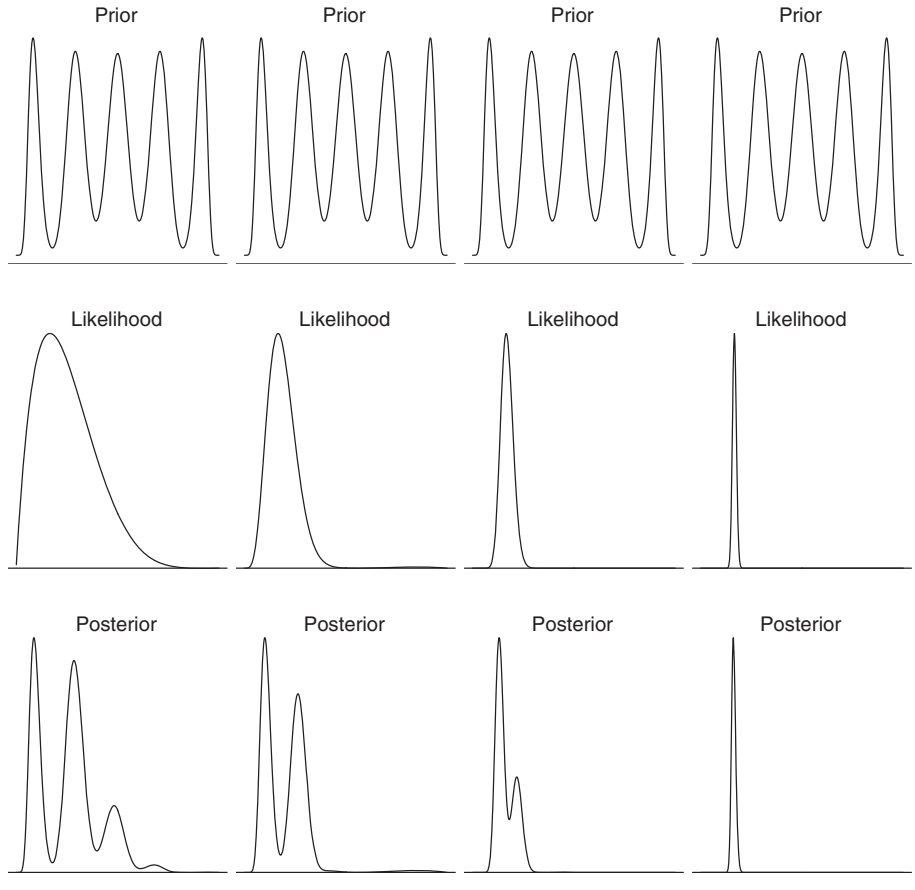

Figure 1.9 Sequence of posterior distributions (2). The prior remains fixed across the sequence, as sample size increases and θ^* is held constant. In this example, $n = 6$, 30, 150, 1500 across the four columns in the figure.

hypothesis testing in the frequentist world: with respect to a continuous parameter, *all point null hypotheses are false*, as the researcher would eventually discover if they were to successively test a point null hypothesis at a pre-specified, non-zero significance level, with increasing amounts of data (a fact that is typically ignored in introductory statistics classes). By concentrating attention on the posterior density, $p(\theta|\mathbf{y})$, the Bayesian approach helps to make clear the logical deficiencies of point null hypothesis testing. Thus, at least for continuous parameters, we don't test point null hypotheses in the Bayesian approach, and for that matter nor should a frequentist.

Instead, suppose we have a continuous parameter $\theta \in \mathbb{R}$, then two, exculsive, exhaustive and non-trivial (non-point) hypotheses are $H_0 : \theta < c$ and the alternative hypothesis $H_1 : \theta \geq c$. Posterior probabilities for these hypotheses are defined as follows:

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

and

$$\Pr(H_1|\mathbf{y}) = \Pr(\theta \geq c|\mathbf{y}) = \int_c^{\infty} p(\theta|\mathbf{y})d\theta.$$

For standard models, where conjugate priors have been deployed, these posterior 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 3.

The posterior probability of a hypothesis is something that only makes sense in a Bayesian framework. There is no such corresponding quantity in a frequentist framework, although this is how a frequentist p-value is often misinterpreted. For a frequentist, θ is a fixed but unknown number, and so hypotheses about θ are either true or false, and $\Pr(H_0|\mathbf{y}) = 1$ if H_0 is true, and zero if it is not. As such, for a frequentist, the falsity or truth of a hypothesis does not depend on the data, and so a quantity such as $\Pr(H_0|\mathbf{y})$ is meaningless. In contrast, for the Bayesian, θ is not fixed, but subject to (subjective prior/posterior) uncertainty, and so too is H_0, and so the posterior probability $\Pr(H_0|\mathbf{y})$ is quite useful. Indeed, one might argue that those types of posterior probability statement are exactly what one wants from a data analysis, letting us make statements of the sort 'how plausible is hypothesis H_0 in light of these data?' A frequentist p-value answers a different question: 'how frequently would I observe a result at least as extreme as the one obtained if H_0 were true?', which is a statement about the plausibility of the data *given* the hypothesis. Turning this assessment into an assessment about the hypothesis requires another step in the frequentist chain of reasoning (e.g. conclude H_0 is false if the p-value falls below some preset level). Contrast the Bayesian procedure, which lets us assess the plausibility of H_0 directly. A long line of papers contrasts p-values with Bayesian posterior probabilities, arguing (as I have here) that many analysts interpret the former as the latter, but that these two quantities can often be very different from one another; especially helpful papers on this score include Dickey (1977), Berger and Sellke (1987) and Berger (2003).

The following example provides a demonstration of Bayesian hypothesis testing using data from a survey. To help understand how Bayesian and frequentist approaches to hypothesis testing differ, a frequentist analysis is also provided.

■ Example 1.9

Attitudes towards abortion. Agresti and Finlay (1997, 133) report that in the 1994 General Social Survey, 1934 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 1934 respondents, 895 reported 'yes' and 1039 said 'no'. Let θ 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.

Frequentist approach. The survey estimate of θ is $\hat{\theta} = 895/1934 \approx .46$, 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. 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 it 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. The standard deviation of the normal sampling distribution (the standard error of $\hat{\theta}$) under H_0 is

$$\text{se}(\hat{\theta}_{H_0}) = \sqrt{\frac{\theta_{H_0}(1-\theta_{H_0})}{n}} = \sqrt{\frac{.50 \times (1-.50)}{1934}} \approx .011.$$

The realized value of $\hat{\theta}$ is $(.5 - .46)/.011 \approx 3.64$ standard errors away from the hypothesized value. Under a normal distribution, this an extremely rare event. Over repeated applications of random sampling, only a small proportion of estimates of θ will lie 3.64 or more standard errors away from the hypothesized mean of the sampling distribution. This proportion is

$$2 \times \int_{3.64}^{\infty} \phi(z)dz = 2 \times [1 - \Phi(3.64)] \approx .00028,$$

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 .00028 and for H_0 against H_B as .00014.

Bayesian approach. The unknown parameter is $\theta \in [0, 1]$ and suppose we bring little or no prior information to the analysis. In such a case, we know that the posterior density has the same shape as the likelihood, which with the large sample used here is well approximated by a normal density (the details of Bayesian estimation and inference for a sample proportion are presented in Chapter 2), specifically, 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 inferences about θ 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

$$\Pr(\theta > .5|y) = \int_{.5}^{\infty} p(\theta|y)d\theta = \int_{.5}^{\infty} \phi\left(\frac{\theta - .46}{.011}\right) d\theta = .00014.$$

That is, there is an apparent symmetry between the frequentist and Bayesian answers: in both instances, the 'answer' involved computing the same tail area probability of a normal distribution, with the 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 1.10. But this similarity really is only superficial. The Bayesian probability is a statement about the researcher's beliefs about θ, obtained via application

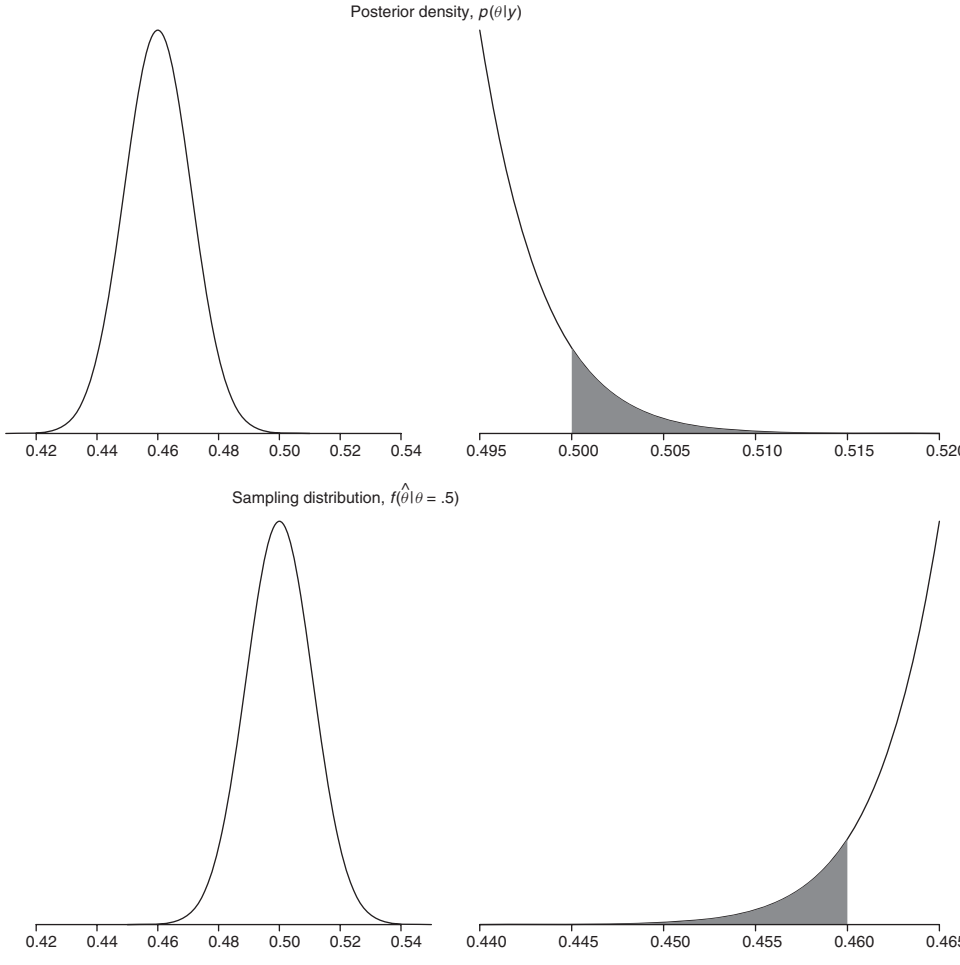

Figure 1.10 Posterior density contrasted with sampling distribution under $H_0 : \theta = .5$, for Example 1.9. The top right panel shows the posterior density in the neighborhood of $\theta = .5$, with the shaded region corresponding to the posterior probability $p(\theta > .5|\mathbf{y}) = \int_{.5}^{\infty} p(\theta|\mathbf{y})d\theta = .00014$. The lower right panel shows the sampling distribution in the neighborhood of $\hat{\theta} = .46$, with the shaded region corresponding to the proportion of times one would observe $\hat{\theta} \leq .46$ if $H_0 : \theta = .5$ were true, corresponding to .00014 of the area under the sampling distribution.

of Bayes Rule, and is $\Pr(H_0|y)$, obtained by computing the appropriate integral of the posterior distribution $p(\theta|y)$. The frequentist p-value 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_0 against the one-sided alternative, the proportion of $\hat{\theta} < .46$ we would see under repeated sampling, with the sampling distribution given by the null hypothesis.

1.8.1 Model choice

Applying Bayes Rule produces a posterior density, $f(\theta|\mathrm{y})$, 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 $\mathbf{y}$. 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 $\mathbf{y}$; 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, $\mathcal{M} = \{M_1, \ldots, M_J\}$; 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, $P(M_i|\mathbf{y})$, and to inform the choice of a particular model. This posterior probability comes via application of Bayes Rule for multiple discrete events, which we encounted earlier as Proposition 1.4. In the specific context of model choice, we have

$$P(M_i|\mathbf{y}) = \frac{P(M_i)\,p(\mathbf{y}|M_i)}{\sum_{j=1}^{J} P(M_j)\,p(\mathbf{y}|M_j)}. \tag{1.5}$$

The expression $p(\mathbf{y}|M_i)$ is the *marginal likelihood*, given by the identity

$$P(\mathbf{y}|M_i) = \int_{\Theta_i} p(\mathbf{y}|\theta_i, M_i)\,p(\theta_i)\,d\theta_i \tag{1.6}$$

i.e. averaging the likelihood for $\mathbf{y}$ under M_i over the prior for the parameters θ_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 $\mathcal{M}$ is the 'best model', M^*, and the researcher possesses the following simple loss function

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

For each model, $P([M_i \neq M^*]|\mathbf{y}) = 1 - P(M_i|\mathbf{y})$, and so the expected posterior loss of choosing model i is $1 - P(M_i|\mathbf{y})$. 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: $H_0 : .5 \leq \theta \leq 1$ and $H_1 : 0 \leq \theta < .5$, which generate priors

$$p_0(\theta_0) \equiv \text{Uniform}(.5, 1) = \begin{cases} 2 & \text{if } .5 \leq \theta_0 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_1(\theta_1) \equiv \text{Uniform}(0, .5) = \begin{cases} 2 & \text{if } 0 \leq \theta_1 < .5 \\ 0 & \text{otherwise} \end{cases}$$

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

$$\begin{aligned} p(y|H_0) &= \int_{.5}^{1} p(y|H_0, \theta_0) p_0(\theta_0) d\theta_0 = 2 \int_{.5}^{1} p(y|H_0, \theta_0) d\theta_0 \\ &= 2 \left(\Phi \left(\frac{1 - .46}{.011} \right) - \Phi \left(\frac{.5 - .46}{.011} \right) \right) = .00028 \end{aligned}$$

Under H_1, the marginal likelihood is

$$\begin{aligned} p(y|H_1) &= \int_{0}^{.5} p(y|H_1, \theta_1) p_1(\theta_1) d\theta_1 = 2 \int_{0}^{.5} p(y|H_1, \theta_1) d\theta_1 \\ &= 2 \left(\Phi \left(\frac{.5 - .46}{.011} \right) - \Phi \left(\frac{-.46}{.011} \right) \right) = 2. \end{aligned}$$

Thus, via Equation 1.5:

$$\begin{aligned} P(H_0|y) &= \frac{\frac{1}{2} \times .00028}{(\frac{1}{2} \times .00028) + (\frac{1}{2} \times 2)} = \frac{.00014}{.00014 + 1} = .00014 \\ P(H_1|y) &= \frac{1}{.00014 + 1} = .99986 \end{aligned}$$

indicating that H_1 is much more plausible than H_0.

1.8.2 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* $p(M_1|\mathbf{y})/p(M_0|\mathbf{y})$. The ratio of these two sets of odds is the Bayes factor:

Definition 1.11 (Bayes Factor). *Given data y and two models M_0 and M_1, the Bayes factor*

$$B_{10} = \frac{p(y|M_1)}{p(y|M_0)} = \left\{ \frac{p(M_1|y)}{p(M_0|y)} \right\} \Big/ \left\{ \frac{p(M_1)}{p(M_0)} \right\} \tag{1.7}$$

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 analagous 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 θ, 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:

B_{10}	2 log B_{10}	Evidence for M_1
<1	<0	negative (support M_0)
1 to 3	0 to 2	barely worth mentioning
3 to 12	2 to 5	positive
12 to 150	5 to 10	strong
>150	>10	very strong

Good (1988) summarizes the history of the Bayes factor, which long predates likelihood ratio 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

Ealier 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 the sense that the probability distribution of the data is written as a function of an unknown parameter θ (a scalar or vector). This approach to statisical 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 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 regarded as metaphysical nonsense. All that is real is the data at hand. We may also possess knowledge (or at least beliefs) about how the data were generated. For instance, are the data real at all, or are they output of a computer simulation? Were the data produced via an experiment with random assignment to treatment and control groups, by random sampling from a specific population, or are the data a 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 $\mathcal{S} = \{\{H, T\}_1 \times \ldots \times \{H, T\}_n\}$, 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 falling back on familar ways to simplify the problem. For example, reaching back to our introductory statistics classes, we would probably inquire 'are the coin flips 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 θ, given the n flips. In a Bayesian analysis we would 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 – using parameteric models to simplify data analysis – 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 only 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 1.12 (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 is finitely exchangeable.

Exchangeability is thus equivalent to the condition that the joint density of the data $\mathbf{y}$ 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, θ (but contrast Problem 1.8). Indeed, this is one of the critical implications of de Finetti's Representation Theorem.

As we shall now see, de Finetti's Thorem 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 Thoerem takes this implication a step further, showing that if I believe the data are infinitely exchangeable, then it is as if there is a parameter θ that drives a stochastic model generating the data, *and* a density over θ that doesn't depend on the data. This density is interpretable as a prior density, since it characterizes beliefs about θ that are not conditioned on the data. That is, the existence of a prior density 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 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 θ, 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 $\mathbf{y} = (y_1, \ldots, y_n)$. 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 θ, 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 θ 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 density over θ, $F(\theta)$, in the sense that it is a density over θ 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 θ, $p(\theta) = dF(\theta)/d\theta$. In this case, the identity in the proposition can be re-written as

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

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 θ, i.e. under independence conditional on θ,

$$\mathcal{L}(\theta; \mathbf{y}) \equiv f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

where

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

The second term in Equation 1.8, $p(\theta)$, is a prior density for θ, The integration in equation 1.8 is how we obtain the marginal density for $\mathbf{y}$, 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 θ and a prior over θ, 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 θ considered in Proposition 1.9 is recognizable as a success probability for independent 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 $\mathbf{y} = (y_1, \ldots, y_n)$, with possible values $s \in \{0, 1, \ldots, n\}$. Since there are $\binom{n}{s}$ 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 θ. Put differently, conditional on θ 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 Finneti's Representation Theorem exists, and here I re-state a relatively simple version of the general form, due to Smith (1984, 252):

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

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

where

$$\mu(F) = \lim_{n\to\infty} P(F_n)$$

and where F_n is the empirical distribution function for $\mathbf{y}$ 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 μ 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 $\boldsymbol{\theta}$; e.g., consider $\boldsymbol{\theta} = (\mu, \sigma^2)$, the mean and variance of a normal density, respectively. Note the *parametric*, *modeling* assumptions being made here. This said, the use of particular parameteric models is not completely *ad hoc*. There is much work outlining the conditions under which exchangeability plus particular *invariance* assumptions imply particular parameteric 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, and so on). Bernardo and Smith (1994, §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. 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 Finneti type representation for $P(\mathbf{y}_n)$ 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).

Still, n can be quite small in social science settings, and exchangeability judgements themselves may not be enough to justify parameteric modeling. In these cases models arise not so much as a consequence of having exchangeable data via de Finetti's Representation Theorem, but exist in the mind of the researcher prior to the analysis. This is perfectly fine, and indeed, corresponds to the way many social scientists go about their business: we look for data sets to tests theories and models, rather than (as may happen more often in statistics) we look for models to fit to the data we've been given to analyze. That is, one can (and ought!) to adopt a Bayesian approach even in the absence of exchangeable data; the point here is that models and prior densities are necessarily implied by accepting that one's data is exchangeable.

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, $\mathbf{y}$, 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 $\tilde{y}$, conditional on the observed sequence of r heads in n flips, $\mathbf{y}$. The probability density (or mass function) we form over this future outcome $\tilde{y}$ is known as the posterior predictive density (or mass function). In this case,

$$P(\tilde{y} = 1|\mathbf{y}) = \frac{P(\tilde{y} = 1, \mathbf{y})}{P(\mathbf{y})}$$

and, by exchangeability,

$$= \frac{\int_0^1 \theta^{r+\tilde{y}}(1-\theta)^{n+1-r-\tilde{y}} p(\theta)d\theta}{\int_0^1 \theta^r(1-\theta)^{n-r} p(\theta)d\theta}$$

$$= \frac{\int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}}\theta^r(1-\theta)^{n-r} p(\theta)d\theta}{\int_0^1 \theta^r(1-\theta)^{n-r} p(\theta)d\theta}$$

$$= \frac{\int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}}\mathcal{L}(\theta; \mathbf{y}) p(\theta)d\theta}{\int_0^1 \mathcal{L}(\theta; \mathbf{y}) p(\theta)d\theta},$$

since up to a constant multiplicative factor (that will cancel across numerator and denominator) $\mathcal{L}(\theta; \mathbf{y}) = \theta^r(1-\theta)^{n-r}$. But, by Bayes Rule (Proposition 1.5),

$$p(\theta|\mathbf{y}) = \frac{\mathcal{L}(\theta; \mathbf{y}) p(\theta)}{\int_0^1 \mathcal{L}(\theta; \mathbf{y}) p(\theta)d\theta}$$

and so $P(\tilde{y} = 1|\mathbf{y}) = \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} p(\theta|\mathbf{y})d\theta = \int_0^1 \theta p(\theta|\mathbf{y})d\theta = E(\theta|\mathbf{y})$. 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 θ. It is provocative to note that θ need not corresponds to anything in the physical world; indeed, the parameter θ may well be nothing more than a convenient fiction we conjure up to make a prediction problem tractable. The general point here is that we will 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

Once again, consider the simple case in Proposition 1.9, where $\mathbf{y} = (y_1, \ldots, y_n)$ 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 θ, and $p(\theta)$ is a prior density over θ.

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 come 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 $\mathbf{y} = (y_1, \ldots, y_n)'$ 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. Linley 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 $j = 1, \ldots, J$,

$$\begin{array}{rcll} y_i|C_i = j & \sim & \text{Bernoulli}(\theta_j) & \text{(likelihoods)} \\ \theta_j & \sim & p_j(\theta_j) & \text{(priors)} \end{array}$$

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 θ_j and the number of respondents in that country n_j:

$$\begin{array}{rcll} r_j|\theta_j, n_j & \sim & \text{Binomial}(\theta_j; n_j) & \text{(likelihood)} \\ \theta_j & \sim & p_j(\theta_j) & \text{(priors)} \end{array}$$

1.9.6 Exchangeability of parameters: hierarchical modeling

The hypothetical multi-country example just considered takes a step in the direction of 'hierarchical models'. That is, idea of exchangeability applies not just to data, but to parameters as well: i.e. note the deliberate use of the general term 'random quantities' rather than 'data' in Propositions 1.9 and 1.10).

Consider the example again. We know that data span J different countries. But that is all we know. Under these conditions, the θ_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 1.12). Put simply, the country labels j do not meaningfully distinguish the countries with respect to their corresponding θ_j. In this case, de Finetti's Representation Theorem implies that the joint density of the θ_j has the representation

$$p(\boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_J) = \int \prod_{j=1}^{m} p(\theta_j|\nu)p(\nu)d\nu \tag{1.9}$$

where ν 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 θ_j:

$$\begin{array}{rcll} \theta_j|\nu & \sim & p(\theta_j|\nu) & \text{(hierarchical model for } \theta_j\text{)} \\ \nu & \sim & p(\nu) & \text{(prior for hyperparameter } \nu\text{)} \end{array}$$

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

$$\begin{aligned} r_j|\theta_j, n_j &\sim \text{Binomial}(\theta_j; n_j) \\ \theta_j|\nu &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

$$\alpha \sim \text{Exponential}(2)$$
$$\beta \sim \text{Exponential}(2)$$

with $\nu = (\alpha, \beta)$ the hyperparameters for this problem. Details on the specific densities come later: e.g. in Chapter 2 we discuss models for proportions in some detail, and a hierarchical model for binomial data is considered in Example 7.9. 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. Information that survey respondents were located in different countries prompted us to revise a belief of exchangeability for them; similarly, information allowing us to distinguish countries from one another might lead us to revisit the exchangeability judgement over the θ_j parameters. In particular, suppose we have variables at the country level, $\mathbf{x}_j$, 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 $\mathbf{x}_j$ might be the following *multi-level* hierarchical model:

$$\begin{aligned} r_j|\theta_j, n_j &\sim \text{Binomial}(\theta_j; n_j) \\ z_j &= \log\left(\frac{\theta_j}{1-\theta_j}\right) \\ z_j|\mathbf{x}_j &\sim N(\mathbf{x}_j\boldsymbol{\beta}, \sigma^2) \\ \boldsymbol{\beta}|\sigma^2 &\sim N(\mathbf{b}, \sigma^2\mathbf{B}) \\ \sigma^2 &\sim \text{Inverse-Gamma}\left(\frac{\nu}{2}, \frac{\nu_0\sigma_0^2}{2}\right). \end{aligned}$$

Again, details on the specific models and densities deployed here come in later chapters; Example 7.10 provides a detailed consideration of a multi-level model. The key idea is that the information in $\mathbf{x}_j$ enters as the independent variables in a regression model for z_j, the log-odds of each country's θ_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 7.

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 his friend, Richard Price (Bayes 1763). This famous essay has been republished many times since (e.g. Bayes 1958).

Several authors have noted that there is some doubt that Bayes actually discovered the theorem named for him; see, for instance, Stigler (1999, Ch 14) and the references

in Fienberg (2006). Nor is it clear that Bayes himself was a 'Bayesian' in the sense that we use the term today (e.g. Stigler 1982).

The subject of Bayes *Essay towards solving a problem in the doctrine of chances* was what we would today recognize as a binomial problem: given x successes in n independent binary trials, what should we infer about π, the underlying probability of success? Bayes himself studied the binomial problem with a uniform prior. In 1774 Laplace (apparently unaware of Bayes work) stated Bayes theorem in its more general form, and also considered non-uniform priors (Laplace 1774). Laplace's article popularized what would later become known as 'Bayesian' statistics. Perhaps because of Laplace's work on the subject, Bayes' essay itself '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, ch. 1), and Stigler (1986a, ch. 3). We return to the relatively simple statistical problem considered by Bayes (drawing inferences given binomial data) in Chapter 2.

The adjective 'Bayesian' did not enter the statistical vernacular until the 20th century. Fienberg (2006) reviews the 'neo-Bayesian revival' of the 20th century, and, via a review by Edwards (2004), traces the first use of 'Bayesian' as an adjective to Fisher (1950), in an introduction to a paper originally written in 1921. Unsurprisingly, Fisher's use of the term was not flattering, since he was at pains to contrast his approach to statistical inference from the subjectivism he disliked in the Bayesian approach. In contrast with Fisher's pejorative use of the term, Fienberg (2006) provides a detailed exposition of how Bayesians themselves came to adopt the 'Bayesian' moniker in the 20th century.

Problems

1.1 Consider a cross-national study of economic development, where the data comprise all OECD countries in 2000. A researcher argues that while these data are the population of OECD countries in 2000, they are nonetheless a random sample from the histories of these countries. Discuss.

1.2 Consider the drug testing problem given in Example 1.1. Consider the false negative rate and the false positive rate of the drug test as two variables.

1. Construct a grid of hypothetical values for these two variables. At each point on the grid, compute the posterior probability of H_U, the hypothesis 'the subject uses the prohibited substance' given the prior on this hypothesis of $P(H_u) = .03$ and a postitive test result. Use a graphical technique such as a contour plot or an image plot to summarize the results.

2. What values for the two error rates of the test give rise to a posterior probability on H_U that exceeds 0.5?

3. Repeat this exercise, but now considering a run of 3 positive tests: what values of the test error rates give rise to a posterior probability for H_U in excess of 0.95?

1.3 Suppose $p(\theta) \equiv \chi^2_2$. Compute a 50 % highest density region for θ. Compare this region with the inter-quartile range of $p(\theta)$.

1.4 Consider a density $p(\theta)$. Under what conditions can a HDR for θ of content α be determined by simply noting the $(1-\alpha)/2$ and $1-(1-\alpha)/2$ quantiles of $p(\theta)$? That is, what must be true about $p(\theta)$ so as to let us compute a HDR this way?

1.5 Consider Example 1.4. Repeat the analysis in the example assuming that it is (a) two times and (b) five times as costly to overestimate applicant ability than it is to underestimate ability.

1.6 A poll of 500 adults in the United States taken in the Spring of 2008 finds that just 29 % of respondents approve of the way that George W. Bush is handling his job as president.

1. Report the posterior probabilities of $H_0 : \theta > .33$ and $H_1 : \theta < .33$. The threshold $\theta = .33$ has some politically interest, say, if we assume that (up to a rough approximation) the electorate is evenly partitioned into Democrat, Independent, and Republican identifiers.
2. Report a Bayes factor for H_0 vs H_1. Comment briefly on your finding.
3. Contrast how a frequentist approach would distinguish between these two hypotheses.

1.7 Consider the poll data in the previous question. Suppose you had the following uniform prior for θ, $\theta \sim \text{Unif}(.4, .6)$. What is your posterior density, given the polling data?

1.8 Is exchangeability merely a Bayesian way of saying 'iid'? That is, establish whether statistical independence is a necessary and sufficient condition for exchangeability. In particular, can you come up with an example where exchangeability holds, but independence does not?