8

Introduction to Bayesian Estimation

The rest of this book deals with Bayesian estimation. This chapter uses examples to illustrate the fundamental concepts of Bayesian point and interval estimation. It also provides an introduction to Chapters 9 and 10 where more advanced examples require computationally intensive methods.

Bayesian and frequentist statistical inference take very different approaches to statistical decision making.

- The frequentist view of probability, and thus of statistical inference, is based on the idea of an experiment that can be repeated many times.
- The Bayesian view of probability and of inference is based on a personal assessment of probability and on observations from a single performance of an experiment.

These different views lead to fundamentally different procedures of estimation, and the interpretations of the resulting estimates are also fundamentally different. In practical applications, both ways of thinking have advantages and disadvantages, some of which we will explore here.

Statistics is a relatively young science. For example, interval estimation has gradually become common in scientific research and business decision making only within the past 75 years. On this time scale it seems strange to talk about "traditional" approaches. However, frequentist viewpoints are currently much better established, particularly in scientific research, than Bayesian ones. Recently, the use of Bayesian methods has been increasing, partly because the Bayesian approach seems to be able to get more useful solutions than frequentist ones in some applications and partly because improvements in computation have made Bayesian methods increasingly convenient to apply in practice. The Gibbs sampler is one computationally intensive method that is broadly applicable in Bayesian estimation.

For some of the very simple examples considered here, Bayesian and frequentist methods give similar results. But that is not the main point. We hope you will gain some appreciation that Bayesian methods are sometimes the most natural and useful ones in practice. Also, we hope you will begin to appreciate the essential role of computation in Bayesian estimation.

For most people, the starkest contrast between frequentist and Bayesian approaches to analyzing an experiment or study is that Bayesian inference provides the opportunity—even imposes the requirement—to take explicit notice of "information" that is available before any data are collected. That is where we begin.

8.1 Prior Distributions

The Bayesian approach to statistical inference treats population parameters as random variables (not as fixed, unknown constants). The distributions of these parameters are called **prior distributions**. Often both expert knowledge and mathematical convenience play a role in selecting a particular type of prior distribution. This is easiest to explain and to understand in terms of examples. Here we introduce four examples that we carry throughout this chapter.

Example 8.1. Election polling. Suppose Proposition A is on the ballot for an upcoming statewide election, and a political consultant has been hired to help manage the campaign for its adoption. The proportion π of prospective voters who currently favor Proposition A is the population parameter of interest here. Based on her knowledge of the politics of the state, the consultant's judgment is that the proposition is almost sure to pass, but not by a large margin. She believes that the most likely proportion π of voters in favor is 55% and that the proportion is not likely to be below 51% or above 59%.

It is reasonable to try to use the beta family of distributions to model the expert's opinion of the proportion in favor because distributions in this family take values in the interval $(0, 1)$, as do proportions. Beta distributions have density functions of the form

$$\begin{aligned} p(\pi) &= K\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1}, \end{aligned}$$

for $0 < \pi < 1$, where α, $\beta > 0$ and K is the constant such that $\int_0^1 p(\pi)\, d\pi = 1$. Here we adopt two conventions that are common in Bayesian discussions: the use of the letter p instead of f to denote a density function, and the use of the symbol $\propto$ (read "proportional to") instead of $=$ so that we can avoid specifying a constant whose exact value is unimportant to the discussion. The essential factor of the density function that remains when the constant is suppressed is called the **kernel** of the density function (or of its distribution).

A member of the beta family that corresponds reasonably well to the expert's opinion has $\alpha_0 = 330$ and $\beta_0 = 270$. (Its density is the fine-line curve in Figure 8.1.) This is a reasonable choice of parameters for several reasons.

- This beta distribution is centered near 55% by any of the common measures of centrality. By analytic methods one can show that the *mean* of this distribution is $\alpha_0/(\alpha_0+\beta_0) = 330/600 = 55.00\%$ and that its *mode* is $(\alpha_0-1)/(\alpha_0+\beta_0-2) = 329/598 = 55.02\%$. Computational methods show the *median* to be 55.01%. (The R function `qbeta(.5, 330, 270)` returns `0.5500556`.) The mean is the most commonly used measure of centrality. Here the mean, median, and mode are so nearly the same that it doesn't make any practical difference which is used.

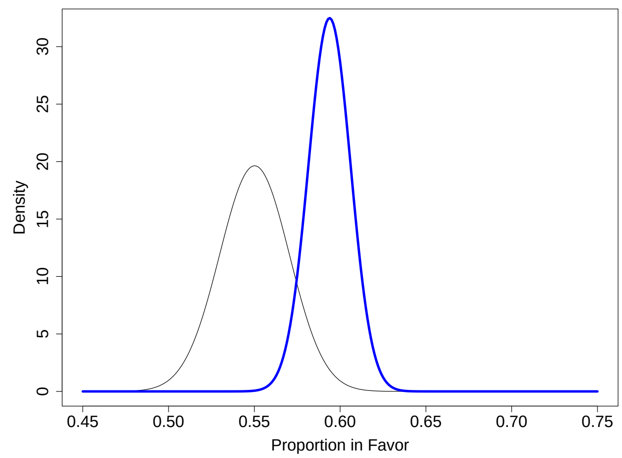

Figure 8.1. Prior and posterior densities for the proportion π of the population in favor of ballot Proposition A (see Examples 8.1 and 8.5). The prior (fine line) is $\mathsf{BETA}(330, 270)$ with mean 55.0%. Based on a poll of 1000 subjects with 62.0% in favor, the more concentrated posterior (heavy) is $\mathsf{BETA}(950, 650)$ with mean 59.5%.

- Numerical integration shows that these parameters match the expert's prior probability interval fairly well: $P\{0.51 < \pi < 0.59\} \approx 0.95$. (The R code `pbeta(.59, 330, 270) - pbeta(.51, 330, 270)` returns 0.9513758.)

Of course, slightly different choices for α_0 and β_0 would match the expert's opinion about as well. It is not necessary to be any fussier in choosing the parameters than the expert was in specifying her hunches. Also, distributional shapes other than the beta might match the expert's opinion just as well. But we choose a member of the beta family because it makes the mathematics relatively easy in what comes later and because we have no reason to believe that the shape of our beta distribution is inappropriate here. (See Problems 8.2 and 8.3.)

If the consultant's judgments about the political situation are correct, then they may be helpful in managing the campaign. If she too often brings bad judgment to her clients, her reputation will suffer and she will be out of the political consulting business before long. Fortunately, as we see in the next section, the details of her judgments become less important if we also have some polling data to rely upon. $\diamondsuit$

Example 8.2. Counting mice. An island in the middle of a river is one of the last known habitats of an endangered kind of mouse. The mice rove about the island in ways that are not fully understood and so are taken as random.

Ecologists are interested in the average number of mice to be found in particular regions of the island. To do the counting in a region they set many traps there at night, using bait that is irresistible to mice at close range. In the morning they count and release the mice caught. It seems reasonable to suppose that almost all of the mice in the region where traps were set during the previous night were caught and that the number of them on any one night has a Poisson distribution. The purpose of the trapping is to estimate the mean λ of this distribution.

Even before the trapping is done the ecologists doing this study have some information about λ. For example, although the mice are quite shy, there have been occasional sightings of them in almost all regions of the island, so it seems likely that $\lambda > 1$. On the other hand, from what is known of the habits of the mice and the food supply in the regions, it seems unlikely that there would be as many as 25 of them in any one region at a given time.

In these circumstances, it seems reasonable to use a gamma distribution as a prior distribution for λ. This gamma distribution has the density

$$p(\lambda) \propto \lambda^{\alpha-1} e^{-\kappa\lambda},$$

for $\lambda > 0$, where the shape parameter α and the rate parameter κ must both be positive. First, we choose a gamma distribution because it puts all of its probability on the positive half line, and λ must surely have a positive value. Second, we choose a member of the gamma family because it simplifies some important computations that we need to do later.

Using straightforward calculus, one can show that a distribution in the gamma family has mean α/κ, mode $(\alpha - 1)/\kappa$, and variance α/κ^2. These distributions are right-skewed, with the skewness decreasing as α increases.

One reasonable choice for a prior distribution on λ is a gamma distribution with $\alpha_0 = 4$ and $\kappa_0 = 1/3$. Reflecting the skewness, the mean 12, median 11.02, and mode 9 are noticeably different. (We obtained the median using R: `qgamma(.5, 4, 1/3)` returns `11.01618`. Also, see Problem 8.8.) Numerical methods also show that $P\{\lambda < 25\} = 0.97$. (In R, `pgamma(25, 4, 1/3)` returns `0.9662266`.) All of these values are consistent with the expert opinions of the ecologists.

It is clear that the experience of the ecologists with the island and its endangered mice will influence the course of this investigation in many ways: dividing the island into meaningful regions, modelling the randomness of mouse movement as Poisson, deciding how many traps to use and where to place them, choosing a kind of bait that will attract mice from a region of interest but not from all over the island, and so on. The expression of some of their background knowledge as a prior distribution is perhaps a relatively small use of their expertise. But a prior distribution is a necessary starting place for Bayesian inference, and it is perhaps the only aspect of expert opinion that will be explicitly tempered by the data that are collected. ◇

Example 8.3. Weighing an object. A construction company buys reinforced concrete beams with a nominal weight of 700 lb. Experience with a particular supplier of these beams has shown that their beams very seldom weigh less than 680 or more than 720 lb. In these circumstances it may be convenient and reasonable to use $\mathsf{NORM}(700, 10)$ as the prior distribution of the weight of a randomly chosen beam from this supplier.

Usually, the exact weight of a beam is not especially important, but there are some situations in which it is crucial to know the weight of a beam more precisely. Then a particular beam is selected and weighed several times on a scale in order to determine its weight more exactly.

Theoretically, a frequentist statistician would ignore "prior" or background experience in doing statistical inference, basing statistical decisions only on the data collected when a beam is weighed. In real life it is not so simple. For example, the design of the weighing experiment will very likely take past experience into account

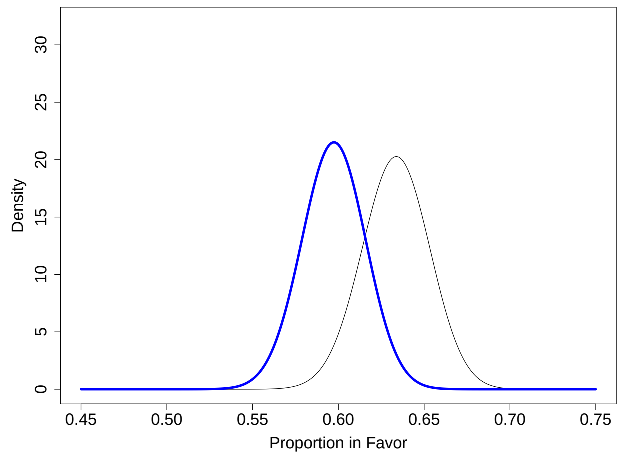

Figure 8.2. Prior and posterior densities for the population proportion π favoring Proposition B (see Problem 8.1). Here the prior (fine line) reflects strong optimism that the proposition is leading. The posterior (heavy line), taking into account results of a relatively small poll with 62% *opposed*, does little to dampen the optimism.

in one way or another. (For example, if you are going to weigh things, then you need to know whether you will be be using a laboratory balance, a truck scale, or some intermediate kind of scale. And if you need more precision than the scale will give in a single measurement, you may need to weigh each object several times and take the average.) For the Bayesian statistician the explicit codification of some kinds of background information into a prior distribution is a required first step. ♢

Example 8.4. Precision of hemoglobin measurements. A hospital has just purchased a device for the assay of hemoglobin (Hgb) in the blood of newborn babies (in g/dl). Considering the claims of the manufacturer and experience with competing methods of measuring Hgb, it seems reasonable to suppose the machine gives unbiased normally distributed results X with a standard deviation σ somewhere between 0.25g/dl and 1g/dl.

For mathematical convenience in Bayesian inference, it is customary to express prior distributions for the variability of a normal distribution in terms of a gamma distribution on the precision $\tau = 1/\sigma^2$. In our example, we might seek a prior distribution on τ with $P\{1/4 < \sigma < 1\} = P\{1/16 < \sigma^2 < 1\} = P\{1 < \tau < 16\} \approx 0.95$. One reasonable choice, under which this interval has probability 0.96, is $\tau \sim \mathsf{GAMMA}(\alpha_0 = 3, \kappa_0 = 0.75)$.

When τ has a gamma prior $\mathsf{GAMMA}(\alpha, \kappa)$, we say that $\theta = 1/\tau = \sigma^2$ has an an **inverse gamma** prior distribution $\mathsf{IG}(\alpha, \kappa)$. This distribution family has density

$$p(\theta) = \frac{\kappa^\alpha}{\Gamma(\alpha)}\, \theta^{-(\alpha+1)}\, e^{-\kappa/\theta} \propto \theta^{-(\alpha+1)}\, e^{-\kappa/\theta},$$

for $\theta > 0$. The mode of this distribution is $\kappa/(\alpha+1)$ and, when $\alpha > 1$, its mean is $\kappa/(\alpha-1)$.

In R, simulated values and quantiles of IG can be found as reciprocals of `rgamma` and `qgamma`, respectively. Cumulative probabilities can be found by using reciprocal arguments in `pgamma`. For example, with $\alpha_0 = 3$, $\kappa_0 = .75$, we find $\mathrm{Med}(\theta) = 1/\mathrm{Med}(\tau) = 0.28$ with the code `1/qgamma(.5, 3, .75)`, and we get 0.50 from `pgamma(1/0.28047, 3, .75)`. ♢

8.2 Data and Posterior Distributions

The second step in Bayesian inference is to collect data and to combine the information in the data with the expert opinion represented by the prior distribution. The result is a posterior distribution that can be used for inference.

Once the data are available, we can use Bayes' Theorem to compute the posterior distribution $\pi|x$. Equation (5.6), repeated here as (8.1), states an elementary version of Bayes' Theorem for an observed event E and a partition $\{A_1, A_2, \ldots, A_k\}$ of the sample space S.

$$P(A_j|E) = \frac{P(A_j)P(E|A_j)}{\sum_{i=1}^{k} P(A_i)P(E|A_i)}. \tag{8.1}$$

This equation expresses a posterior probability $P(A_j|E)$ in terms of the prior probabilities $P(A_i)$ and the conditional probabilities $P(E|A_i)$.

Here we use a more general version of Bayes' Theorem involving data x and a parameter π:

$$p(\pi|x) = \frac{p(\pi)p(x|\pi)}{\int p(\pi)p(x|\pi)\,d\pi} \propto p(\pi)p(x|\pi), \tag{8.2}$$

where the integral is taken over all values of π for which the integrand is possible. The proportionality symbol $\propto$ is appropriate because the integral is a constant. (In case the distribution of π is discrete, the integral is interpreted as a sum.)

Thus the posterior distribution of $\pi|x$ is found from the prior distribution of π and the distribution of the data x given π. If π is a known constant, $p(x|\pi)$ is the density function of x; we might integrate it with respect to x to evaluate the probability $P(x \in A) = \int_A p(x)\,dx$. However, when we use (8.2) to find a posterior, we know the data x, and we view $p(x|\pi)$ as a function of π. When viewed in this way, $p(x|\pi)$ is called the **likelihood function** of π. (Technically, the likelihood function is defined only up to a positive constant.)

A convenient summary of of our procedure for finding the posterior distribution with relationship (8.2) is to say

$$\text{POSTERIOR} \propto \text{PRIOR} \times \text{LIKELIHOOD}.$$

We now illustrate this procedure for each of the examples of the previous section.

Example 8.5. Election Polling (continued). Suppose n randomly selected registered voters express opinions on Proposition A. What is the likelihood function, and how do we use it to find the posterior distribution?

If the value of π were known, then the number x of the respondents in favor of Proposition A is a random variable with the binomial distribution: $\binom{n}{x}\pi^x(1-\pi)^{n-x}$,

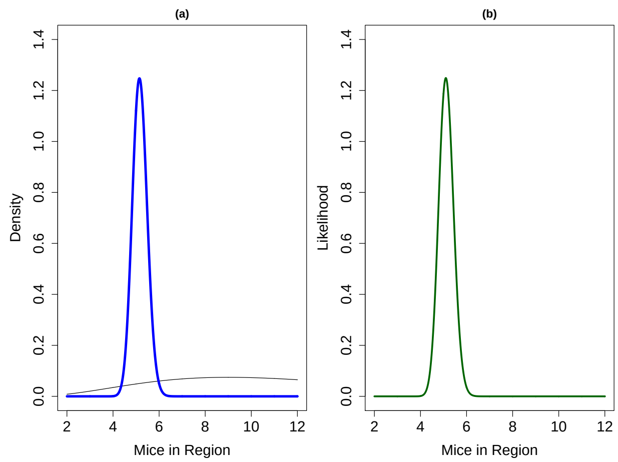

Figure 8.3. (a) Prior and posterior densities for the number of mice in a region. (b) Likelihood function of the mouse data: 50 nights with a total of 256 mice trapped. Because the prior density (fine line) is relatively flat, the data largely determine the mode 5.166 of the posterior (heavy). The MLE $\hat{\lambda} = 256/50 = 5.120$ (mode of the likelihood) is not far from the posterior mode. (See Examples 8.2 and 8.6.)

for $x = 0, 1, 2, \ldots, n$. Now that we have data x, the likelihood function of π becomes $p(x|\pi) \propto \pi^x (1-\pi)^{n-x}$.

Furthermore, display (8.2) shows how to find the posterior distribution

$$\begin{aligned} p(\pi|x) &\propto \pi^{\alpha_0 - 1}(1-\pi)^{\beta_0 - 1} \times \pi^x (1-\pi)^{n-x} \\ &= \pi^{\alpha_0 + x - 1}(1-\pi)^{\beta_0 + n - x - 1} = \pi^{\alpha_n - 1}(1-\pi)^{\beta_n - 1}, \end{aligned}$$

where we recognize the last line as the kernel of a beta distribution with parameters $\alpha_n = \alpha_0 + x$ and $\beta_n = \beta_0 + n - x$. It is easy to find the posterior in this case because the (beta) prior distribution we selected has a functional form that is similar to that of the (binomial) distribution of the data, yielding a (beta) posterior. In this case we say that the beta is a **conjugate prior** for binomial data. (When nonconjugate priors are used, special computational methods are often necessary; see Problems 8.5 and 8.6.)

Recall that the parameters of the prior beta distribution are $\alpha_0 = 330$ and $\beta_0 = 270$. If $x = 620$ of the $n = 1000$ respondents favor Proposition A, then the posterior has a beta distribution with parameters $\alpha_n = \alpha_0 + x = 950$ and $\beta_n = \beta_0 + n - x = 650$. Look at Figure 8.1 for a visual comparison of the prior and posterior distributions. The density curves were plotted with the following R script. (By using `lines` we can plot the prior curve on the same axes as the posterior.)

```
x = seq(.45, .7, .001)
prior = dbeta(x, 330, 270)
post = dbeta(x, 950, 650)
plot(x, post, type="l", ylim=c(0, 35), lwd=2,
     xlab="Proportion in Favor", ylab="Density")
lines(x, prior)
```

The posterior mean is $950/(950+650) = 59.4\%$, a Bayesian point estimate of the actual proportion of the population currently in favor of Proposition A. Also, according to the posterior distribution, $P\{0.570 < \pi < 0.618\} = 0.95$, so that a 95% **posterior probability interval** for the proportion in favor is (57.0%, 61.8%). (In R, `qbeta(.025, 950, 650)` returns `0.5695848`, and `qbeta(.975, 950, 650)` returns `0.6176932`.)

This probability interval resulting from Bayesian estimation is a straightforward probability statement. Based on the combined information from her prior distribution and from the polling data, the political consultant now believes it is very likely that between 57% and 62% of the population currently favors Proposition A. In contrast to a frequentist "confidence" interval, the consultant can use the probability interval without the need to view the poll as a repeatable experiment. ◇

Example 8.6. Counting mice (continued). Suppose that a region of the island is selected where the gamma distribution with parameters $\alpha_0 = 4$ and $\kappa_0 = 1/3$ is a reasonable prior for λ. The prior density is $p(\lambda) \propto \lambda^{\alpha_0 - 1} e^{-\kappa_0 \lambda}$.

Over a period of about a year, traps are set out on $n = 50$ nights with the total number of captures $t = \sum_{i=1}^{50} x_i = 256$ for an average of 5.12 mice captured per night. Thus the Poisson likelihood function of the data is

$$p(\mathbf{x}|\lambda) \propto \prod_{i=1}^{n} \lambda^{x_i} e^{-\lambda} = \lambda^t e^{-n\lambda},$$

and the posterior distribution is

$$\begin{aligned} p(\lambda|\mathbf{x}) &\propto \lambda^{\alpha_0 - 1} e^{-\kappa_0 \lambda} \times \lambda^t e^{-n\lambda} \\ &= \lambda^{\alpha_0 + t - 1} e^{-(\kappa_0 + n)\lambda}, \end{aligned}$$

in which we recognize the kernel of the gamma distribution with parameters $\alpha_n = \alpha_0 + t$ and $\kappa_n = \kappa_0 + n$. Thus the posterior mean for our particular prior and data is

$$\frac{\alpha_n}{\kappa_n} = \frac{\alpha_0 + t}{\kappa_0 + n} = \frac{4 + 256}{1/3 + 50} = \frac{260}{50.33} = 5.166,$$

the posterior mode is $(\alpha_n - 1)/\kappa_n = 259/50.33 = 5.146$, and the posterior median is 5.159. Based on this posterior distribution, a 95% probability interval for λ is $(4.56, 5.81)$. (In R, `qgamma(.025, 260, 50.33)` returns `4.557005`, and `qgamma(.975, 260, 50.33)` returns `5.812432`.) The prior and posterior densities are shown in Figure 8.3. ◇

Example 8.7. Weighing a beam (continued). Suppose that a particular beam is selected from among the beams available. Recall that, according to our prior distribution, the weights of beams in this population is $\mathsf{NORM}(700, 10)$, so $\mu_0 = 700$ pounds and $\sigma_0 = 10$ pounds. The beam is weighed $n = 5$ times on a balance that gives unbiased, normally distributed readings with a standard deviation of $\sigma = 1$ pound. Denote the data by $\mathbf{x} = (x_1, \ldots, x_n)$, where the x_i are independent $\mathsf{NORM}(\mu, \sigma)$, and μ is the parameter to be estimated. Such data have the likelihood function

$$p(\mathbf{x}|\mu) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right],$$

where the distribution of μ is determined by the prior, and $\sigma = 1$ is known. Then after some algebra (see Problem 8.12), the posterior is seen to be

$$p(\mu|\mathbf{x}) \propto p(\mu)p(\mathbf{x}|\mu) \propto \exp[-(\mu - \mu_n)^2/2\sigma_n^2],$$

which is the kernel of $\mathsf{NORM}(\mu_n, \sigma_n)$, where

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}.$$

It is common to use the term **precision** to refer to the reciprocal of a variance. If we define $\tau_0 = 1/\sigma_0^2$, $\tau = 1/\sigma^2$, and $\tau_n = 1/\sigma_n^2$, then we have

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{x} \quad \text{and} \quad \tau_n = \tau_0 + n\tau.$$

Thus, we say that the posterior precision is the sum of the precisions of the prior and the data, and that the posterior mean is a precision-weighted average of the means of the prior and the data.

In our example, $\tau_0 = 0.01$, $\tau = 1$, and $\tau_n = 5.01$. Thus the weights are $0.01/5.01 \approx 0.002$ for the prior mean μ_0 and $5/5.01 \approx 0.998$ for the mean $\bar{x}$ of the data. We see that the posterior precision is almost entirely due to the precision of the data, and the value of the posterior mean is almost entirely due to the mean of the sample. In this case, the sample of five relatively high-precision observations is enough to concentrate the posterior and diminish the impact of the prior. (See Problem 8.11 and Figure 8.4 for the computation of the posterior mean and a posterior probability interval.) ◇

Example 8.8. Precision of hemoglobin measurements (continued). Suppose researchers use the new device to make Hgb determinations v_i on blood samples from $n = 42$ randomly chosen newborns, and also make extremely precise corresponding laboratory determinations w_i on the same samples. Based in part on assumptions in Example 8.4, we assume $x_i = v_i - w_i \sim \mathsf{NORM}(0, \sigma)$. Assuming the laboratory measurements to be of "gold standard" quality, we ignore their errors and take $\tau = 1/\sigma^2$ to be a useful measure of the precision of the new device. If we observe $s = \sqrt{\sum_i x_i^2/n} = 0.34$ and use the prior distribution $\tau \sim \mathsf{GAMMA}(3, 0.75)$ of Example 8.4, then what posterior probability intervals can we give for τ and for σ?

The likelihood function of the data $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$p(\mathbf{x}|\theta) \propto \prod_{i=1}^{n} \theta^{-1/2} \exp\left(-\frac{x_i^2}{2\theta}\right) = \theta^{-n/2} \exp\left(-\frac{ns^2}{2\theta}\right),$$

where we denote $\sigma^2 = \theta$, and the posterior distribution of θ is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \theta^{-(\alpha_0+1)} \exp\left(-\frac{\kappa_0}{\theta}\right) \times \theta^{-n/2} \exp\left(-\frac{ns^2}{2\theta}\right) \\ &= \theta^{-(\alpha_n+1)} \exp\left(-\frac{\kappa_n}{\theta}\right), \end{aligned}$$

where $\alpha_n = \alpha_0 + n/2$ and $\kappa_n = \kappa_0 + ns^2/2$. We recognize this as the kernel of the $\mathsf{IG}(\alpha_n, \kappa_n)$ density function. Notice that the posterior has a relatively simple

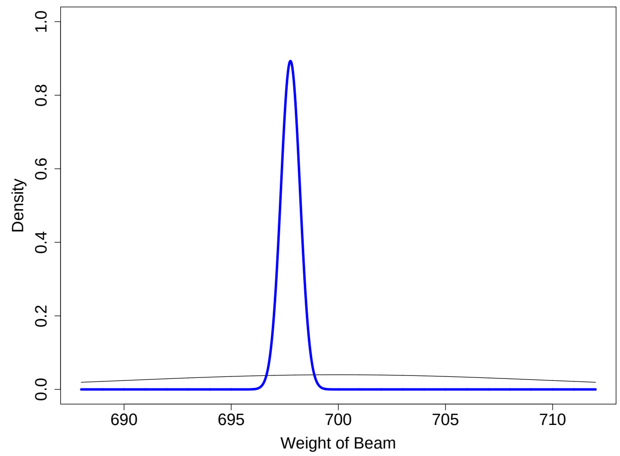

Figure 8.4. Prior density and posterior density for the weight of a beam. The normal prior (fine line) is so flat that the normal posterior (heavy) is overwhelmingly influenced by the data, obtained by repeated weighing of the beam on a scale of relatively high precision. (See Examples 8.3 and 8.7, and Problem 8.11.)

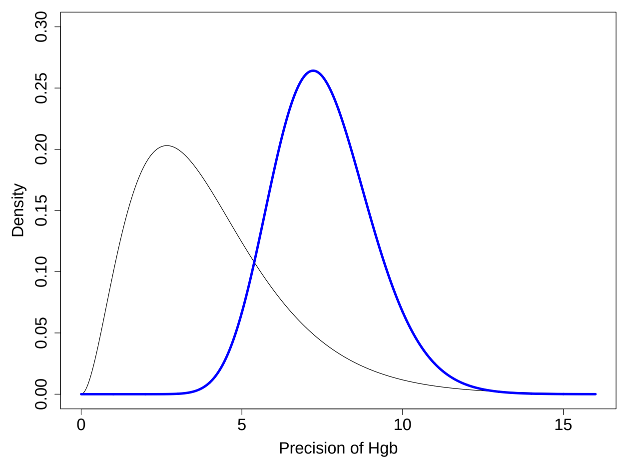

Figure 8.5. Prior density and posterior density for the precision of hemoglobin measurements. The gamma prior (fine line) contributes information corresponding to six measurements. The posterior (heavy) combines this information with data on 42 subjects to give greater precision. (See Examples 8.4 and 8.8.)

form because θ appears in the denominator of the exponential factor of the inverse-gamma prior. If we had used a gamma prior for the variance θ (instead of the precision $\tau = 1/\theta$), then θ would have appeared in the numerator of the exponential factor, making the posterior density unwieldy.

For our data $\alpha_n = 3 + 42/2 = 24$ and $\kappa_n = 0.75 + 42(0.34)^2/2 = 3.178$, so that a 95% posterior probability interval for τ is $(4.84, 10.86)$, computed in R as `qgamma(c(.025, .975), 24, 3.18)`. The corresponding interval for σ is $(0.303, 0.455)$. The frequentist 95% confidence interval for $\sigma = \sqrt{\theta}$ is based on $ns^2/\theta \sim \mathsf{CHISQ}(n)$ is $(0.280, 0.432)$, and can be computed in R as `sqrt(42*(.34)^2/qchisq(c(.975,.025), 42))`. The gamma prior and posterior distributions for the precision τ are shown in Figure 8.5 for τ in the interval $(1, 16)$.

Notes: (1) Because the normal mean is assumed known, $\mu = 0$, we have $ns^2/\sigma^2 = \sum(x_i - \mu)^2/\sigma^2 = \sum x_i^2/\sigma^2$ distributed as chi-squared with n (not $n-1$) degrees of freedom. (2) This example is loosely based on a real situation reported in [HF94] and used as an extended example in Unit 14 of [Tru02]. In this study, $s = 0.34$ based on $n = 42$ subjects. Complications in practice are that readings from the new device appear to be slightly biased and that the laboratory determinations, while more precise than those from the new device, are hardly free of measurement error. Fortunately, in this clinical setting the precision of both kinds of measurements is much better than it needs to be. ◇

In the next two chapters we look at Bayesian estimation problems where computationally intensive methods are required to find posterior distributions. Specifically, ideas of continuous Markov Chains from Chapter 7 are used to implement Gibbs Samplers.

8.3 Problems

Problems Related to Examples 8.1 and 8.5 (Binomial Data)

8.1 In a situation similar to Example 8.1, suppose a political consultant chooses the prior $\mathsf{BETA}(380, 220)$ to reflect his assessment of the proportion of the electorate favoring Proposition B.

a) In terms of a most likely value for π and a 95% probability interval for π, describe this consultant's view of the prospects for Proposition B.
b) If a poll of 100 randomly chosen registered voters shows 62% *opposed* to Proposition B, do you think the consultant (a believer in Bayesian inference) now fears Proposition B will fail? Quantify your answer with specific information about the posterior distribution. Recall that in Example 8.5 a poll of 1000 subjects showed 62% in favor of Proposition A. Contrast that situation with the current one.
c) Modify the R code of Example 8.5 to make a version of Figure 8.2 (p191) that describes this problem.
d) Pollsters sometimes report the margin of sampling error for a poll with n subjects as roughly given by the formula $100/\sqrt{n}\,\%$. According to this formula, what is the (frequentist's) margin of error for the poll in part (b)? How do you suppose the formula is derived?

Hints: (a) Use R code `qbeta(c(.025,.975), 380, 220)` to find one 95% prior probability interval. (b) One response: $P\{\pi < 0.55\} < 1\%$. (d) A standard formula for an interval with roughly 95% confidence is $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$, where n is "large" and $\hat{p}$ is the sample proportion in favor (see Example 1.6). What value of π maximizes $\pi(1-\pi)$? What if $\pi = 0.4$ or 0.6?

8.2 In Example 8.1, we require a prior distribution with $\mathrm{E}(\pi) \approx 0.55$ and $P\{0.51 < \pi < 0.59\} \approx 0.95$. Here we explore how one might find suitable parameters α and β for such a beta distributed prior.

a) For a beta distribution, the mean is $\mu = \alpha/(\alpha+\beta)$, and the variance is $\sigma^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Also, a beta distribution with large enough values of α and β is roughly normal, so that $P\{\mu - 2\sigma < \pi < \mu + 2\sigma\} \approx 0.95$. Use these facts to find values of α and β that approximately satisfy the requirements.
b) The following R script finds values of α and β that may come close to satisfying the requirements, and then checks to see how well they succeed.

```
alpha = 1:2000       # trial values of alpha
beta = .818*alpha    # corresponding values of beta

# Vector of probabilities for interval (.51, .59)
prob = pbeta(.59, alpha, beta) - pbeta(.51, alpha, beta)
prob.err = abs(.95 - prob)  # errors for probabilities

# Results: Target parameter values
t.al = alpha[prob.err==min(prob.err)]
t.be = round(.818*t.al)
t.al; t.be

# Checking: Achieved mean and probability
a.mean = t.al/(t.al + t.be)
a.mean
a.prob = pbeta(.59, t.al, t.be) - pbeta(.51, t.al, t.be)
a.prob
```

What assumptions about α and β are inherent in the script? Why do we use $\beta = 0.818\alpha$? What values of α and β are returned? For the values of the parameters considered, how close do we get to the desired values of $\mathrm{E}(\pi)$ and $P\{0.51 < \pi < 0.59\}$?
c) If the desired mean is 0.56 and the desired probability in the interval $(0, 51, 0.59)$ is 90%, what values of the parameters are returned by a suitably modified script?

8.3 In practice, the beta family of distributions offers a rich variety of shapes for modeling priors to match expert opinion.

a) Beta densities $p(\pi)$ are defined on the *open* unit interval. Observe that parameter α controls behavior of the density function near 0. In particular, find the value $p(0^+)$ and the slope $p'(0^+)$ in each of the following five cases: $\alpha < 1$, $\alpha = 1$, $1 < \alpha < 2$, $\alpha = 2$, and $\alpha > 2$. Evaluate each limit as being 0, positive and finite, ∞, or $-\infty$. (As usual, 0^+ means to take the limit as the argument approaches 0 through positive values.)

b) By symmetry, parameter β controls behavior of the density function near 1. Thus, combinations of the parameters yield 25 cases, each with its own "shape" of density. In which of these 25 cases does the density have a unique mode in $(0, 1)$? The number of possible inflection points of a beta density curve is 0, 1, or 2. For each of the 25 cases, give the number of inflection points.

c) The R script below plots examples of each of the 25 cases, scaled vertically (with `top`) to show the properties in parts (a) and (b) about as well as can be done and yet show most of each curve.

```
alpha = c(.5, 1, 1.2, 2, 5);  beta = alpha
op = par(no.readonly = TRUE)  # records existing parameters
par(mfrow=c(5, 5))            # formats 5 x 5 matrix of plots
par(mar=rep(2, 4), pty="m")   # sets margins
x = seq(.001, .999, .001)

for (i in 1:5)
{
   for (j in 1:5)  {
      top = .2 + 1.2 * max(dbeta(c(.05, .2, .5, .8, .95),
        alpha[j], beta[i]))
      plot(x,dbeta(x, alpha[i], beta[j]),
        type="l", ylim=c(0, top), xlab="", ylab="",
        main=paste("BETA(",alpha[j],",", beta[i],")", sep=""))  }
}
par(op)                       # restores former parameters
```

Run the code and compare the resulting matrix of plots with your results above (α-cases are rows, β columns). What symmetries within and among the 25 plots are lost if we choose `beta = c(.7, 1, 1.7, 2, 7)`?

8.4 In Example 8.1, we require a prior distribution with $E(\pi) \approx 0.55$ and $P\{0.51 < \pi < 0.59\} \approx 0.95$. If we are willing to use nonbeta priors, how might we find ones that meet these requirements?

a) If we use a normal distribution, what parameters μ and σ would satisfy the requirements?

b) If we use a density function in the shape of an isosceles triangle, show that it should have vertices at $(0.4985, 0)$,$(0.55, 19.43)$, and $(0.6015, 0)$.

c) Plot three priors on the same axes: $\mathsf{BETA}(330, 270)$ of Example 8.1 and the results of parts (a) and (b).

d) Do you think the expert would object to any of these priors as an expression of her feelings about the distribution of π?

Notes: (c) Plot: Your result should be similar to Figure 8.7. Use the method in Example 8.5 to put several plots on the same axes. Experiment: If `v = c(.51, .55, .59)` and `w = c(0, 10, 0)`, then what does `lines(v, w)` add to an existing `plot`? (d) The

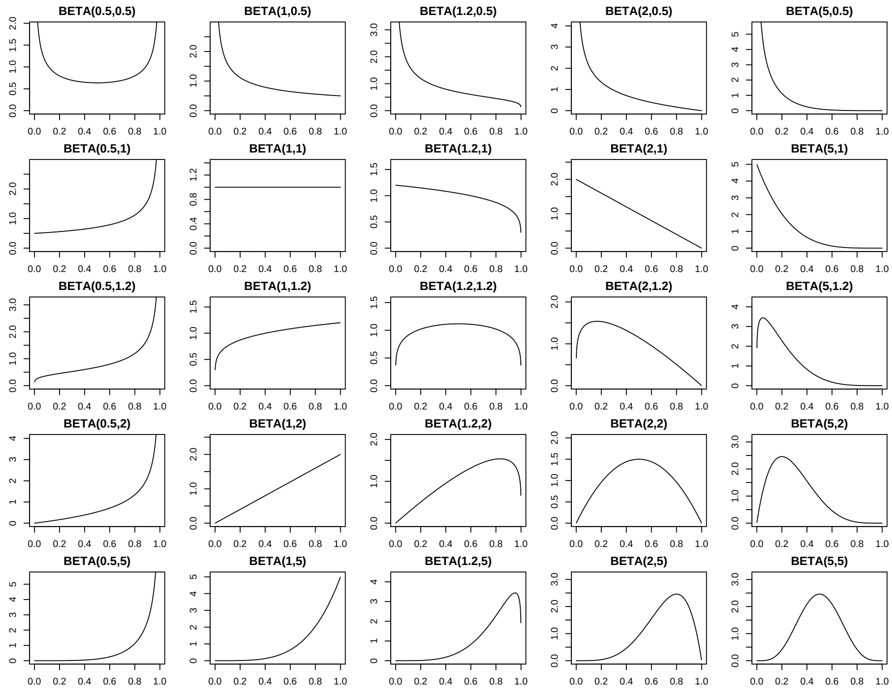

Figure 8.6. Shapes of beta density functions. Shape parameters α and β control the behavior of the density near 0 and 1, respectively; 25 fundamentally different shapes are shown here. (See Problem 8.3.)

triangular prior would be agreeable only if she thinks values of π below 0.4985 or above 0.6015 are *absolutely* impossible.

8.5 Computational methods are often necessary if we multiply the kernels of the prior and likelihood and then can't recognize the result as the kernel of a known distribution. This can occur, for example, when we don't use a conjugate prior. We illustrate several computational methods using the polling situation of Examples 8.1 and 8.5 where we seek to estimate the parameter π.

To begin, suppose we know the beta prior $p(\pi)$ (with $\alpha = 330$ and $\beta = 270$) and the binomial likelihood $p(x|\pi)$ (for $x = 620$ subjects in favor out of $n = 1000$ responding). But we have *not* been clever enough to notice the convenient beta form of the posterior $p(\pi|x)$. We wish to compute the posterior estimate of centrality $\mathrm{E}(\pi|x)$ and the posterior probability $P\{\pi > .6|x\}$ of a "big margin" in favor of the ballot proposition. From the *equation* in (8.2), we have $\mathrm{E}(\pi|x) = \int_0^1 \pi p(\pi)p(x|\pi)\,d\pi/D$ and $P(\pi > 0.6|x) = \int_{0.6}^1 p(\pi)p(x|\pi)\,d\pi/D$, where the denominator of the posterior is $D = \int_0^1 p(\pi)p(x|\pi)\,d\pi$. You should verify these equations for yourself before going on.

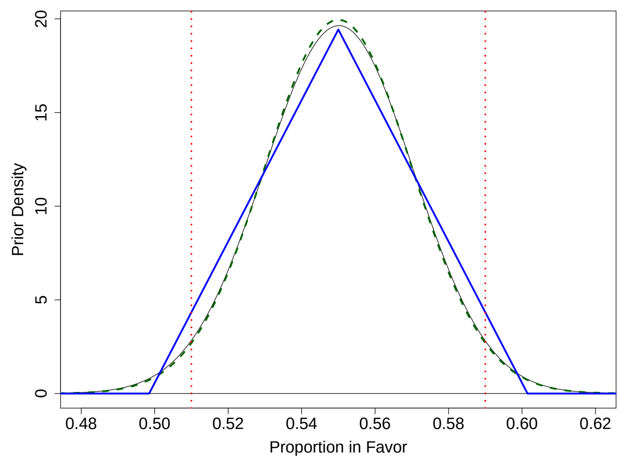

Figure 8.7. Two nonbeta priors. One (thick lines) has an isosceles triangle as its density. The other, NORM$(.55, .02)$ (dashed), is hardly distinguishable from BETA$(330, 270)$ of Example 8.1 (thin). For all three priors $P\{.51 < \pi < .59\} \approx 95\%$. Only the beta prior is conjugate with binomial data. Bayesian inference using the nonbeta priors requires special numerical methods. (See Problems 8.4, 8.5 and 8.6.)

a) The following R script uses Riemann approximation to obtain the desired posterior information. Match key quantities in the program with those in the equations above. Also, interpret the last two lines of code. Run the program and compare the results with those obtainable directly from the known beta posterior of Example 8.5. (In R, `pi` means 3.1416, so we use `pie` for the grid points of parameter π.)

```
x = 620;   n = 1000                                 # data
m = 10000;   pie = seq(0, 1, length=m)              # grid points
igd = dbeta(pie, 330, 270) * dbinom(x, n, pie)      # integrand
d = mean(igd);   d                                  # denominator

# Results
post.mean = mean(pie*igd)/d;  post.mean
post.prob.bigwin = (1/m)*sum(igd[pie > .6])/d;
post.prob.bigwin
post.cum = cumsum((igd/d)/m)
min(pie[post.cum > .025])
min(pie[post.cum > .975])
```

b) Now suppose we choose the prior NORM$(0.55, 0.02)$ to match the expert's impression that the prior should be centered at $\pi = 55\%$ and put 95% of its probability in the interval $51\% < \pi < 59\%$. The shape of this distribution is very similar to BETA$(330, 270)$ (see Problem 8.4). However, the normal prior is *not a conjugate prior.* Write the kernel of the posterior, and say why the method of Example 8.5 is intractable. Modify the program above to use the normal prior (substituting a `dnorm` function for the `dbeta` function). Run the modified program. Compare the results with those in part (a).

c) The scripts in parts (a) and (b) above are "wasteful" because grid values of π are generated throughout $(0, 1)$, but both prior densities are very nearly 0 outside of $(0.45, 0.65)$. Modify the program in part (b) to integrate over this shorter interval.
Strictly speaking, you need to divide `d`, `post.pi.mean`, and so on, by 5 because you are integrating over a region of length 1/5. (Observe the change in `b` if you shorten the interval without dividing by 5.) Nevertheless, show that this correction factor "cancels out" in the main results. Compare your results with those obtained above.

d) Modify the R script of part (c) to do the computation for a normal prior by Monte Carlo integration. Increase the number of iterations to $m \geq 100\,000$, and use `pie = sort(runif(m, .45, .65))`. Part of the program depends on having the π-values sorted in order. Which part? Why? Compare your results with those obtained by Riemann approximation. (If this were a multidimensional integration, some sort of Monte Carlo integration would probably be the method of choice.)

e) (Advanced) Modify part (d) to generate normally distributed values of `pie` (with sorted `rnorm(m, .55,.02)`), removing the `dnorm` factor from the integrand. Explain why this works, and compare the results with those above.
This method is efficient because it concentrates π values in the "important" part of $(0, 1)$ where computed quantities are largest. So there would be no point in restricting the range of integration as in parts (c) and (d). This is an elementary example of **importance sampling**.

8.6 *Metropolis algorithm.* In Section 7.5 we illustrated the Metropolis algorithm as a way to sample from a bivariate normal distribution having a known density function. In Problem 8.5 we considered some methods of computing posterior probabilities that arise from nonconjugate prior distributions. Here we use the Metropolis algorithm in a more serious way than before to sample from posterior distributions arising from the nonconjugate prior distributions of Problem 8.4.

a) Use the Metropolis algorithm to sample from the posterior distribution of π arising from the prior $\mathsf{NORM}(0.55, 0.02)$ and a binomial sample of size $n = 1000$ with $x = 620$ respondents in favor. Simulate $m = 100\,000$ observations from the posterior to find a 95% Bayesian probability interval for π. Also, if you did Problem 8.5, find the posterior probability $P\{\pi > 0.6|x\}$.
The R code below implements this computation using a symmetrical uniform jump function, and compares results with those from the very similar conjugate prior $\mathsf{BETA}(330, 270)$. See the top panel in Figure 8.8.

```
set.seed(1234)
m = 100000
piec = numeric(m);  piec[1] = 0.7              # states of chain
for (i in 2:m)  {
  piec[i] = piec[i-1]                          # if no jump
  piep = runif(1, piec[i-1]-.05, piec[i-1]+.05) # proposal
  nmtr = dnorm(piep, .55, .02)*dbinom(620, 1000, piep)
  dmtr = dnorm(piec[i-1], .55, .02)*dbinom(620, 1000, piec[i-1])
  r = nmtr/dmtr;  acc = (min(r,1) > runif(1))   # accept prop.?
  if(acc) {piec[i] = piep}  }
pie = piec[(m/2+1):m]                          # after burn-in
quantile(pie, c(.025,.975));  mean(pie > .6)
qbeta(c(.025,.975), 950, 650);  1-pbeta(.6, 950, 650)
hist(pie, prob=T, col="wheat", main="")
  xx = seq(.5, .7, len=1000)
  lines(xx, dbeta(xx, 950, 650), lty="dashed", lwd=2)
```

b) Modify the program of part (a) to find the posterior corresponding to the "isosceles" prior of Problem 8.2. Make sure your initial value is within the support of this prior, and use the the following lines of code for the numerator and denominator of the ratio of densities. Notice that, in this ratio, the constant of integration cancels, so it is not necessary to know the height of the triangle. In some more advanced applications of the Metropolis algorithm, the ability to ignore the constant of integration is an important advantage. Explain why results here differ considerably from those in part (a). See the bottom panel in Figure 8.8.

```
nmtr = max(.0515-abs(piep-.55), 0)*dbinom(620, 1000, piep)
dmtr = max(.0515-abs(piec[i-1]-.55), 0)*
         dbinom(620, 1000, piec[i-1])
```

Note: (b) Even though the isosceles prior may seem superficially similar to the beta and normal priors, it puts no probability above 0.615, so the posterior can put no probability there either. In contrast, the data show 620 out of 1000 respondents are in favor.

8.7 A commonly used frequentist principle of estimation provides a point estimate of a parameter by finding the value of the parameter that maximizes the likelihood function. The result is called a **maximum likelihood estimate** (MLE). Here we explore one example of an MLE and its similarity to a particular Bayesian estimate.

Suppose we observe $x = 620$ successes in $n = 1000$ binomial trials and wish to estimate the probability π of success. The likelihood function is $p(x|\pi) \propto \pi^x(1-\pi)^{n-x}$ taken as a function of π.

a) Find the MLE $\hat{\pi}$. A common way to maximize $p(x|\pi)$ in π is to maximize $\ell(\pi) = \ln p(x|\pi)$. Solve $d\ell(\pi)/d\pi = 0$ for π, and verify that you have found an absolute maximum. State the general formula for $\hat{\pi}$ and then its value for $x = 620$ and $n = 1000$.

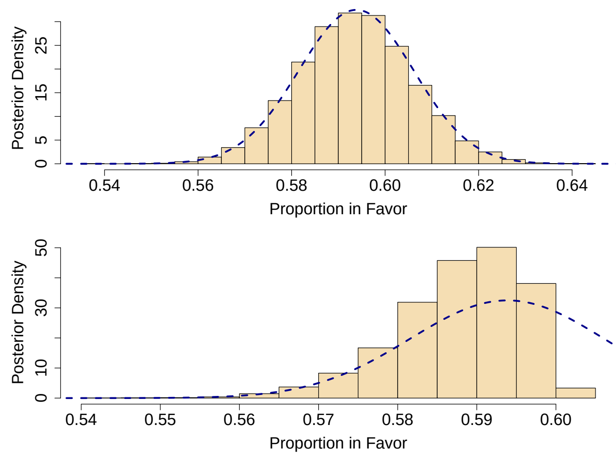

Figure 8.8. Posteriors from nonconjugate priors. Data: 620 subjects in favor out of 1000. Top: The simulated posterior distribution from the prior $\mathsf{NORM}(.55, .02)$ is nearly the same as the posterior $\mathsf{BETA}(950, 650)$ (dashed) from the conjugate prior $\mathsf{BETA}(330, 270)$. Bottom: In contrast, support of the posterior from the "isosceles" prior in Figure 8.7 cannot extend beyond $(0.485, 0.615)$. (See Problem 8.6.)

b) Plot the likelihood function for $n = 1000$ and $x = 620$. Approximate its maximum value from the graph. Then do a numerical maximization with the R script below. Compare with the answer in part (a).

```
pie = seq(.001, .999, .001)      # avoid 'pi' (3.1416)
like = dbinom(620, 1000, pie)
plot(like, type="l");  pie[like==max(like)]
```

c) The interval $\tilde{\pi} \pm 1.96\sqrt{\tilde{\pi}(1-\tilde{\pi})/(n+4)}$, where $\tilde{\pi} = (x+2)/(n+4)$, has approximately 95% confidence for estimating π. (This interval is based on the normal approximation to the binomial; see Example 1.6, p13 and Problems 1.16 and 1.17.) Evaluate its endpoints for 620 successes in 1000 trials.
d) Now we return to Bayesian estimation. A prior distribution that provides little, if any, definite information about the parameter to be estimated is called a **noninformative prior**. A commonly used noninformative beta prior has $\alpha_0 = \beta_0 = 1$, which is the same as $\mathsf{UNIF}(0, 1)$. For this prior and data consisting of x successes in n trials, find the posterior distribution and its mode.
e) For the particular case with $n = 1000$ and $x = 620$, find the posterior mode and a 95% probability interval.

Note: In many estimation problems, the MLE is in close numerical agreement with the Bayesian point estimate based on a noninformative prior and on the posterior mode. Also, a confidence interval based on the MLE may be numerically similar to a Bayesian probability interval from a noninformative prior. But the underlying philosophies of frequentists and Bayesians differ, and so the ways they interpret results in practice may also differ.

Problems Related to Examples 8.2 and 8.6 (Poisson data)

8.8 In a situation similar to that Examples 8.2 and 8.6, suppose that we want to begin with a prior distribution on the parameter λ that has $\mathrm{E}(\lambda) \approx 8$ and $P\{\lambda < 12\} \approx 0.95$. Subsequently, we count a total of $t = 158$ mice in $n = 12$ trappings.

a) To find the parameters of a gamma prior that satisfy the above requirements, **write a program** analogous to the one in Problem 8.2. (You can come very close with α_0 an integer, but don't restrict κ_0 to integer values.)
b) Find the gamma posterior that results from the prior in part (a) and the data given above. Find the posterior mean and a 95% posterior probability interval for λ.
c) As in Figure 8.3(a), plot the prior and the posterior. Why is the posterior here less concentrated than the one in Figure 8.3(a)?
d) The ultimate *noninformative* gamma prior is the *improper* one with $\alpha_0 = \kappa_0 = 0$ (see Problems 8.7 and 8.11 for definitions). Using this prior and the data above, find the posterior mean and a 95% posterior probability interval for λ. Compare with the interval in part (c)?

Partial answers: In (a) you can use a prior with $\alpha_0 = 13$. Our posterior intervals in (c) and (d) agree when rounded to integer endpoints: $(11, 15)$. But not when expressed to one or two-place accuracy—as you should do.

8.9 In this chapter we have computed 95% posterior probability intervals by finding values that cut off 2.5% from each tail. This method is computationally relatively simple and gives satisfactory intervals for most purposes. However, for skewed posterior densities, it does not give the *shortest* interval with 95% probability.

The following R script finds the shortest interval for a gamma posterior. (The vectors `p.low` and `p.up` show endpoints of enough 95% intervals that we can come very close to finding the one for which the length, `long`, is a minimum.)

```
alp = 5; kap = 1
p.lo = seq(.001,.05, .00001)
p.up = .95 + p.lo
q.lo = qgamma(p.lo, alp, kap)
q.up = qgamma(p.up, alp, kap)
long = q.up - q.lo          # avoid confusion with function ‘length’
c(q.lo[long==min(long)], q.up[long==min(long)])
```

a) Compare the length of the shortest interval with that of the usual (probability-symmetric) interval. What probability does the shortest interval put in each tail?

b) Use the same method to find the shortest 95% posterior probability interval in Example 8.6. Compare it with the probability interval given there. Repeat, using suitably modified code, for 99% intervals.
c) Suppose a posterior density function has a single mode and decreases monotonically as the distance away from the mode increases (for example, a gamma density with $\alpha > 1$). Then the shortest 95% posterior probability interval is also the 95% probability interval corresponding to the highest values of the posterior: a **highest posterior density** interval. Explain why this is true. For the 95% intervals in parts (a) and (b), verify that the heights of the posterior density curve are indeed the same at each end of the interval (as far as allowed by the spacing `0.00001` of the probability values used in the script).

8.10 *Mark-recapture estimation of population size.* In order to estimate the number ν of fish in a lake, investigators capture r of these fish at random, tag them, and then release them. Later (leaving time for mixing, but not for significant population change), they capture s fish at random from the lake and observe the number x of tagged fish among them.

Suppose $r = 900$, $s = 1100$, and we observe $x = 103$. [This is similar to the situation described in Problem 4.27 (p112), partially reprised here in parts (a) and (b).]

a) Method of moments estimate (MME). At recapture, an unbiased estimate of the true proportion r/ν of tagged fish in the lake is x/s. That is, $\mathrm{E}(x/s) = r/\nu$. To find the MME of ν, equate the observed value x/s to its expectation and solve for ν. (It is customary to truncate to an integer.)
b) Maximum likelihood estimate (MLE). For known r, s, and ν, the hypergeometric probability distribution function $p_{r,s}(x|\nu) = \binom{r}{x}\binom{\nu-r}{s-x}/\binom{\nu}{s}$ gives the probability of observing x tagged fish at recapture. Once x is observed $p_{r,s}(x|\nu)$, considered as a function of ν, is the likelihood function. Find the MLE; that is, the value of ν that maximizes $p_{r,s}(x|\nu)$.
c) Bayesian interval estimate. Suppose we believe ν lies in $(6000, 14000)$ and are willing to take the prior distribution of ν as uniform on this interval. Use the R code below to find the cumulative posterior distribution of $\nu|x$ and thence a 95% Bayesian interval estimate of ν. Explain the code.

```
r = 900;  s = 1100;  x = 103
nu = 6000:14000;  n = length(nu)
prior = rep(1/n, n)
like = dhyper(x, r, nu-r, s)
denom = sum(prior*like)
post = prior*like/denom;  cumpost = cumsum(post)
c(min(nu[cumpost >= .025]), max(nu[cumpost <= .975]))
```

d) Use the negative binomial prior: `prior = dnbinom(nu-150, 150, .014)`. Compare the resulting Bayesian interval with that of part (c) and with a bootstrap confidence interval obtained as in Problem 4.27.

Problems Related to Examples 8.3 and 8.7 (Normal data, σ known)

8.11 In Example 8.7 we show formulas for the mean and precision of the posterior distribution. Suppose five measurements of the weight of the beam, using a scale known to have precision $\tau = 1$, are: 698.54, 698.45, 696.09, 697.14, 698.62 ($\bar{x} = 697.76$).

a) Based on these data and the prior distribution of Example 8.3, what is the posterior mean of μ? Does it matter whether we choose the mean, the median, or the mode of the posterior distribution as our point estimate? (Explain.) Find a 95% posterior probability interval for μ. Also, suppose we are unwilling to use this beam if it weighs more than 699 pounds; what are the chances of that?
b) Modify the R script shown in Example 8.5 to plot the prior and posterior densities on the same axes. (Your result should be similar to Figure 8.4.)
c) Taking a frequentist point of view, use the five observations given above and the known variance of measurements produced by our scale to give a 95% confidence interval for the true weight of the beam. Compare with the results of part (a) and comment.
d) The prior distribution in this example is very "flat" compared with the posterior: its precision is small. A practically noninformative normal prior is one with precision τ_0 that is much smaller than the precision of the data. As τ_0 decreases, the effect of μ_0 diminishes. Specifically,

$$\lim_{\tau_0 \to 0} \mu_n = \bar{x} \quad \text{and} \quad \lim_{\tau_0 \to 0} \tau_n = n\tau.$$

The *effect* is as if we had used $p(\mu) \propto 1$ as the prior. Of course, such a prior distribution is not strictly possible because $\int_{-\infty}^{\infty} p(\mu)\, d\mu$ would be ∞. However, it is convenient to use such an **improper prior** as shorthand for understanding what happens to a posterior as the prior gets less and less informative. What posterior mean and 95% probability interval result from using an improper prior with our data? Compare with the results of part (c).
e) Now change the example: Suppose that our vendor supplies us with a more consistent product so that the prior $\mathsf{NORM}(701, 5)$ is realistic and that our data above come from a scale with known precision $\tau = 0.4$ Repeat parts (a) and (b) for this situation.

8.12 The purpose of this problem is to derive the posterior distribution $p(\mu|\mathbf{x})$ resulting from the prior $\mathsf{NORM}(\mu_0, \sigma_0)$ and n independent observations $x_i \sim \mathsf{NORM}(\mu, \sigma)$. (See Example 8.7.)

a) Show that the *likelihood* is

$$f(\mathbf{x}|\mu) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \propto \exp\left[-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(\bar{x} - \mu)^2\right].$$

To obtain the first expression above, recall that the likelihood function is the joint density function of $\mathbf{x} = (x_1, \ldots, x_n)|\mu$. To obtain the second, write $(x_i - \mu)^2 = [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$, expand the square, and sum over i. On distributing the sum, you should obtain three terms. One of them provides the desired result, another is 0, and the third is irrelevant because it does not contain the variable μ. (A constant term in the exponential is a constant factor of the likelihood, which is not included in the kernel.)

b) To derive the expression for the kernel of the *posterior,* multiply the kernels of the prior and the likelihood, and expand the squares in each. Then put everything in the exponential over a common denominator, and collect terms in μ^2 and μ. Terms in the exponent that do not involve μ are constant factors of the posterior density that may be adjusted as required in completing the square to obtain the desired posterior kernel.

Problems Related to Examples 8.3 and 8.7 (Normal data, $\mu = 0$)

8.13 For a pending American football game, the "point spread" is established by experts as a measure of the difference in the ability of the two teams. The point spread is often of interest to gamblers. Roughly speaking, the favored team is thought to be just as likely to win by more than the point spread as to win by less or to lose. So ideally a fair bet that the favored team "beats the spread" could be made at even odds. Here we are interested in the difference $x = v - w$ between the point spread v, which might be viewed as the favored team's predicted lead, and the actual point difference w (favored team's score minus opponent's) when the game is played.

a) Suppose an amateur gambler, perhaps interested in bets that would not have even odds, is interested in the precision of x and is willing to assume $x \sim \mathsf{NORM}(0, \sigma)$. Also, recalling relatively few instances with $|x| > 30$, he decides to use a prior distribution on σ that satisfies $P\{10 < \sigma < 20\} = P\{100 < \sigma^2 = 1/\tau < 400\} = P\{1/400 < \tau < 1/100\} = 0.95$. Find parameters α_0 and κ_0 for a gamma-distributed prior on τ that approximately satisfy this condition. (Use a program similar to the one in Problem 8.2.)
b) Suppose data for point spreads and scores of 146 professional football games show $s = (\sum x_i^2/n)^{1/2} = 13.3$. Under the prior distribution of part (a), what 95% posterior probability intervals for τ and σ result from these data?
c) Use the noninformative improper prior distribution with $\alpha_0 = \kappa_0 = 0$ and the data of part (b) to find 95% posterior probability intervals for τ and σ. Also, use these data to find the frequentist 95% confidence interval for σ based on the distribution $\mathsf{CHISQ}(146)$, and compare it with the posterior probability interval for σ.

Notes and hints: (a) Parameters $\alpha_0 = 11, \kappa_0 = 2500$ give probability 0.945 and might be used for part (b), but a properly written program will give integers that come closer to 95%. (b) The data $\mathbf{x}$ in part (b), taken from more extensive data available online [Ste92], are for 1992 NFL home games; $\bar{x} \approx 0$ and the data pass standard tests for normality. For a more detailed discussion and analysis of point spreads see [Ste91]. (c) The two intervals for σ agree closely, roughly (12, 15). You should report results to one decimal place.

8.14 We want to know the precision of an analytic device. We believe its readings to be normally distributed and unbiased. We have five standard specimens of known value to use in testing the device, so we can observe the error x_i that the device makes for each specimen. Thus we assume that the x_i are independent $\mathsf{NORM}(0, \sigma)$, and we wish to estimate $\sigma = 1/\sqrt{\tau}$.

a) We use information from the manufacturer of the device to determine a gamma-distributed prior for τ. This information is provided in terms of σ. Specifically, we want the prior to be consistent with a median of about 0.65 for σ and with $P\{\sigma < 1\} \approx 0.95$. If a gamma prior distribution on τ has parameter $\alpha_0 = 5$, then what value of the parameter κ_0 comes close to meeting these requirements?
b) The following five errors are observed when analyzing test specimens: $-2.65, 0.52, 1.82, -1.41, 1.13$. Based on the prior distribution in part (a) and these data, find the posterior distribution, the posterior *median* value of τ, and a 95% posterior probability interval for τ. Use these to give the posterior median value of σ and a 95% posterior probability interval for σ.
c) On the same axes, make plots of the prior and posterior distributions of τ. Comment.
d) Taking a frequentist approach, find the maximum likelihood estimate (MLE) $\hat{\tau}$ of τ based on the data given in part (b). Also, find 95% confidence intervals for σ^2, σ, and τ. Use the fact that $\sum_{i=1}^{n} x_i^2/\sigma^2 \sim \mathsf{CHISQ}(n) = \mathsf{GAMMA}(n/2, 1/2)$. Compare these with the Bayesian results in part (b).

Notes: The invariance principle of MLEs states that $\hat{\tau} = 1/\widehat{\sigma^2} = 1/\hat{\sigma}^2$, where "hats" indicate MLEs of the respective parameters. Also, the median of a random variable is invariant under any monotone transformation. Thus, for the prior or posterior distribution of τ (always positive), $\text{Med}(\tau) = 1/\text{Med}(\sigma^2) = 1/[\text{Med}(\sigma)]^2$. But, in general, expectation is invariant only under *linear* transformations. For example, $\text{E}(\tau) \neq 1/\text{E}(\sigma^2)$ and $\text{E}(\sigma^2) \neq [\text{E}(\sigma)]^2$.