Chapter 7
Bayesian Inference

CHAPTER OUTLINE

In Chapter 5, we introduced the basic concepts of inference. At the heart of the theory of inference is the concept of the statistical model $\{f_\theta : \theta \in \Omega\}$ that describes the statistician's uncertainty about how the observed data were produced. Chapter 6 dealt with the analysis of this uncertainty based on the model and the data alone. In some cases, this seemed quite successful, but we note that we only dealt with some of the simpler contexts there.

If we accept the principle that, to be amenable to analysis, all uncertainties need to be described by probabilities, then the prescription of a model alone is incomplete, as this does not tell us how to make probability statements about the unknown true value of θ. In this chapter, we complete the description so that all uncertainties are described by probabilities. This leads to a probability distribution for θ, and, in essence, we are in the situation of Section 5.2, with the parameter now playing the role of the unobserved response. This is the Bayesian approach to inference.

Many statisticians prefer to develop statistical theory without the additional ingredients necessary for a full probability description of the unknowns. In part, this is motivated by the desire to avoid the prescription of the additional model ingredients necessary for the Bayesian formulation. Of course, we would prefer to have our statistical analysis proceed based on the fewest and weakest model assumptions possible. For example, in Section 6.4, we introduced distribution-free methods. A price is paid for this weakening, however, and this typically manifests itself in ambiguities about how inference should proceed. The Bayesian formulation in essence removes the ambiguity, but at the price of a more involved model.

The Bayesian approach to inference is sometimes presented as antagonistic to methods that are based on repeated sampling properties (often referred to as *frequentist*

methods), as discussed, for example, in Chapter 6. The approach taken in this text, however, is that the Bayesian model arises naturally from the statistician assuming more ingredients for the model. It is up to the statistician to decide what ingredients can be justified and then use appropriate methods. We must be wary of *all* model assumptions, because using inappropriate ones may invalidate our inferences. Model checking will be taken up in Chapter 9.

7.1 The Prior and Posterior Distributions

The *Bayesian model* for inference contains the statistical model $\{f_\theta : \theta \in \Omega\}$ for the data $s \in S$ and adds to this the *prior probability measure* Π for θ. The prior describes the statistician's beliefs about the true value of the parameter θ *a priori*, i.e., before observing the data. For example, if $\Omega = [0, 1]$ and θ equals the probability of getting a head on the toss of a coin, then the prior density π plotted in Figure 7.1.1 indicates that the statistician has some belief that the true value of θ is around 0.5. But this information is not very precise.

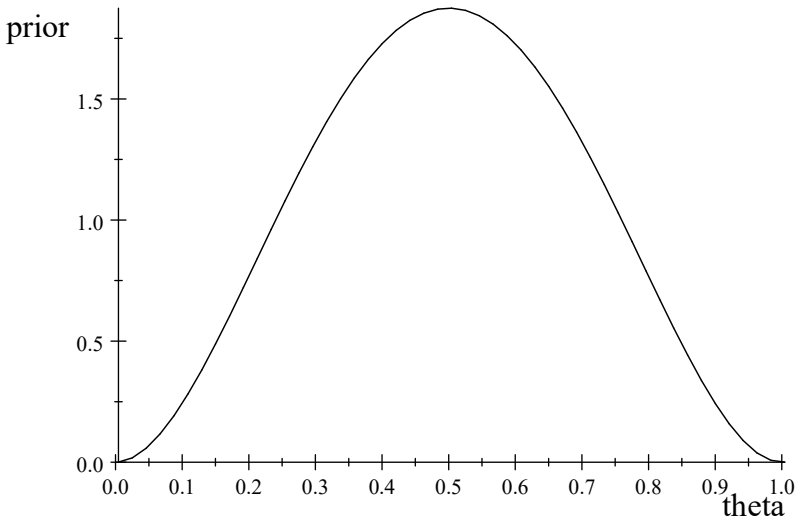

Figure 7.1.1: A fairly diffuse prior on [0,1].

On the other hand, the prior density π plotted in Figure 7.1.2indicates that the statistician has very precise information about the true value of θ. In fact, if the statistician knows nothing about the true value of θ, then using the uniform distribution on [0, 1] might be appropriate.

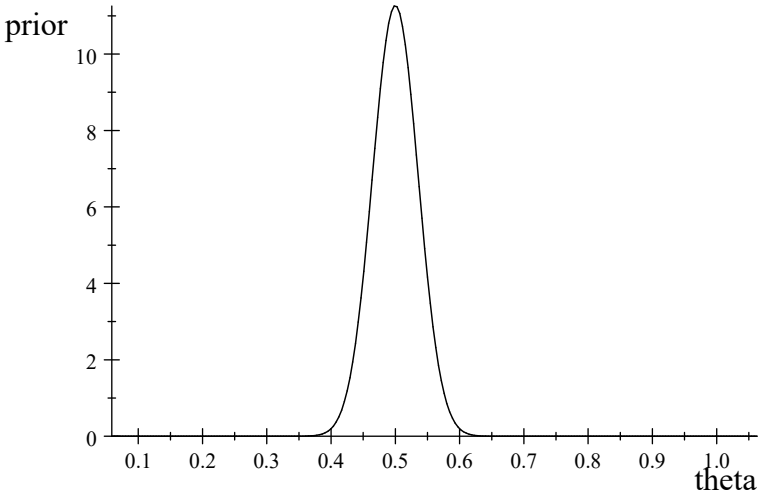

Figure 7.1.2: A fairly precise prior on [0,1].

It is important to remember that the probabilities prescribed by the prior represent beliefs. They do not in general correspond to long-run frequencies, although they could in certain circumstances. A natural question to ask is: Where do these beliefs come from in an application? An easy answer is to say that they come from previous experience with the random system under investigation or perhaps with related systems. To be honest, however, this is rarely the case, and one has to admit that the prior, as well as the statistical model, is often a somewhat arbitrary construction used to drive the statistician's investigations. This raises the issue as to whether or not the inferences derived have any relevance to the practical context, if the model ingredients suffer from this arbitrariness. This is where the concept of model checking comes into play, a topic we will discuss in Chapter 9. At this point, we will assume that all the ingredients make sense, but remember that in an application, these must be checked if the inferences taken are to be practically meaningful.

We note that the ingredients of the Bayesian formulation for inference prescribe a marginal distribution for θ, namely, the prior Π, and a set of conditional distributions for the data s given θ, namely, $\{f_\theta : \theta \in \Omega\}$. By the law of total probability (Theorems 2.3.1 and 2.8.1), these ingredients specify a joint distribution for (s, θ), namely,

$$\pi(\theta) f_\theta(s),$$

where π denotes the probability or density function associated with Π. When the prior distribution is absolutely continuous, the marginal distribution for s is given by

$$m(s) = \int_\Omega \pi(\theta) f_\theta(s)\, d\theta$$

and is referred to as the *prior predictive distribution* of the data. When the prior distribution of θ is discrete, we replace (as usual) the integral by a sum.

If we did not observe any data, then the prior predictive distribution is the relevant distribution for making probability statements about the unknown value of s. Similarly, the prior π is the relevant distribution to use in making probability statements about θ, before we observe s. Inference about these unobserved quantities then proceeds as described in Section 5.2.

Recall now the principle of conditional probability; namely, $P(A)$ is replaced by $P(A \mid C)$ after we are told that C is true. Therefore, after observing the data, the relevant distribution to use in making probability statements about θ is the conditional distribution of θ given s. We denote this conditional probability measure by $\Pi(\cdot \mid s)$ and refer to it as the posterior distribution of θ. Note that the density (or probability function) of the posterior is obtained immediately by taking the joint density $\pi(\theta) f_\theta(s)$ of (s, θ) and dividing it by the marginal $m(s)$ of s.

Definition 7.1.1 The *posterior distribution* of θ is the conditional distribution of θ, given s. The *posterior density*, or *posterior probability function* (whichever is relevant), is given by

$$\pi(\theta \mid s) = \frac{\pi(\theta) f_\theta(s)}{m(s)}. \tag{7.1.1}$$

Sometimes this use of conditional probability is referred to as an application of Bayes' theorem (Theorem 1.5.2). This is because we can think of a value of θ being selected first according to π, and then s is generated from f_θ. We then want to make probability statements about the first stage, having observed the outcome of the second stage. It is important to remember, however, that choosing to use the posterior distribution for probability statements about θ is an axiom, or principle, not a theorem.

We note that in (7.1.1) the prior predictive of the data s plays the role of the *inverse normalizing constant* for the posterior density. By this we mean that the posterior density of θ is proportional to $\pi(\theta) f_\theta(s)$, as a function of θ; to convert this into a proper density function, we need only divide by $m(s)$. In many examples, we do not need to compute the inverse normalizing constant. This is because we recognize the functional form, as a function of θ, of the posterior from the expression $\pi(\theta) f_\theta(s)$ and so immediately deduce the posterior probability distribution of θ. Also, there are Monte Carlo methods, such as those discussed in Chapter 4, that allow us to sample from $\pi(\theta \mid s)$ without knowing $m(s)$ (also see Section 7.3).

We consider some applications of Bayesian inference.

EXAMPLE 7.1.1 *Bernoulli Model*

Suppose that we observe a sample $(x_1, \ldots, x_n)$ from the Bernoulli(θ) distribution with $\theta \in [0, 1]$ unknown. For the prior, we take π to be equal to a Beta(α, β) density (see Problem 2.4.16). Then the posterior of θ is proportional to the likelihood

$$\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{n\bar{x}} (1-\theta)^{n(1-\bar{x})}$$

times the prior

$$B^{-1}(\alpha, \beta)\, \theta^{\alpha-1} (1-\theta)^{\beta-1}.$$

This product is proportional to

$$\theta^{n\bar{x}+\alpha-1}\,(1-\theta)^{n(1-\bar{x})+\beta-1}\,.$$

We recognize this as the unnormalized density of a Beta$(n\bar{x}+\alpha, n\,(1-\bar{x})+\beta)$ distribution. So in this example, we did not need to compute $m(x_1,\ldots,x_n)$ to obtain the posterior.

As a specific case, suppose that we observe $n\bar{x} = 10$ in a sample of $n = 40$ and $\alpha = \beta = 1$, i.e., we have a uniform prior on θ. Then the posterior of θ is given by the Beta$(11, 31)$ distribution. We plot the posterior density in Figure 7.1.3 as well as the prior.

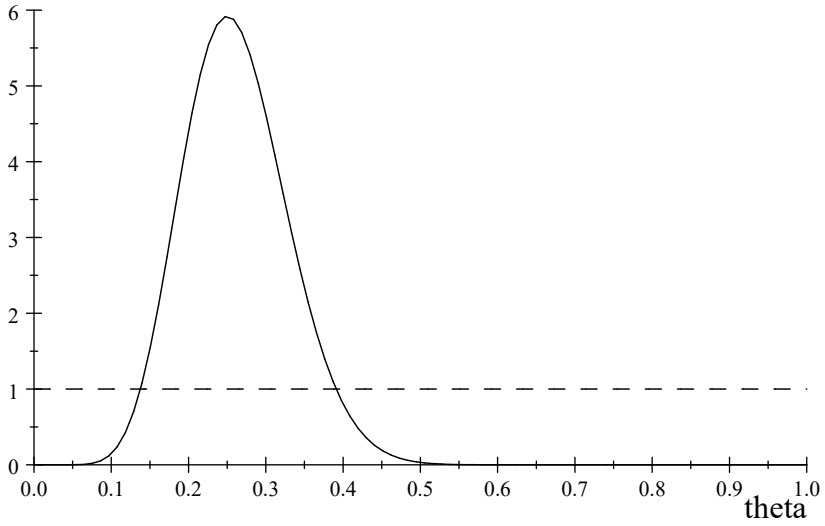

Figure 7.1.3: Prior (dashed line) and posterior densities (solid line) in Example 7.1.1.

The spread of the posterior distribution gives us some idea of the precision of any probability statements we make about θ. Note how much information the data have added, as reflected in the graphs of the prior and posterior densities. ∎

EXAMPLE 7.1.2 *Location Normal Model*
Suppose that $(x_1,\ldots,x_n)$ is a sample from an $N(\mu,\sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known. The likelihood function is then given by

$$L\,(\mu \mid x_1,\ldots,x_n) = \exp\left(-\frac{n}{2\sigma_0^2}\,(\bar{x}-\mu)^2\right).$$

Suppose we take the prior distribution of μ to be an $N(\mu_0,\tau_0^2)$ for some specified choice of μ_0 and τ_0^2. The posterior density of μ is then proportional to

$$
\begin{aligned}
&\exp\left\{-\frac{1}{2\tau_0^2}\left(\mu-\mu_0\right)^2\right\}\exp\left\{-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu\right)^2\right\}\\
&=\exp\left\{-\frac{1}{2\tau_0^2}\left(\mu^2-2\mu\mu_0+\mu_0^2\right)-\frac{n}{2\sigma_0^2}\left(\bar{x}^2-2\mu\bar{x}+\mu^2\right)\right\}\\
&=\exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\left[\mu^2-2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)\mu\right]\right\}\\
&\quad\times\exp\left\{-\frac{\mu_0^2}{2\tau_0^2}-\frac{n\bar{x}^2}{2\sigma_0^2}\right\}\\
&=\exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\left(\mu-\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)\right)^2\right\}\\
&\quad\times\exp\left\{\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)^2\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{n\bar{x}^2}{\sigma_0^2}\right)\right\}. \qquad (7.1.2)
\end{aligned}
$$

We immediately recognize this, as a function of μ, as being proportional to the density of an

$$
N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right),\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)
$$

distribution.

Notice that the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{x}$, with weights

$$
\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\frac{1}{\tau_0^2}\ \text{and}\ \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\frac{n}{\sigma_0^2},
$$

respectively. This implies that the posterior mean lies between the prior mean and the sample mean.

Furthermore, the posterior variance is smaller than the variance of the sample mean. So if the information expressed by the prior is accurate, inferences about μ based on the posterior will be more accurate than those based on the sample mean alone. Note that the more diffuse the prior is — namely, the larger τ_0^2 is — the less influence the prior has. For example, when $n = 20$ and $\sigma_0^2 = 1$, $\tau_0^2 = 1$, then the ratio of the posterior variance to the sample mean variance is $20/21 \approx 0.95$. So there has been a 5% improvement due to the use of prior information.

For example, suppose that $\sigma_0^2 = 1, \mu_0 = 0, \tau_0^2 = 2$, and that for $n = 10$, we observe $\bar{x} = 1.2$. Then the prior is an $N(0, 2)$ distribution, while the posterior is an

$$N\left(\left(\frac{1}{2} + \frac{10}{1}\right)^{-1}\left(\frac{0}{2} + \frac{10}{1}1.2\right), \left(\frac{1}{2} + \frac{10}{1}\right)^{-1}\right) = N(1.1429, 9.523\,8 \times 10^{-2})$$

distribution. These densities are plotted in Figure 7.1.4. Notice that the posterior is quite concentrated compared to the prior, so we have learned a lot from the data. ∎

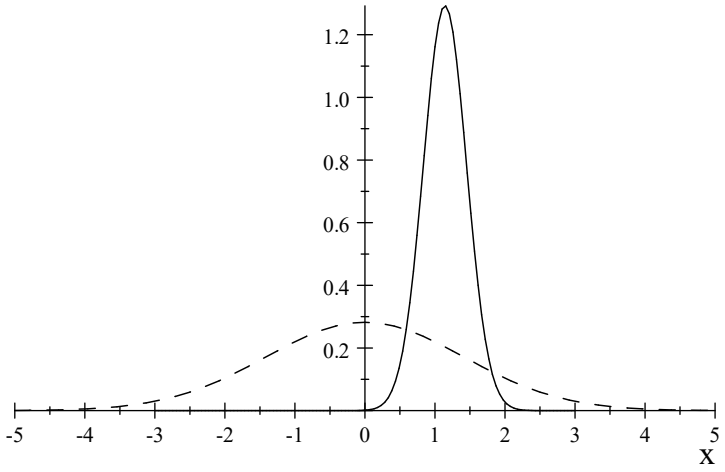

Figure 7.1.4: Plot of the $N(0, 2)$ prior (dashed line) and the $N\left(1.1429, 9.523\,8 \times 10^{-2}\right)$ posterior (solid line) in Example 7.1.2.

EXAMPLE 7.1.3 *Multinomial Model*

Suppose we have a categorical response s that takes k possible values, say, $s \in S = \{1, \ldots, k\}$. For example, suppose we have a bowl containing chips labelled one of $1, \ldots, k$. A proportion θ_i of the chips are labelled i, and we randomly draw a chip, observing its label.

When the θ_i are unknown, the statistical model is given by

$$\{p_{(\theta_1,\ldots,\theta_k)} : (\theta_1, \ldots, \theta_k) \in \Omega\},$$

where $p_{(\theta_1,\ldots,\theta_k)}(i) = P(s = i) = \theta_i$ and

$$\Omega = \{(\theta_1, \ldots, \theta_k) : 0 \leq \theta_i \leq 1, i = 1, \ldots, k \text{ and } \theta_1 + \cdots + \theta_k = 1\}.$$

Note that the parameter space is really only $(k-1)$-dimensional because, for example, $\theta_k = 1 - \theta_1 - \cdots - \theta_{k-1}$, namely, once we have determined $k - 1$ of the θ_i, the remaining value is specified.

Now suppose we observe a sample $(s_1, \ldots, s_n)$ from this model. Let the frequency (count) of the ith category in the sample be denoted by x_i. Then, from Example 2.8.5, we see that the likelihood is given by

$$L(\theta_1, \ldots, \theta_k \mid (s_1, \ldots, s_n)) = \theta_1^{x_1}\theta_2^{x_2} \cdots \theta_k^{x_k}.$$

For the prior we assume that $(\theta_1, \ldots, \theta_{k-1}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ with density (see Problem 2.7.13) given by

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{7.1.3}$$

for $(\theta_1, \ldots, \theta_k) \in \Omega$ (recall that $\theta_k = 1 - \theta_1 - \cdots - \theta_{k-1}$). The α_i are nonnegative constants chosen by the statistician to reflect her beliefs about the unknown value of $(\theta_1, \ldots, \theta_k)$. The choice $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$ corresponds to a uniform distribution, as then (7.1.3) is constant on Ω.

The posterior density of $(\theta_1, \ldots, \theta_{k-1})$ is then proportional to

$$\theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \cdots \theta_k^{x_k + \alpha_k - 1}$$

for $(\theta_1, \ldots, \theta_k) \in \Omega$. From (7.1.3), we immediately deduce that the posterior distribution of $(\theta_1, \ldots, \theta_{k-1})$ is $\text{Dirichlet}(x_1 + \alpha_1, x_2 + \alpha_2, \ldots, x_k + \alpha_k)$. ∎

EXAMPLE 7.1.4 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown. The likelihood function is then given by

$$L\left(\mu, \sigma^2 \,|\, x_1, \ldots, x_n\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2} s^2\right).$$

Suppose we put the following prior on (μ, σ^2). First, we specify that

$$\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2 \sigma^2),$$

i.e., the conditional prior distribution of μ given σ^2 is normal with mean μ_0 and variance $\tau_0^2 \sigma^2$. Then we specify the marginal prior distribution of σ^2 as

$$\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0). \tag{7.1.4}$$

Sometimes (7.1.4) is referred to by saying that σ^2 is distributed *inverse Gamma*. The values μ_0, τ_0^2, α_0, and β_0 are selected by the statistician to reflect his prior beliefs.

From this, we can deduce (see Section 7.5 for the full derivation) that the posterior distribution of (μ, σ^2) is given by

$$\mu \,|\, \sigma^2, x_1, \ldots, x_n \sim N\left(\mu_x, \left(n + \frac{1}{\tau_0^2}\right)^{-1} \sigma^2\right) \tag{7.1.5}$$

and

$$\frac{1}{\sigma^2} \,|\, x_1, \ldots, x_n \sim \text{Gamma}(\alpha_0 + n/2, \beta_x) \tag{7.1.6}$$

where

$$\mu_x = \left(n + \frac{1}{\tau_0^2}\right)^{-1} \left(\frac{\mu_0}{\tau_0^2} + n\bar{x}\right) \tag{7.1.7}$$

and

$$\beta_x = \beta_0 + \frac{n-1}{2}s^2 + \frac{1}{2}\frac{n(\bar{x}-\mu_0)^2}{1+n\tau_0^2}. \tag{7.1.8}$$

To generate a value (μ, σ^2) from the posterior, we can make use of the method of composition (see Problem 2.10.13) by first generating σ^2 using (7.1.6) and then using (7.1.5) to generate μ. We will discuss this further in Section 7.3.

Notice that as $\tau_0 \to \infty$, i.e., as the prior on μ becomes increasingly diffuse, the conditional posterior distribution of μ given σ^2 converges in distribution to an $N(\bar{x}, \sigma^2/n)$ distribution because

$$\mu_x \to \bar{x} \tag{7.1.9}$$

and

$$\left(n + \frac{1}{\tau_0^2}\right)^{-1} \to \frac{1}{n}. \tag{7.1.10}$$

Furthermore, as $\tau_0 \to \infty$ and $\beta_0 \to 0$, the marginal posterior of $1/\sigma^2$ converges in distribution to a Gamma$\left(\alpha_0 + n/2, (n-1)s^2/2\right)$ distribution because

$$\beta_x \to (n-1)s^2/2. \tag{7.1.11}$$

Actually, it does not really seem to make sense to let $\tau_0 \to \infty$ and $\beta_0 \to 0$ in the prior distribution of (μ, σ^2), as the prior does not converge to a proper probability distribution. The idea here, however, is that we think of taking τ_0 large and β_0 small, so that the posterior inferences are approximately those obtained from the limiting posterior. There is still a need to choose α_0, however, even in the diffuse case, as the limiting inferences are dependent on this quantity. ∎

Summary of Section 7.1

- Bayesian inference adds the prior probability distribution to the sampling model for the data as an additional ingredient to be used in determining inferences about the unknown value of the parameter.
- Having observed the data, the principle of conditional probability leads to the posterior distribution of the parameter as the basis for inference.
- Inference about marginal parameters is handled by marginalizing the full posterior.

EXERCISES

7.1.1 Suppose that $S = \{1, 2\}$, $\Omega = \{1, 2, 3\}$, and the class of probability distributions for the response s is given by the following table.

	$s=1$	$s=2$
$f_1(s)$	1/2	1/2
$f_2(s)$	1/3	2/3
$f_3(s)$	3/4	1/4

If we use the prior $\pi(\theta)$ given by the table

	$\theta = 1$	$\theta = 2$	$\theta = 3$
$\pi(\theta)$	1/5	2/5	2/5

,

then determine the posterior distribution of θ for each possible sample of size 2.

7.1.2 In Example 7.1.1, determine the posterior mean and variance of θ.

7.1.3 In Example 7.1.2, what is the posterior probability that μ is positive, given that $n = 10, \bar{x} = 1$ when $\sigma_0^2 = 1, \mu_0 = 0$, and $\tau_0^2 = 10$? Compare this with the prior probability of this event.

7.1.4 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Poisson(λ) distribution with $\lambda \geq 0$ unknown. If we use the prior distribution for λ given by the Gamma(α, β) distribution, then determine the posterior distribution of λ.

7.1.5 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Uniform$[0, \theta]$ distribution with $\theta > 0$ unknown. If the prior distribution of θ is Gamma(α, β) , then obtain the form of the posterior density of θ.

7.1.6 Find the posterior mean and variance of θ_i in Example 7.1.3 when $k = 3$. (Hint: See Problems 3.2.16 and 3.3.20.)

7.1.7 Suppose we have a sample

6.56	6.39	3.30	3.03	5.31	5.62	5.10	2.45	8.24	3.71
4.14	2.80	7.43	6.82	4.75	4.09	7.95	5.84	8.44	9.36

from an $N(\mu, \sigma^2)$ distribution and we determine that a prior specified by $\mu \mid \sigma^2 \sim N(3, 4\sigma^2), \sigma^{-2} \sim$ Gamma(1, 1) is appropriate. Determine the posterior distribution of $(\mu, 1/\sigma^2)$.

7.1.8 Suppose that the prior probability of θ being in a set $A \subset \Omega$ is 0.25 and the posterior probability of θ being in A is 0.80.

(a) Explain what effect the data have had on your beliefs concerning the true value of θ being in A.

(b) Explain why a posterior probability is more relevant to report than is a prior probability.

7.1.9 Suppose you toss a coin and put a Uniform[0.4, 0.6] prior on θ, the probability of getting a head on a single toss.

(a) If you toss the coin n times and obtain n heads, then determine the posterior density of θ.

(b) Suppose the true value of θ is, in fact, 0.99. Will the posterior distribution of θ ever put any probability mass around $\theta = 0.99$ for any sample of n?

(c) What do you conclude from part (b) about how you should choose a prior ?

7.1.10 Suppose that for statistical model $\{f_\theta : \theta \in R^1\}$, we assign the prior density π. Now suppose that we reparameterize the model via the function $\psi = \Psi(\theta)$, where $\Psi : R^1 \to R^1$ is differentiable and strictly increasing.

(a) Determine the prior density of ψ.

(b) Show that $m(x)$ is the same whether we parameterize the model by θ or by ψ.

7.1.11 Suppose that for statistical model $\{f_\theta : \theta \in \Omega\}$, where $\Omega = \{-2, -1, 0, 1, 2, 3\}$, we assign the prior probability function π, which is uniform on Ω. Now suppose we are interested primarily in making inferences about $|\theta|$.

(a) Determine the prior probability distribution of $|\theta|$. Is this distribution uniform?

(b) A uniform prior distribution is sometimes used to express complete ignorance about the value of a parameter. Does complete ignorance about the value of a parameter imply complete ignorance about a function of a parameter? Explain.

7.1.12 Suppose that for statistical model $\{f_\theta : \theta \in [0, 1]\}$, we assign the prior density π, which is uniform on $\Omega = [0, 1]$. Now suppose we are interested primarily in making inferences about θ^2.

(a) Determine the prior density of θ^2. Is this distribution uniform?

(b) A uniform prior distribution is sometimes used to express complete ignorance about the value of a parameter. Does complete ignorance about the value of a parameter imply complete ignorance about a function of a parameter? Explain.

COMPUTER EXERCISES

7.1.13 In Example 7.1.2, when $\mu_0 = 2, \tau_0^2 = 1, \sigma_0^2 = 1, n = 20$, and $\bar{x} = 8.2$, generate a sample of 10^4 (or as large as possible) from the posterior distribution of μ and estimate the posterior probability that the coefficient of variation is greater than 0.125, i.e., the posterior probability that $\sigma_0/\mu > 0.125$. Estimate the error in your approximation.

7.1.14 In Example 7.1.2, when $\mu_0 = 2, \tau_0^2 = 1, \sigma_0^2 = 1, n = 20$, and $\bar{x} = 8.2$, generate a sample of 10^4 (or as large as possible) from the posterior distribution of μ and estimate the posterior expectation of the coefficient of variation σ_0/μ. Estimate the error in your approximation.

7.1.15 In Example 7.1.1, plot the prior and posterior densities on the same graph and compare them when $n = 30, \bar{x} = 0.73, \alpha = 3$, and $\beta = 3$. (Hint: Calculate the logarithm of the posterior density and then exponentiate this. You will need the *log-gamma function* defined by $\ln \Gamma(\alpha)$ for $\alpha > 0$.)

PROBLEMS

7.1.16 Suppose the prior of a real-valued parameter θ is given by the $N(\theta_0, \tau^2)$ distribution. Show that this distribution does not converge to a probability distribution as $\tau \to \infty$. (Hint: Consider the limits of the distribution functions.)

7.1.17 Suppose that $(x_1, \ldots, x_n)$ is a sample from $\{f_\theta : \theta \in \Omega\}$ and that we have a prior π. Show that if we observe a further sample $(x_{n+1}, \ldots, x_{n+m})$, then the posterior you obtain from using the posterior $\pi(\cdot \mid x_1, \ldots, x_n)$ as a prior, and then conditioning on $(x_{n+1}, \ldots, x_{n+m})$, is the same as the posterior obtained using the prior π and conditioning on $(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})$. This is the *Bayesian updating* property.

7.1.18 In Example 7.1.1, determine $m(x)$. If you were asked to generate a value from this distribution, how would you do it? (Hint: For the generation part, use the theorem of total probability.)

7.1.19 Prove that the posterior distribution depends on the data only through the value of a sufficient statistic.

COMPUTER PROBLEMS

7.1.20 For the data of Exercise 7.1.7, plot the prior and posterior densities of σ^2 over $(0, 10)$ on the same graph and compare them. (Hint: Evaluate the logarithms of the densities first and then plot the exponential of these values.)

7.1.21 In Example 7.1.4, when $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1, n = 20, \bar{x} = 8.2$, and $s^2 = 2.1$, generate a sample of 10^4 (or as large as is feasible) from the posterior distribution of σ^2 and estimate the posterior probability that $\sigma > 2$. Estimate the error in your approximation.

7.1.22 In Example 7.1.4, when $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1, n = 20, \bar{x} = 8.2$, and $s^2 = 2.1$, generate a sample of 10^4 (or as large as is feasible) from the posterior distribution of (μ, σ^2) and estimate the posterior expectation of σ. Estimate the error in your approximation.

DISCUSSION TOPICS

7.1.23 One of the objections raised concerning Bayesian inference methodology is that it is subjective in nature. Comment on this and the role of subjectivity in scientific investigations.

7.1.24 Two statisticians are asked to analyze a data set x produced by a system under study. Statistician I chooses to use a sampling model $\{f_\theta : \theta \in \Omega\}$ and prior π_I, while statistician II chooses to use a sampling model $\{g_\psi : \psi \in \Psi\}$ and prior π_{II}. Comment on the fact that these ingredients can be completely different and so the subsequent analyses completely different. What is the relevance of this for the role of subjectivity in scientific analyses of data?

7.2 Inferences Based on the Posterior

In Section 7.1, we determined the posterior distribution of θ as a fundamental object of Bayesian inference. In essence, the principle of conditional probability asserts that the posterior distribution $\pi(\theta \mid s)$ contains all the relevant information in the sampling model $\{f_\theta : \theta \in \Omega\}$, the prior π and the data s, about the unknown true value of θ. While this is a major step forward, it does not completely tell us how to make the types of inferences we discussed in Section 5.5.3.

In particular, we must specify how to compute estimates, credible regions, and carry out hypothesis assessment — which is what we will do in this section. It turns out that there are often several plausible ways of proceeding, but they all have the common characteristic that they are based on the posterior.

In general, we are interested in specifying inferences about a real-valued characteristic of interest $\psi(\theta)$. One of the great advantages of the Bayesian approach is that inferences about ψ are determined in the same way as inferences about the full parameter θ, but with the marginal posterior distribution for ψ replacing the full posterior.

This situation can be compared with the likelihood methods of Chapter 6, where it is not always entirely clear how we should proceed to determine inferences about ψ based upon the likelihood. Still, we have paid a price for this in requiring the addition of another model ingredient, namely, the prior.

So we need to determine the posterior distribution of ψ. This can be a difficult task in general, even if we have a closed-form expression for $\pi(\theta \mid s)$. When the posterior distribution of θ is discrete, the posterior probability function of ψ is given by

$$\omega(\psi_0 \mid s) = \sum_{\{\theta:\psi(\theta)=\psi_0\}} \pi(\theta \mid s).$$

When the posterior distribution of θ is absolutely continuous, we can often find a complementing function $\lambda(\theta)$ so that $h(\theta) = (\psi(\theta), \lambda(\theta))$ is 1–1, and such that the methods of Section 2.9.2 can be applied. Then, denoting the inverse of this transformation by $\theta = h^{-1}(\psi, \lambda)$, the methods of Section 2.9.2 show that the marginal posterior distribution of ψ has density given by

$$\omega\left(\psi_0 \mid s\right) = \int \pi(h^{-1}(\psi_0, \lambda) \mid s)|J(h^{-1}(\psi_0, \lambda))|^{-1}\, d\lambda, \tag{7.2.1}$$

where J denotes the Jacobian derivative of this transformation (see Problem 7.2.35). Evaluating (7.2.1) can be difficult, and we will generally avoid doing so here. An example illustrates how we can sometimes avoid directly implementing (7.2.1) and still obtain the marginal posterior distribution of ψ.

EXAMPLE 7.2.1 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown, and we use the prior given in Example 7.1.4. The posterior distribution for $\left(\mu, \sigma^2\right)$ is then given by (7.1.5) and (7.1.6).

Suppose we are primarily interested in $\psi(\mu, \sigma^2) = \sigma^2$. We see immediately that the marginal posterior of σ^2 is prescribed by (7.1.6) and thus have no further work to do, unless we want a form for the marginal posterior density of σ^2. We can use the methods of Section 2.6 for this (see Exercise 7.2.4).

If we want the marginal posterior distribution of $\psi(\mu, \sigma^2) = \mu$, then things are not quite so simple because (7.1.5) only prescribes the conditional posterior distribution of μ given σ^2. We can, however, avoid the necessity to implement (7.2.1). Note that (7.1.5) implies that

$$Z = \frac{\mu - \mu_x}{\left(n + 1/\tau_0^2\right)^{-1/2}\sigma} \mid \sigma^2, x_1, \ldots, x_n \sim N(0, 1),$$

where μ_x is given in (7.1.7). Because this distribution does not involve σ^2, the posterior distribution of Z is independent of the posterior distribution of σ. Now if $X \sim \text{Gamma}(\alpha, \beta)$, then $Y = 2\beta X \sim \text{Gamma}(\alpha, 1/2) = \chi^2(2\alpha)$ (see Problem 4.6.16 for the definition of the general chi-squared distribution) and so, from (7.1.6),

$$2\frac{\beta_x}{\sigma^2} \mid x_1, \ldots, x_n \sim \chi^2(2\alpha_0 + n),$$

where β_x is given in (7.1.8). Therefore (using Problem 4.6.14), as we are dividing an $N(0, 1)$ variable by the square root of an independent $\chi^2(2\alpha_0 + n)$ random variable divided by its degrees of freedom, we conclude that the posterior distribution of

$$T = \frac{Z}{\sqrt{\left(2\frac{\beta_x}{\sigma^2}\right) / (2\alpha_0 + n)}} = \frac{\mu - \mu_x}{\sqrt{\frac{2\beta_x}{(2\alpha_0+n)(n+1/\tau_0^2)}}}$$

is $t(2\alpha_0 + n)$. Equivalently, we can say the posterior distribution of μ is the same as

$$\mu_x + \sqrt{\frac{1}{2\alpha_0 + n}}\sqrt{\frac{2\beta_x}{n + 1/\tau_0^2}}\ T,$$

where $T \sim t(2\alpha_0 + n)$. By (7.1.9), (7.1.10), and (7.1.11), we have that the posterior distribution of μ converges to the distribution of

$$\bar{x} + \sqrt{\frac{n-1}{(2\alpha_0 + n)}}\frac{s}{\sqrt{n}}\ T$$

as $\tau_0 \to \infty$ and $\beta_0 \to 0$.

In other cases, we cannot avoid the use of (7.2.1) if we want the marginal posterior density of ψ. For example, suppose we are interested in the posterior distribution of the coefficient of variation (we exclude the line given by $\mu = 0$ from the parameter space)

$$\psi = \psi(\mu, \sigma^{-2}) = \frac{\sigma}{\mu} = \frac{1}{\mu}\left(\frac{1}{\sigma^2}\right)^{-1/2}.$$

Then a complementing function to ψ is given by

$$\lambda = \lambda(\mu, \sigma^{-2}) = \frac{1}{\sigma^2},$$

and it can be shown (see Section 7.5) that

$$J(\theta(\psi, \lambda)) = \psi^{-2}\lambda^{-1/2}.$$

If we let $\pi(\cdot \mid \lambda^{-1}, x_1, \ldots, x_n)$ and $\rho(\cdot \mid x_1, \ldots, x_n)$ denote the posterior densities of μ given λ, and the posterior density of λ, respectively, then, from (7.2.1), the marginal density of ψ is given by

$$\psi^{-2}\int_0^\infty \pi(\psi^{-1}\lambda^{-1/2} \mid \lambda^{-1}, x_1, \ldots, x_n)\rho(\lambda \mid x_1, \ldots, x_n)\ \lambda^{-1/2}\, d\lambda. \tag{7.2.2}$$

Without writing this out (see Problem 7.2.22), we note that we are left with a rather messy integral to evaluate. ∎

In some cases, integrals such as (7.2.2) can be evaluated in closed form; in other cases, they cannot. While it is convenient to have a closed form for a density, often this is not necessary, as we can use Monte Carlo methods to approximate posterior

probabilities and expectations of interest. We will return to this in Section 7.3. We should always remember that our goal, in implementing Bayesian inference methods, is not to find the marginal posterior densities of quantities of interest, but rather to have a computational algorithm that allows us to implement our inferences.

Under fairly weak conditions, it can be shown that the posterior distribution of θ converges, as the sample size increases, to a distribution degenerate at the true value. This is very satisfying, as it indicates that Bayesian inference methods are consistent.

7.2.1 Estimation

Suppose now that we want to calculate an estimate of a characteristic of interest $\psi(\theta)$. We base this on the posterior distribution of this quantity. There are several different approaches to this problem.

Perhaps the most natural estimate is to obtain the posterior density (or probability function when relevant) of ψ and use the *posterior mode* $\hat{\psi}$, i.e., the point where the posterior probability or density function of ψ takes its maximum. In the discrete case, this is the value of ψ with the greatest posterior probability; in the continuous case, it is the value that has the greatest amount of posterior probability in short intervals containing it.

To calculate the posterior mode, we need to maximize $\omega(\psi \mid s)$ as a function of ψ. Note that it is equivalent to maximize $m(s)\omega(\psi \mid s)$ so that we do not need to compute the inverse normalizing constant to implement this. In fact, we can conveniently choose to maximize any function that is a 1–1 increasing function of $\omega(\cdot \mid s)$ and get the same answer. In general, $\omega(\cdot \mid s)$ may not have a unique mode, but typically there is only one.

An alternative estimate is commonly used and has a natural interpretation. This is given by the posterior mean

$$E(\psi(\theta) \mid s),$$

whenever this exists. When the posterior distribution of ψ is symmetrical about its mode, and the expectation exists, then the posterior expectation is the same as the posterior mode; otherwise, these estimates will be different. If we want the estimate to reflect where the central mass of probability lies, then in cases where $\omega(\cdot \mid s)$ is highly skewed, perhaps the mode is a better choice than the mean. We will see in Chapter 8, however, that there are other ways of justifying the posterior mean as an estimate.

We now consider some examples.

EXAMPLE 7.2.2 *Bernoulli Model*

Suppose we observe a sample $(x_1, \ldots, x_n)$ from the Bernoulli(θ) distribution with $\theta \in [0, 1]$ unknown and we place a Beta(α, β) prior on θ. In Example 7.1.1, we determined the posterior distribution of θ to be Beta($n\bar{x} + \alpha, n(1 - \bar{x}) + \beta$). Let us suppose that the characteristic of interest is $\psi(\theta) = \theta$.

The posterior expectation of θ is given by

$$
\begin{aligned}
&E(\theta \mid x_1, \ldots, x_n) \\
&= \int_0^1 \theta \, \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \, \theta^{n\bar{x}+\alpha-1} \, (1-\theta)^{n(1-\bar{x})+\beta-1} \, d\theta \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \int_0^1 \theta^{n\bar{x}+\alpha} \, (1-\theta)^{n(1-\bar{x})+\beta-1} \, d\theta \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \frac{\Gamma(n\bar{x}+\alpha+1)\Gamma(n(1-\bar{x})+\beta)}{\Gamma(n+\alpha+\beta+1)} \\
&= \frac{n\bar{x}+\alpha}{n+\alpha+\beta}.
\end{aligned}
$$

When we have a uniform prior, i.e., $\alpha = \beta = 1$, the posterior expectation is given by

$$E(\theta \mid x) = \frac{n\bar{x}+1}{n+2}.$$

To determine the posterior mode, we need to maximize

$$\ln \theta^{n\bar{x}+\alpha-1} \, (1-\theta)^{n(1-\bar{x})+\beta-1} = (n\bar{x}+\alpha-1) \ln \theta + (n\,(1-\bar{x})+\beta-1) \ln (1-\theta) \, .$$

This function has first derivative

$$\frac{n\bar{x}+\alpha-1}{\theta} - \frac{n\,(1-\bar{x})+\beta-1}{1-\theta}$$

and second derivative

$$-\frac{n\bar{x}+\alpha-1}{\theta^2} - \frac{n\,(1-\bar{x})+\beta-1}{(1-\theta)^2}.$$

Setting the first derivative equal to 0 and solving gives the solution

$$\hat{\theta} = \frac{n\bar{x}+\alpha-1}{n+\alpha+\beta-2}.$$

Now, if $\alpha \geq 1, \beta \geq 1$, we see that the second derivative is always negative, and so $\hat{\theta}$ is the unique posterior mode. The restriction on the choice of $\alpha \geq 1, \beta \geq 1$ implies that the prior has a mode in (0, 1) rather than at 0 or 1. Note that when $\alpha = 1, \beta = 1$, namely, when we put a uniform prior on θ, the posterior mode is $\hat{\theta} = \bar{x}$. This is the same as the maximum likelihood estimate (MLE).

The posterior is highly skewed whenever $n\bar{x}+\alpha$ and $n\,(1-\bar{x})+\beta$ are far apart (plot Beta densities to see this). Thus, in such a case, we might consider the posterior mode as a more sensible estimate of θ. Note that when n is large, the mode and the mean will be very close together and in fact very close to the MLE $\bar{x}$. ∎

EXAMPLE 7.2.3 *Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known, and we take the prior distribution on μ to be $N(\mu, \tau_0^2)$. Let us suppose, that the characteristic of interest is $\psi(\mu) = \mu$.

In Example 7.1.2 we showed that the posterior distribution of μ is given by the

$$N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right),\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)$$

distribution. Because this distribution is symmetric about its mode, and the mean exists, the posterior mode and mean agree and equal

$$\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right).$$

This is a weighted average of the prior mean and the sample mean and lies between these two values.

When n is large, we see that this estimator is approximately equal to the sample mean $\bar{x}$, which we also know to be the MLE for this situation. Furthermore, when we take the prior to be very diffuse, namely, when τ_0^2 is very large, then again this estimator is close to the sample mean.

Also observe that the ratio of the sampling variance of $\bar{x}$ to the posterior variance of μ is

$$\frac{\sigma_0^2}{n}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)=1+\frac{\sigma_0^2}{n\tau_0^2},$$

is always greater than 1. The closer τ_0^2 is to 0, the larger this ratio is. Furthermore, as $\tau_0^2 \to 0$, the Bayesian estimate converges to μ_0.

If we are pretty confident that the population mean μ is close to the prior mean μ_0, we will take τ_0^2 small so that the bias in the Bayesian estimate will be small and its variance will be much smaller than the sampling variance of $\bar{x}$. In such a situation, the Bayesian estimator improves on accuracy over the sample mean. Of course, if we are not very confident that μ is close to the prior mean μ_0, then we choose a large value for τ_0^2, and the Bayesian estimator is basically the MLE. ∎

EXAMPLE 7.2.4 *Multinomial Model*

Suppose we have a sample $(s_1, \ldots, s_n)$ from the model discussed in Example 7.1.3 and we place a Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$ distribution on $(\theta_1, \ldots, \theta_{k-1})$. The posterior distribution of $(\theta_1, \ldots, \theta_{k-1})$ is then

$$\text{Dirichlet}(x_1+\alpha_1, x_2+\alpha_2, \ldots, x_k+\alpha_k),$$

where x_i is the number of responses in the ith category.

Now suppose we are interested in estimating $\psi(\theta) = \theta_1$, the probability that a response is in the first category. It can be shown (see Problem 7.2.25) that, if $(\theta_1, \ldots, \theta_{k-1})$ is distributed Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$, then θ_i is distributed

$$\text{Dirichlet}(\alpha_i, \alpha_{-i}) = \text{Beta}(\alpha_i, \alpha_{-i})$$

where $\alpha_{-i} = \alpha_1 + \alpha_2 + \cdots + \alpha_k - \alpha_i$. This result implies that the marginal posterior distribution of θ_1 is

$$\text{Beta}(x_1+\alpha_1, x_2+\cdots+x_k+\alpha_2+\cdots+\alpha_k).$$

Then, assuming that each $\alpha_i \geq 1$, and using the argument in Example 7.2.2 and $x_1 + \cdots + x_k = n$, the marginal posterior mode of θ_1 is

$$\hat{\theta}_1 = \frac{x_1 + \alpha_1 - 1}{n - 2 + \alpha_1 + \cdots + \alpha_k}.$$

When the prior is the uniform, namely, $\alpha_1 = \cdots = \alpha_k = 1$, then

$$\hat{\theta}_1 = \frac{x_1}{n + k - 2}.$$

As in Example 7.2.2, we compute the posterior expectation to be

$$E(\theta_1 \mid x) = \frac{x_1 + \alpha_1}{n + \alpha_1 + \cdots + \alpha_k}.$$

The posterior distribution is highly skewed whenever $x_1 + \alpha_1$ and $x_2 + \cdots + x_k + \alpha_2 + \cdots + \alpha_k$ are far apart.

From Problem 7.2.26, we have that the plug-in MLE of θ_1 is x_1/n. When n is large, the Bayesian estimates are close to this value, so there is no conflict between the estimates. Notice, however, that when the prior is uniform, then $\alpha_1 + \cdots + \alpha_k = k$, hence the plug-in MLE and the Bayesian estimates will be quite different when k is large relative to n. In fact, the posterior mode will always be smaller than the plug-in MLE when $k > 2$ and $x_1 > 0$. This is a situation in which the Bayesian and frequentist approaches to inference differ.

At this point, the decision about which estimate to use is left with the practitioner, as theory does not seem to provide a clear answer. We can be comforted by the fact that the estimates will not differ by much in many contexts of practical importance. ∎

EXAMPLE 7.2.5 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown, and we use the prior given in Example 7.1.4. Let us suppose that the characteristic of interest is $\psi(\mu, \sigma^2) = \mu$.

In Example 7.2.1, we derived the marginal posterior distribution of μ to be the same as the distribution of

$$\mu_x + \sqrt{\frac{1}{2\alpha_0 + n}} \sqrt{\frac{2\beta_x}{n + 1/\tau_0^2}}\, T,$$

where $T \sim t(n + 2\alpha_0)$. This is a $t\,(n + 2\alpha_0)$ distribution relocated to have its mode at μ_x and rescaled by the factor

$$\sqrt{\frac{1}{2\alpha_0 + n}} \sqrt{\frac{2\beta_x}{n + 1/\tau_0^2}}.$$

So the marginal posterior mode of μ is

$$\mu_x = \left(n + \frac{1}{\tau_0^2}\right)^{-1} \left(\frac{\mu_0}{\tau_0^2} + n\bar{x}\right).$$

Because a t distribution is symmetric about its mode, this is also the posterior mean of μ, provided that $n + 2\alpha_0 > 1$, as a $t(\lambda)$ distribution has a mean only when $\lambda > 1$ (see Problem 4.6.16). This will always be the case as the sample size $n \geq 1$. Again, μ_x is a weighted average of the prior mean μ_0 and the sample average $\bar{x}$.

The marginal posterior mode and expectation can also be obtained for $\psi(\mu, \sigma^2) = \sigma^2$. These computations are left to the reader (see Exercise 7.2.4). ∎

One issue that we have not yet addressed is how we will assess the accuracy of Bayesian estimates. Naturally, this is based on the posterior distribution and how concentrated it is about the estimate being used. In the case of the posterior mean, this means that we compute the posterior variance as a measure of spread for the posterior distribution of ψ about its mean. For the posterior mode, we will discuss this issue further in Section 7.2.3.

EXAMPLE 7.2.6 *Posterior Variances*
In Example 7.2.2, the posterior variance of θ is given by (see Exercise 7.2.6)

$$\frac{(n\bar{x} + \alpha)\,(n\,(1 - \bar{x}) + \beta)}{(n + \alpha + \beta)^2\,(n + \alpha + \beta + 1)}.$$

Notice that the posterior variance converges to 0 as $n \to \infty$.

In Example 7.2.3, the posterior variance is given by $(1/\tau_0^2 + n/\sigma_0^2)^{-1}$. Notice that the posterior variance converges to 0 as $\tau_0^2 \to 0$ and converges to σ_0^2/n, the sampling variance of $\bar{x}$, as $\tau_0^2 \to \infty$.

In Example 7.2.4, the posterior variance of θ_1 is given by (see Exercise 7.2.7)

$$\frac{(x_1 + \alpha_1)\,(x_2 + \cdots + x_k + \alpha_2 + \cdots + \alpha_k)}{(n + \alpha_1 + \cdots + \alpha_k)^2\,(n + \alpha_1 + \cdots + \alpha_k + 1)}.$$

Notice that the posterior variance converges to 0 as $n \to \infty$.

In Example 7.2.5, the posterior variance of μ is given by (see Problem 7.2.28)

$$\left(\frac{1}{n + 2\alpha_0}\right)\left(\frac{2\beta_x}{n + 1/\tau_0^2}\right)\left(\frac{n + 2\alpha_0}{n + 2\alpha_0 - 2}\right) = \left(\frac{2\beta_x}{n + 1/\tau_0^2}\right)\left(\frac{1}{n + 2\alpha_0 - 2}\right),$$

provided $n + 2\alpha_0 > 2$, because the variance of a $t(\lambda)$ distribution is $\lambda/(\lambda - 2)$ when $\lambda > 2$ (see Problem 4.6.16). Notice that the posterior variance goes to 0 as $n \to \infty$. ∎

7.2.2 Credible Intervals

A *credible interval*, for a real-valued parameter $\psi(\theta)$, is an interval $C(s) = [l(s), u(s)]$ that we believe will contain the true value of ψ. As with the sampling theory approach, we specify a probability γ and then find an interval $C(s)$ satisfying

$$\Pi(\psi(\theta) \in C(s) \mid s) = \Pi\left(\{\theta : l(s) \leq \psi(\theta) \leq u(s)\} \mid s\right) \geq \gamma\,. \tag{7.2.3}$$

We then refer to $C(s)$ as a γ-credible interval for ψ.

Naturally, we try to find a γ-credible interval $C(s)$ so that $\Pi(\psi(\theta) \in C(s) \mid s)$ is as close to γ as possible, and such that $C(s)$ is as short as possible. This leads to the consideration of *highest posterior density (HPD) intervals*, which are of the form

$$C(s) = \{\psi : \omega(\psi \mid s) \geq c\},$$

where $\omega(\cdot \mid s)$ is the marginal posterior density of ψ and where c is chosen as large as possible so that (7.2.3) is satisfied. In Figure 7.2.1, we have plotted an example of an HPD interval for a given value of c.

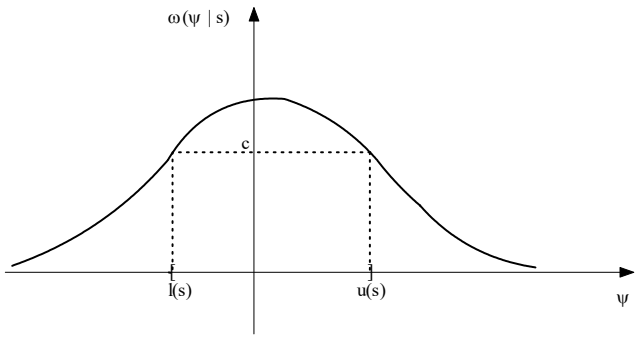

Figure 7.2.1: An HPD interval $C(s) = [l(s), u(s)] = \{\psi : \omega(\psi \mid s) \geq c\}$.

Clearly, $C(s)$ contains the mode whenever $c \leq \max_{\psi} \omega(\psi \mid s)$. We can take the length of an HPD interval as a measure of the accuracy of the mode of $\omega(\cdot \mid s)$ as an estimator of $\psi(\theta)$. The length of a 0.95-credible interval for ψ will serve the same purpose as the margin of error does with confidence intervals.

Consider now some applications of the concept of credible interval.

EXAMPLE 7.2.7 *Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known, and we take the prior distribution on μ to be $N(\mu_0, \tau_0^2)$. In Example 7.1.2, we showed that the posterior distribution of μ is given by the

$$N\left(\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma_0^2}\bar{x}\right), \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}\right)$$

distribution. Since this distribution is symmetric about its mode (also mean) $\hat{\mu}$, a shortest γ-HPD interval is of the form

$$\hat{\mu} \pm \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1/2} c,$$

where c is such that

$$\begin{aligned}\gamma &= \Pi\left(\mu \in \left[\hat{\mu} \pm \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1/2} c\right] \,\middle|\, x_1,\ldots,x_n\right)\\ &= \Pi\left(-c \le \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{1/2}(\mu-\hat{\mu}) \le c \,\middle|\, x_1,\ldots,x_n\right).\end{aligned}$$

Since

$$\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{1/2}(\mu-\hat{\mu}) \mid x_1,\ldots,x_n \sim N(0,1),$$

we have $\gamma = \Phi(c) - \Phi(-c)$, where Φ is the standard normal cumulative distribution function (cdf). This immediately implies that $c = z_{(1+\gamma)/2}$ and the γ-HPD interval is given by

$$\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right) \pm \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1/2} z_{(1+\gamma)/2}.$$

Note that as $\tau_0^2 \to \infty$, namely, as the prior becomes increasingly diffuse, this interval converges to the interval

$$\bar{x} \pm \frac{\sigma_0}{\sqrt{n}} z_{(1+\gamma)/2},$$

which is also the γ-confidence interval derived in Chapter 6 for this problem. So under a diffuse normal prior, the Bayesian and frequentist approaches agree. ∎

EXAMPLE 7.2.8 *Location-Scale Normal Model*
Suppose that $(x_1,\ldots,x_n)$ is a sample from an $N(\mu,\sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma \ge 0$ are unknown, and we use the prior given in Example 7.1.4. In Example 7.2.1, we derived the marginal posterior distribution of μ to be the same as

$$\mu_x + \sqrt{\frac{1}{2\alpha_0+n}}\sqrt{\frac{2\beta_x}{n+1/\tau_0^2}}\, T,$$

where $T \sim t(2\alpha_0+n)$. Because this distribution is symmetric about its mode μ_x, a γ-HPD interval is of the form

$$\mu_x \pm \sqrt{\frac{1}{2\alpha_0+n}}\sqrt{\frac{2\beta_x}{n+1/\tau_0^2}}\, c,$$

where c satisfies

$$\begin{aligned}\gamma &= \Pi\left(\mu \in \left[\mu_x \pm \sqrt{\frac{1}{2\alpha_0+n}}\sqrt{\frac{2\beta_x}{n+1/\tau_0^2}}\,c\right] \,\middle|\, x_1,\ldots,x_n\right)\\ &= \Pi\left(-c \le \left(\frac{2\beta_x}{(2\alpha_0+n)\left(n+1/\tau_0^2\right)}\right)^{-1/2}\left(\mu-\hat{\mu}\right) \le c \,\middle|\, x_1,\ldots,x_n\right)\\ &= G_{2\alpha_0+n}(c) - G_{2\alpha_0+n}(-c).\end{aligned}$$

Here, $G_{2\alpha_0+n}$ is the $t\,(2\alpha_0+n)$ cdf, and therefore $c = t_{(1+\gamma)/2}(2\alpha_0+n)$.

Using (7.1.9), (7.1.10), and (7.1.11) we have that this interval converges to the interval

$$\bar{x} \pm \sqrt{\frac{n-1}{(2\alpha_0+n)}}\frac{s}{\sqrt{n}}\,t\,(n+2\alpha_0)$$

as $\tau_0 \to \infty$ and $\beta_0 \to 0$. Note that this is a little different from the γ-confidence interval we obtained for μ in Example 6.3.8, but when α_0/n is small, they are virtually identical. ∎

In the examples we have considered so far, we could obtain closed-form expressions for the HPD intervals. In general, this is not the case. In such situations, we have to resort to numerical methods to obtain the HPD intervals, but we do not pursue this topic further here.

There are other methods of deriving credible intervals. For example, a common method of obtaining a γ-credible interval for ψ is to take the interval $[\psi_l, \psi_r]$ where ψ_l is a $(1-\gamma)/2$ quantile for the posterior distribution of ψ and ψ_r is a $1-(1-\gamma)/2$ quantile for this distribution. Alternatively, we could form one-sided intervals. These credible intervals avoid the more extensive computations that may be needed for HPD intervals.

7.2.3 Hypothesis Testing and Bayes Factors

Suppose now that we want to assess the evidence in the observed data concerning the hypothesis $H_0 : \psi(\theta) = \psi_0$. It seems clear how we should assess this, namely, compute the posterior probability

$$\Pi(\psi(\theta) = \psi_0 \,|\, s). \tag{7.2.4}$$

If this is small, then conclude that we have evidence against H_0. We will see further justification for this approach in Chapter 8.

EXAMPLE 7.2.9

Suppose we want to assess the evidence concerning whether or not $\theta \in A$. If we let $\psi = I_A$, then we are assessing the hypothesis $H_0 : \psi(\theta) = 1$ and

$$\Pi(\psi(\theta) = 1 \,|\, s) = \Pi(A \,|\, s).$$

So in this case, we simply compute the posterior probability that $\theta \in A$. ∎

There can be a problem, however, with using (7.2.4) to assess a hypothesis. For when the prior distribution of ψ is absolutely continuous, then $\Pi(\psi(\theta) = \psi_0 \,|\, s) = 0$ for all data s. Therefore, we would always find evidence against H_0 no matter what is observed, which does not make sense. In general, if the value ψ_0 is assigned small prior probability, then it can happen that this value also has a small posterior probability no matter what data are observed.

To avoid this problem, there is an alternative approach to hypothesis assessment that is sometimes used. Recall that, if ψ_0 is a surprising value for the posterior distribution of ψ, then this is evidence that H_0 is false. The value ψ_0 is surprising whenever it occurs in a region of low probability for the posterior distribution of ψ. A region of low probability will correspond to a region where the posterior density $\omega(\cdot \,|\, s)$ is relatively low. So, one possible method for assessing this is by computing the *(Bayesian) P-value*

$$\Pi(\{\theta : \omega(\psi(\theta) \,|\, s) \leq \omega(\psi_0 \,|\, s)\} \,|\, s). \tag{7.2.5}$$

Note that when $\omega(\cdot \,|\, s)$ is unimodal, (7.2.5) corresponds to computing a tail probability. If the probability (7.2.5) is small, then ψ_0 is surprising, at least with respect to our posterior beliefs. When we decide to reject H_0 whenever the P-value is less than $1 - \gamma$, then this approach is equivalent to computing a γ-HPD region for ψ and rejecting H_0 whenever ψ_0 is not in the region.

EXAMPLE 7.2.10 *(Example 7.2.9 continued)*
Applying the P-value approach to this problem, we see that $\psi(\theta) = I_A(\theta)$ has posterior given by the Bernoulli$(\Pi(A \,|\, s))$ distribution. Therefore, $\omega(\cdot \,|\, s)$ is defined by $\omega(0 \,|\, s) = 1 - \Pi(A \,|\, s) = \Pi(A^c \,|\, s)$ and $\omega(1 \,|\, s) = \Pi(A \,|\, s)$.

Now $\psi_0 = 1$, so

$$\begin{aligned} \{\theta : \omega(\psi(\theta) \,|\, s) \leq \omega(1 \,|\, s)\} &= \{\theta : \omega(I_A(\theta) \,|\, s) \leq \Pi(A \,|\, s)\} \\ &= \begin{cases} \Omega & \Pi(A \,|\, s) \geq \Pi(A^c \,|\, s) \\ A & \Pi(A \,|\, s) < \Pi(A^c \,|\, s). \end{cases} \end{aligned}$$

Therefore, (7.2.5) becomes

$$\Pi(\{\theta : \omega(\psi(\theta) \,|\, s) \leq \omega(1 \,|\, s)\} \,|\, s) = \begin{cases} 1 & \Pi(A \,|\, s) \geq \Pi(A^c \,|\, s) \\ \Pi(A \,|\, s) & \Pi(A \,|\, s) < \Pi(A^c \,|\, s), \end{cases}$$

so again we have evidence against H_0 whenever $\Pi(A \,|\, s)$ is small. ∎

We see from Examples 7.2.9 and 7.2.10 that computing the P-value (7.2.5) is essentially equivalent to using (7.2.4), whenever the marginal parameter ψ takes only two values. This is not the case whenever ψ takes more than two values, however, and the statistician has to decide which method is more appropriate in such a context.

As previously noted, when the prior distribution of ψ is absolutely continuous, then (7.2.4) is always 0, no matter what data are observed. As the following example illustrates, there is also a difficulty with using (7.2.5) in such a situation.

EXAMPLE 7.2.11
Suppose that the posterior distribution of θ is Beta(2, 1), i.e., $\omega(\theta \,|\, s) = 2\theta$ when $0 \leq \theta \leq 1$, and we want to assess $H_0 : \theta = 3/4$. Then $\omega(\theta \,|\, s) \leq \omega(3/4 \,|\, s)$ if and

only if $\theta \leq 3/4$, and (7.2.5) is given by

$$\int_0^{3/4} 2\theta \, d\theta = 9/16.$$

On the other hand, suppose we make a 1–1 transformation to $\rho = \theta^2$ so that the hypothesis is now $H_0 : \rho = 9/16$. The posterior distribution of ρ is Beta(1, 1). Since the posterior density of ρ is constant, this implies that the posterior density at every possible value is less than or equal to the posterior density evaluated at $9/16$. Therefore, (7.2.5) equals 1, and we would never find evidence against H_0 using this parameterization.

This example shows that our assessment of H_0 via (7.2.5) depends on the parameterization used, which does not seem appropriate. ∎

The difficulty in using (7.2.5), as demonstrated in Example 7.2.11, only occurs with continuous posterior distributions. So, to avoid this problem, it is often recommended that the hypothesis to be tested always be assigned a positive prior probability. As demonstrated in Example 7.2.10, the approach via (7.2.5) is then essentially equivalent to using (7.2.4) to assess H_0.

In problems where it seems natural to use continuous priors, this is accomplished by taking the prior Π to be a mixture of probability distributions, as discussed in Section 2.5.4, namely, the prior distribution equals

$$\Pi = p\Pi_1 + (1 - p)\,\Pi_2,$$

where $\Pi_1(\psi(\theta) = \psi_0) = 1$ and $\Pi_2(\psi\,(\theta) = \psi_0) = 0$, i.e., Π_1 is degenerate at ψ_0 and Π_2 is continuous at ψ_0. Then

$$\Pi(\psi(\theta) = \psi_0) = p\Pi_1(\psi(\theta) = \psi_0) + (1 - p)\,\Pi_2(\psi(\theta) = \psi_0) = p > 0$$

is the prior probability that H_0 is true.

The prior predictive for the data s is then given by

$$m(s) = pm_1(s) + (1 - p)m_2(s),$$

where m_i is the prior predictive obtained via prior Π_i (see Problem 7.2.34). This implies (see Problem 7.2.34) that the posterior probability measure for θ, when using the prior Π, is

$$\Pi\,(A \mid s) = \frac{pm_1(s)}{pm_1(s) + (1 - p)m_2(s)}\Pi_1(A \mid s) + \frac{(1 - p)\,m_2(s)}{pm_1(s) + (1 - p)m_2(s)}\Pi_2(A \mid s) \quad (7.2.6)$$

where $\Pi_i(\cdot \mid s)$ is the posterior measure obtained via the prior Π_i. Note that this a mixture of the posterior probability measures $\Pi_1(\cdot \mid s)$ and $\Pi_2(\cdot \mid s)$ with mixture probabilities

$$\frac{pm_1(s)}{pm_1(s) + (1 - p)m_2(s)} \text{ and } \frac{(1 - p)\,m_2(s)}{pm_1(s) + (1 - p)m_2(s)}.$$

Now $\Pi_1(\cdot \mid s)$ is degenerate at ψ_0 (if the prior is degenerate at a point then the posterior must be degenerate at that point too) and $\Pi_2(\cdot \mid s)$ is continuous at ψ_0. Therefore,

$$\Pi(\psi(\theta) = \psi_0 \mid s) = \frac{pm_1(s)}{pm_1(s) + (1-p)m_2(s)}, \tag{7.2.7}$$

and we use this probability to assess H_0.

The following example illustrates this approach.

EXAMPLE 7.2.12 *Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known, and we want to assess the hypothesis $H_0 : \mu = \mu_0$. As in Example 7.1.2, we will take the prior for μ to be an $N(\mu_0, \tau_0^2)$ distribution. Given that we are assessing whether or not $\mu = \mu_0$, it seems reasonable to place the mode of the prior at the hypothesized value. The choice of the hyperparameter τ_0^2 then reflects the degree of our prior belief that H_0 is true. We let Π_2 denote this prior probability measure, i.e., Π_2 is the $N(\mu_0, \sigma_0^2)$ probability measure.

If we use Π_2 as our prior, then, as shown in Example 7.1.2, the posterior distribution of μ is absolutely continuous. This implies that (7.2.4) is 0. So, following the preceding discussion, we consider instead the prior $\Pi = p\Pi_1 + (1-p)\Pi_2$ obtained by mixing Π_2 with a probability measure Π_1 degenerate at μ_0. Then $\Pi_1(\{\mu_0\}) = 1$ and so $\Pi(\{\mu_0\}) = p$. As shown in Example 7.1.2, under Π_2 the posterior distribution of μ is

$$N\left(\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma_0^2}\bar{x}\right), \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}\right),$$

while the posterior under Π_1 is the distribution degenerate at μ_0. We now need to evaluate (7.2.7), and we will do this in Example 7.2.13. ∎

Bayes Factors

Bayes factors comprise another method of hypothesis assessment and are defined in terms of odds.

Definition 7.2.1 In a probability model with sample space S and probability measure P, the *odds in favor* of event $A \subset S$ is defined to be $P(A)/P(A^c)$, namely, the ratio of the probability of A to the probability of A^c.

Obviously, large values of the odds in favor of A indicate a strong belief that A is true. Odds represent another way of presenting probabilities that are convenient in certain contexts, e.g., horse racing. Bayes factors compare posterior odds with prior odds.

Definition 7.2.2 The *Bayes factor* BF_{H_0} *in favor of the hypothesis* $H_0 : \psi(\theta) = \psi_0$ is defined, whenever the prior probability of H_0 is not 0 or 1, to be the ratio of the *posterior odds* in favor of H_0 to the *prior odds* in favor of H_0, or

$$BF_{H_0} = \left\{\frac{\Pi(\psi(\theta) = \psi_0 \mid s)}{1 - \Pi(\psi(\theta) = \psi_0 \mid s)}\right\} \Big/ \left\{\frac{\Pi(\psi(\theta) = \psi_0)}{1 - \Pi(\psi(\theta) = \psi_0)}\right\}. \tag{7.2.8}$$

So the Bayes factor in favor of H_0 is measuring the degree to which the data have changed the odds in favor of the hypothesis. If BF_{H_0} is small, then the data are providing evidence against H_0 and evidence in favor of H_0 when BF_{H_0} is large.

There is a relationship between the posterior probability of H_0 being true and BF_{H_0}. From (7.2.8), we obtain

$$\Pi(\psi(\theta) = \psi_0 \,|\, s) = \frac{r BF_{H_0}}{1 + r BF_{H_0}}, \tag{7.2.9}$$

where

$$r = \frac{\Pi(\psi(\theta) = \psi_0)}{1 - \Pi(\psi(\theta) = \psi_0)}$$

is the prior odds in favor of H_0. So, when BF_{H_0} is small, then $\Pi(\psi(\theta) = \psi_0 | s)$ is small and conversely.

One reason for using Bayes factors to assess hypotheses is the following result. This establishes a connection with likelihood ratios.

Theorem 7.2.1 If the prior Π is a mixture $\Pi = p\Pi_1 + (1-p)\,\Pi_2$, where $\Pi_1(A) = 1$, $\Pi_2(A^C) = 1$, and we want to assess the hypothesis $H_0 : \theta \in A$, then

$$BF_{H_0} = m_1(s)/m_2(s),$$

where m_i is the prior predictive of the data under Π_i.

PROOF Recall that, if a prior concentrates all of its probability on a set, then the posterior concentrates all of its probability on this set, too. Then using (7.2.6), we have

$$BF_{H_0} = \frac{\Pi(A \,|\, s)}{1 - \Pi(A \,|\, s)} \Big/ \frac{\Pi(A)}{1 - \Pi(A)} = \frac{p m_1(s)}{(1-p)\, m_2(s)} \Big/ \frac{p}{1-p} = \frac{m_1(s)}{m_2(s)}. \blacksquare$$

Interestingly, Theorem 7.2.1 indicates that the Bayes factor is independent of p. We note, however, that it is not immediately clear how to interpret the value of BF_{H_0}. In particular, how large does BF_{H_0} have to be to provide strong evidence in favor of H_0? One approach to this problem is to use (7.2.9), as this gives the posterior probability of H_0, which is directly interpretable. So we can calibrate the Bayes factor. Note, however, that this requires the specification of p.

EXAMPLE 7.2.13 *Location Normal Model (Example 7.2.12 continued)*
We now compute the prior predictive under Π_2. We have that the joint density of $(x_1, \ldots, x_n)$ given μ equals

$$(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{n-1}{2\sigma_0^2} s^2\right) \exp\left(-\frac{n}{2\sigma_0^2} (\bar{x} - \mu)^2\right)$$

and so

$$
\begin{aligned}
&m_2(x_1,\ldots,x_n)\\
&=\int_{-\infty}^{\infty}\left\{\begin{array}{c}\left(2\pi\sigma_0^2\right)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu\right)^2\right)\\ \times\left(2\pi\tau_0^2\right)^{-1/2}\exp\left(-\frac{1}{2\tau_0^2}\left(\mu-\mu_0\right)^2\right)\end{array}\right\}d\mu\\
&=(2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\\
&\quad\times\tau_0^{-1}(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu\right)^2\right)\exp\left(-\frac{1}{2\tau_0^2}\left(\mu-\mu_0\right)^2\right)d\mu.
\end{aligned}
$$

Then using (7.1.2), we have

$$
\begin{aligned}
&\tau_0^{-1}(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu\right)^2\right)\exp\left(-\frac{1}{2\tau_0^2}\left(\mu-\mu_0\right)^2\right)d\mu\\
&=\tau_0^{-1}\exp\left(\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)^2\right)\\
&\quad\times\exp\left(-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{n\bar{x}^2}{\sigma_0^2}\right)\right)\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}\right)^{-1/2}. \qquad (7.2.10)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&m_2(x_1,\ldots,x_n)\\
&=(2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\tau_0^{-1}\exp\left(\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{x}\right)^2\right)\\
&\quad\times\exp\left(-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{n\bar{x}^2}{\sigma_0^2}\right)\right)\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}\right)^{-1/2}.
\end{aligned}
$$

Because Π_1 is degenerate at μ_0, it is immediate that the prior predictive under Π_1 is given by

$$
m_1(x_1,\ldots,x_n)=(2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right)\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu_0\right)^2\right).
$$

Therefore, BF_{H_0} equals

$$
\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu_0\right)^2\right)
$$

divided by (7.2.10).

For example, suppose that $\mu_0 = 0$, $\tau_0^2 = 2$, $\sigma_0^2 = 1$, $n = 10$, and $\bar{x} = 0.2$. Then

$$\exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}-\mu_0\right)^2\right) = \exp\left(-\frac{10}{2}(0.2)^2\right) = 0.81873,$$

while (7.2.10) equals

$$\frac{1}{\sqrt{2}}\exp\left(\frac{1}{2}\left(\frac{1}{2}+10\right)^{-1}(10\,(0.2))^2\right)\exp\left(-\frac{10\,(0.2)^2}{2}\right)\left(10+\frac{1}{2}\right)^{-1/2}$$
$$= 0.21615.$$

So

$$BF_{H_0} = \frac{0.81873}{0.21615} = 3.7878,$$

which gives some evidence in favor of $H_0 : \mu = \mu_0$. If we suppose that $p = 1/2$, so that we are completely indifferent between H_0 being true and not being true, then $r = 1$, and (7.2.9) gives

$$\Pi(\mu = 0 \mid x_1, \ldots, x_n) = \frac{3.7878}{1+3.7878} = 0.79114,$$

indicating a large degree of support for H_0. ∎

7.2.4 Prediction

Prediction problems arise when we have an unobserved response value t in a sample space T and observed response $s \in S$. Furthermore, we have the statistical model $\{P_\theta : \theta \in \Omega\}$ for s and the conditional statistical model $\{Q_\theta(\cdot \mid s) : \theta \in \Omega\}$ for t given s. We assume that both models have the same true value of $\theta \in \Omega$. The objective is to construct a *prediction* $\tilde{t}(s) \in T$, of the unobserved value t, based on the observed data s. The value of t could be unknown simply because it represents a future outcome.

If we denote the conditional density or probability function (whichever is relevant) of t by $q_\theta(\cdot \mid s)$, the joint distribution of (θ, s, t) is given by

$$q_\theta(t \mid s) f_\theta(s) \pi(\theta).$$

Then, once we have observed s (assume here that the distributions of θ and t are absolutely continuous; if not, we replace integrals by sums), the conditional density of (t, θ), given s, is

$$\frac{q_\theta(t \mid s) f_\theta(s) \pi(\theta)}{\int_\Omega \int_T q_\theta(t \mid s) f_\theta(s) \pi(\theta)\, dt\, d\theta} = \frac{q_\theta(t \mid s) f_\theta(s) \pi(\theta)}{\int_\Omega f_\theta(s) \pi(\theta)\, d\theta} = \frac{q_\theta(t \mid s) f_\theta(s) \pi(\theta)}{m(s)}.$$

Then the marginal posterior distribution of t, known as the *posterior predictive* of t, is

$$q(t \mid s) = \int_\Omega \frac{q_\theta(t \mid s) f_\theta(s) \pi(\theta)}{m(s)}\, d\theta = \int_\Omega q_\theta(t \mid s) \pi(\theta \mid s)\, d\theta.$$

Notice that the posterior predictive of t is obtained by averaging the conditional density of t, given (θ, s), with respect to the posterior distribution of θ.

Now that we have obtained the posterior predictive distribution of t, we can use it to select an estimate of the unobserved value. Again, we could choose the posterior mode $\hat{t}$ or the posterior expectation $E(t \mid x) = \int_T t q(t \mid s)\, dt$ as our prediction, whichever is deemed most relevant.

EXAMPLE 7.2.14 *Bernoulli Model*
Suppose we want to predict the next independent outcome X_{n+1}, having observed a sample $(x_1, \ldots, x_n)$ from the Bernoulli(θ) and $\theta \sim \text{Beta}(\alpha, \beta)$. Here, the future observation is independent of the observed data. The posterior predictive probability function of X_{n+1} at t is then given by

$$
\begin{aligned}
& q(t \mid x_1, \ldots, x_n) \\
&= \int_0^1 \theta^t (1-\theta)^{1-t} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \theta^{n\bar{x}+\alpha-1} (1-\theta)^{n(1-\bar{x})+\beta-1}\, d\theta \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \int_0^1 \theta^{n\bar{x}+\alpha+t-1} (1-\theta)^{n(1-\bar{x})+\beta+(1-t)-1}\, d\theta \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\bar{x}+\alpha)\Gamma(n(1-\bar{x})+\beta)} \frac{\Gamma(n\bar{x}+\alpha+t)\Gamma(n(1-\bar{x})+\beta+1-t)}{\Gamma(n+\alpha+\beta+1)} \\
&= \begin{cases} \frac{n\bar{x}+\alpha}{n+\alpha+\beta} & t=1 \\ \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta} & t=0, \end{cases}
\end{aligned}
$$

which is the probability function of a Bernoulli($(n\bar{x}+\alpha)\,/\,(n+\alpha+\beta)$) distribution.

Using the posterior mode as the predictor, i.e., maximizing $q(t \mid x_1, \ldots, x_n)$ for t, leads to the prediction

$$
\hat{t} = \begin{cases} 1 & \text{if } \frac{n\bar{x}+\alpha}{n+\alpha+\beta} \geq \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta}, \\ 0 & \text{otherwise.} \end{cases}
$$

The posterior expectation predictor is given by

$$
E(t \mid x_1, \ldots, x_n) = \frac{n\bar{x}+\alpha}{n+\alpha+\beta}.
$$

Note that the posterior mode takes a value in $\{0, 1\}$, and the future X_{n+1} will be in this set, too. The posterior mean can be any value in $[0, 1]$. ∎

EXAMPLE 7.2.15 *Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known, and we use the prior given in Example 7.1.2. Suppose we want to predict a future observation X_{n+1}, but this time X_{n+1} is from the

$$
N\left(\bar{x}, \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1} \sigma_0^2\right) \tag{7.2.11}
$$

distribution. So, in this case, the future observation is not independent of the observed data, but it is independent of the parameter. A simple calculation (see Exercise 7.2.9) shows that (7.2.11) is the posterior predictive distribution of t and so we would predict t by $\bar{x}$, as this is both the posterior mode and mean. ∎

We can also construct a *γ-prediction region* $C(s)$ for a future value t from the model $\{q_\theta(\cdot \mid s) : \theta \in \Omega\}$. A γ-prediction region for t satisfies $Q(C(s) \mid s) \geq \gamma)$, where $Q(\cdot \mid s)$ is the posterior predictive measure for t. One approach to constructing $C(s)$ is to apply the HPD concept to $q(t \mid s)$. We illustrate this via several examples.

EXAMPLE 7.2.16 *Bernoulli Model (Example 7.2.14 continued)*
Suppose we want a γ-prediction region for a future value X_{n+1}. In Example 7.2.14, we derived the posterior predictive distribution of X_{n+1} to be

$$\text{Bernoulli}\left(\frac{n\bar{x}+\alpha}{n+\alpha+\beta}\right).$$

Accordingly, a γ-prediction region for t, derived via the HPD concept, is given by

$$C(x_1,\ldots,x_n) = \begin{cases} \{0,1\} & \text{if } \max\left\{\frac{n\bar{x}+\alpha}{n+\alpha+\beta}, \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta}\right\} \leq \gamma, \\ \{1\} & \text{if } \gamma \leq \max\left\{\frac{n\bar{x}+\alpha}{n+\alpha+\beta}, \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta}\right\} = \frac{n\bar{x}+\alpha}{n+\alpha+\beta}, \\ \{0\} & \text{if } \gamma \leq \max\left\{\frac{n\bar{x}+\alpha}{n+\alpha+\beta}, \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta}\right\} = \frac{n(1-\bar{x})+\beta}{n+\alpha+\beta}. \end{cases}$$

We see that this predictive region contains just the mode or encompasses all possible values for X_{n+1}. In the latter case, this is not an informative inference. ∎

EXAMPLE 7.2.17 *Location Normal Model (Example 7.2.15 continued)*
Suppose we want a γ-prediction interval for a future observation X_{n+1} from a

$$N\left(\bar{x}, \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\sigma_0^2\right)$$

distribution. As this is also the posterior predictive distribution of X_{n+1} and is symmetric about $\bar{x}$, a γ-prediction interval for X_{n+1}, derived via the HPD concept, is given by

$$\bar{x} \pm \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1/2} \sigma_0 z_{(1+\gamma)/2}.$$ ∎

Summary of Section 7.2

- Based on the posterior distribution of a parameter, we can obtain estimates of the parameter (posterior modes or means), construct credible intervals for the parameter (HPD intervals), and assess hypotheses about the parameter (posterior probability of the hypothesis, Bayesian P-values, Bayes factors).

- A new type of inference was discussed in this section, namely, prediction problems where we are concerned with predicting an unobserved value from a sampling model.

EXERCISES

7.2.1 For the model discussed in Example 7.1.1, derive the posterior mean of $\psi = \theta^m$ where $m > 0$.

7.2.2 For the model discussed in Example 7.1.2, determine the posterior distribution of the third quartile $\psi = \mu + \sigma_0 z_{0.75}$. Determine the posterior mode and the posterior expectation of ψ.

7.2.3 In Example 7.2.1, determine the posterior expectation and mode of $1/\sigma^2$.

7.2.4 In Example 7.2.1, determine the posterior expectation and mode of σ^2. (Hint: You will need the posterior density of σ^2 to determine the mode.)

7.2.5 Carry out the calculations to verify the posterior mode and posterior expectation of θ_1 in Example 7.2.4.

7.2.6 Establish that the variance of the θ in Example 7.2.2 is as given in Example 7.2.6. Prove that this goes to 0 as $n \to \infty$.

7.2.7 Establish that the variance of θ_1 in Example 7.2.4 is as given in Example 7.2.6. Prove that this goes to 0 as $n \to \infty$.

7.2.8 In Example 7.2.14, which of the two predictors derived there do you find more sensible? Why?

7.2.9 In Example 7.2.15, prove that the posterior predictive distribution for X_{n+1} is as stated. (Hint: Write the posterior predictive distribution density as an expectation.)

7.2.10 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Exponential(λ) distribution, where $\lambda > 0$ is unknown and $\lambda \sim$ Gamma(α_0, β_0). Determine the mode of posterior distribution of λ. Also determine the posterior expectation and posterior variance of λ.

7.2.11 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Exponential(λ) distribution where $\lambda > 0$ is unknown and $\lambda \sim$ Gamma(α_0, β_0). Determine the mode of posterior distribution of a future independent observation X_{n+1}. Also determine the posterior expectation of X_{n+1} and posterior variance of X_{n+1}. (Hint: Problems 3.2.16 and 3.3.20.)

7.2.12 Suppose that in a population of students in a course with a large enrollment, the mark, out of 100, on a final exam is approximately distributed $N(\mu, 9)$. The instructor places the prior $\mu \sim N(65, 1)$ on the unknown parameter. A sample of 10 marks is obtained as given below.

46	68	34	86	75	56	77	73	53	64

(a) Determine the posterior mode and a 0.95-credible interval for μ. What does this interval tell you about the accuracy of the estimate?

(b) Use the 0.95-credible interval for μ to test the hypothesis $H_0 : \mu = 65$.

(c) Suppose we assign prior probability 0.5 to $\mu = 65$. Using the mixture prior $\Pi = 0.5\Pi_1 + 0.5\Pi_2$, where Π_1 is degenerate at $\mu = 65$ and Π_2 is the $N(65, 1)$ distribution, compute the posterior probability of the null hypothesis.

(d) Compute the Bayes factor in favor of $H_0 : \mu = 65$ when using the mixture prior.

7.2.13 A manufacturer believes that a machine produces rods with lengths in centimeters distributed $N(\mu_0, \sigma^2)$, where μ_0 is known and $\sigma^2 > 0$ is unknown, and that the prior distribution $1/\sigma^2 \sim \text{Gamma}(\alpha_0, \beta_0)$ is appropriate.

(a) Determine the posterior distribution of σ^2 based on a sample $(x_1, \ldots, x_n)$.

(b) Determine the posterior mean of σ^2.

(c) Indicate how you would assess the hypothesis $H_0 : \sigma^2 \leq \sigma_0^2$.

7.2.14 Consider the sampling model and prior in Exercise 7.1.1.

(a) Suppose we want to estimate θ based upon having observed $s = 1$. Determine the posterior mode and posterior mean. Which would you prefer in this situation? Explain why.

(b) Determine a 0.8 HPD region for θ based on having observed $s = 1$.

(c) Suppose instead interest was in $\psi(\theta) = I_{\{1,2\}}(\theta)$. Identify the prior distribution of ψ. Identify the posterior distribution of ψ based on having observed $s = 1$. Determine a 0.5 HPD region for ψ.

7.2.15 For an event A, we have that $P(A^c) = 1 - P(A)$.

(a) What is the relationship between the odds in favor of A and the odds in favor of A^c?

(b) When A is a subset of the parameter space, what is the relationship between the Bayes factor in favor of A and the Bayes factor in favor of A^c?

7.2.16 Suppose you are told that the odds in favor of a subset A are 3 to 1. What is the probability of A? If the Bayes factor in favor of A is 10 and the prior probability of A is 1/2, then determine the posterior probability of A.

7.2.17 Suppose data s is obtained. Two statisticians analyze these data using the same sampling model but different priors, and they are asked to assess a hypothesis H_0. Both statisticians report a Bayes factor in favor of H_0 equal to 100. Statistician I assigned prior probability 1/2 to H_0 whereas statistician II assigned prior probability 1/4 to H_0. Which statistician has the greatest posterior degree of belief in H_0 being true?

7.2.18 You are told that a 0.95-credible interval, determined using the HPD criterion, for a quantity $\psi(\theta)$ is given by $(-3.3, 2.6)$. If you are asked to assess the hypothesis $H_0 : \psi(\theta) = 0$, then what can you say about the Bayesian P-value? Explain your answer.

7.2.19 What is the range of possible values for a Bayes factor in favor of $A \subset \Omega$? Under what conditions will a Bayes factor in favor of $A \subset \Omega$ take its smallest value?

PROBLEMS

7.2.20 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Uniform$[0, \theta]$ distribution, where $\theta > 0$ is unknown, and we have $\theta \sim \text{Gamma}(\alpha_0, \beta_0)$. Determine the mode of the posterior distribution of θ. (Hint: The posterior is not differentiable at $\theta = x_{(n)}$.)

7.2.21 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Uniform$[0, \theta]$ distribution, where $\theta \in (0, 1)$ is unknown, and we have $\theta \sim \text{Uniform}[0, 1]$. Determine the form of the γ-credible interval for θ based on the HPD concept.

7.2.22 In Example 7.2.1, write out the integral given in (7.2.2).

7.2.23 (MV) In Example 7.2.1, write out ithe integral that you would need to evaluate if you wanted to compute the posterior density of the third quartile of the population distribution, i.e., $\psi = \mu + \sigma z_{0.75}$.

7.2.24 Consider the location normal model discussed in Example 7.1.2 and the population coefficient of variation $\psi = \sigma_0/\mu$.

(a) Show that the posterior expectation of ψ does not exist. (Hint: Show that we can write the posterior expectation as

$$\int_{-\infty}^{\infty} \frac{\sigma_0}{a + bz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz,$$

where $b > 0$, and show that this integral does not exist by considering the behavior of the integrand at $z = -a/b$.)

(b) Determine the posterior density of ψ.

(c) Show that you can determine the posterior mode of ψ by evaluating the posterior density at two specific points. (Hint: Proceed by maximizing the logarithm of the posterior density using the methods of calculus.)

7.2.25 (MV) Suppose that $(\theta_1, \ldots, \theta_{k-1}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

(a) Prove that $(\theta_1, \ldots, \theta_{k-2}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_{k-1} + \alpha_k)$. (Hint: In the integral to integrate out θ_{k-1}, make the transformation $\theta_{k-1} \to \theta_{k-1}/(1-\theta_1-\cdots-\theta_{k-2})$.)

(b) Prove that $\theta_1 \sim \text{Beta}(\alpha_1, \alpha_2 + \cdots + \alpha_k)$. (Hint: Use part (a).)

(c) Suppose $(i_1, \ldots, i_k)$ is a permutation of $(1, \ldots, k)$. Prove that $(\theta_{i_1}, \ldots, \theta_{i_{k-1}}) \sim \text{Dirichlet}(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_k})$. (Hint: What is the Jacobian of this transformation?)

(d) Prove that $\theta_i \sim \text{Beta}(\alpha_i, \alpha_{-i})$. (Hint: Use parts (b) and (c).)

7.2.26 (MV) In Example 7.2.4, show that the plug-in MLE of θ_1 is given by x_1/n, i.e., find the MLE of $(\theta_1, \ldots, \theta_k)$ and determine the first coordinate. (Hint: Show there is a unique solution to the score equations and then use the facts that the log-likelihood is bounded above and goes to $-\infty$ whenever $\theta_i \to 0$.)

7.2.27 Compare the results obtained in Exercises 7.2.3 and 7.2.4. What do you conclude about the invariance properties of these estimation procedures? (Hint: Consider Theorem 6.2.1.)

7.2.28 In Example 7.2.5, establish that the posterior variance of μ is as stated in Example 7.2.6. (Hint: Problem 4.6.16.)

7.2.29 In a prediction problem, as described in Section 7.2.4, derive the form of the prior predictive density for t when the joint density of (θ, s, t) is $q_\theta(t \mid s) f_\theta(s) \pi(\theta)$ (assume s and θ are real-valued).

7.2.30 In Example 7.2.16, derive the posterior predictive probability function of (X_{n+1}, X_{n+2}), having observed $x_1, \ldots, x_n$ when $X_1, \ldots, X_n, X_{n+1}, X_{n+2}$ are independently and identically distributed (i.i.d.) Bernoulli(θ).

7.2.31 In Example 7.2.15, derive the posterior predictive distribution for X_{n+1}, having observed $x_1, \ldots, x_n$ when $X_1, \ldots, X_n, X_{n+1}$ are i.i.d. $N(\mu, \sigma_0^2)$. (Hint: We can write $X_{n+1} = \mu + \sigma_0 Z$, where $Z \sim N(0, 1)$ is independent of the posterior distribution of μ.)

7.2.32 For the context of Example 7.2.1, prove that the posterior predictive distribution of an additional future observation X_{n+1} from the population distribution has the same distribution as

$$\mu_x + \sqrt{\frac{2\beta_x\left(\left(n + 1/\tau_0^2\right)^{-1} + 1\right)}{(2\alpha_0 + n)}}\, T,$$

where $T \sim t(2\alpha_0 + n)$. (Hint: Note that we can write $X_{n+1} = \mu + \sigma U$, where $U \sim N(0, 1)$ independent of $X_1, \ldots, X_n, \mu, \sigma$ and then reason as in Example 7.2.1.)

7.2.33 In Example 7.2.1, determine the form of an exact γ-prediction interval for an additional future observation X_{n+1} from the population distribution, based on the HPD concept. (Hint: Use Problem 7.2.32.)

7.2.34 Suppose that Π_1 and Π_2 are discrete probability distributions on the parameter space Ω. Prove that when the prior Π is a mixture $\Pi = p\Pi_1 + (1 - p)\,\Pi_2$, then the prior predictive for the data s is given by $m(s) = pm_1(s) + (1 - p)m_2(s)$, and the posterior probability measure is given by (7.2.6).

7.2.35 (MV) Suppose that $\theta = (\theta_1, \theta_2) \in R^2$ and $h(\theta_1, \theta_2) = (\psi(\theta), \lambda(\theta)) \in R^2$. Assume that h satisfies the necessary conditions and establish (7.2.1). (Hint: Theorem 2.9.2.)

CHALLENGES

7.2.36 Another way to assess the null hypothesis $H_0 : \psi(\theta) = \psi_0$ is to compute the P-value

$$\Pi\left(\frac{\omega(\psi(\theta) \mid s)}{\omega(\psi(\theta))} \leq \frac{\omega(\psi_0 \mid s)}{\omega(\psi_0)} \,\middle|\, s\right) \tag{7.2.12}$$

where ω is the marginal prior density or probability function of ψ. We call (7.2.12) the *observed relative surprise* of H_0.

The quantity $\omega(\psi_0 \mid s)/\omega(\psi_0)$ is a measure of how the data s have changed our *a priori* belief that ψ_0 is the true value of ψ. When (7.2.12) is small, ψ_0 is a surprising value for ψ, as this indicates that the data have increased our belief more for other values of ψ.

(a) Prove that (7.2.12) is invariant under 1–1 continuously differentiable transformations of ψ.

(b) Show that a value ψ_0 that makes (7.2.12) smallest, maximizes $\omega(\psi_0 \mid s)/\omega(\psi_0)$. We call such a value a *least relative suprise estimate* of ψ.

(c) Indicate how to use (7.2.12) to form a γ-credible region, known as a γ-relative surprise region, for ψ.

(d) Suppose that ψ is real-valued with prior density ω and posterior density $\omega(\cdot \mid s)$ both continuous and positive at ψ_0. Let $A_\epsilon = (\psi_0 - \epsilon, \psi_0 + \epsilon)$. Show that $BF_{A_\epsilon} \to \omega(\psi_0 \mid s)/\omega(\psi_0)$ as $\epsilon \downarrow 0$. Generalize this to the case where ψ takes its values in an open subset of R^k. This shows that we can think of the observed relative surprise as a way of calibrating Bayes factors.

7.3 Bayesian Computations

In virtually all the examples in this chapter so far, we have been able to work out the exact form of the posterior distributions and carry out a number of important computations using these. It often occurs, however, that we cannot derive any convenient form for the posterior distribution. Furthermore, even when we *can* derive the posterior distribution, there computations might arise that cannot be carried out exactly — e.g., recall the discussion in Example 7.2.1 that led to the integral (7.2.2). These calculations involve evaluating complicated sums or integrals. Therefore, when we apply Bayesian inference in a practical example, we need to have available methods for approximating these quantities.

The subject of approximating integrals is an extensive topic that we cannot deal with fully here.[1] We will, however, introduce several approximation methods that arise very naturally in Bayesian inference problems.

7.3.1 Asymptotic Normality of the Posterior

In many circumstances, it turns out that the posterior distribution of $\theta \in R^1$ is approximately normally distributed. We can then use this to compute approximate credible regions for the true value of θ, carry out hypothesis assessment, etc. One such result says that, under conditions that we will not describe here, when $(x_1, \ldots, x_n)$ is a sample from f_θ, then

$$\Pi\left(\frac{\theta - \hat{\theta}(x_1, \ldots, x_n)}{\hat{\sigma}(x_1, \ldots, x_n)} \leq z \;\middle|\; x_1, \ldots, x_n \right) \to \Phi(z)$$

as $n \to \infty$, where $\hat{\theta}(x_1, \ldots, x_n)$ is the posterior mode, and

$$\hat{\sigma}^2(x_1, \ldots, x_n) = \left(-\frac{\partial^2 \ln(L(\theta \mid x_1, \ldots, x_n)\pi(\theta))}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \right)^{-1}.$$

Note that this result is similar to Theorem 6.5.3 for the MLE. Actually, we can replace $\hat{\theta}(x_1, \ldots, x_n)$ by the MLE and replace $\hat{\sigma}^2(x_1, \ldots, x_n)$ by the observed information (see Section 6.5), and the result still holds. When θ is k-dimensional, there is a similar but more complicated result.

7.3.2 Sampling from the Posterior

Typically, there are many things we want to compute as part of implementing a Bayesian analysis. Many of these can be written as expectations with respect to the posterior distribution of θ. For example, we might want to compute the posterior probability content of a subset $A \subset \Omega$, namely,

$$\Pi(A \mid s) = E(I_A(\theta) \mid s).$$

[1] See, for example, *Approximating Integrals via Monte Carlo and Deterministic Methods,* by M. Evans and T. Swartz (Oxford University Press, Oxford, 2000).

More generally, we want to be able to compute the posterior expectation of some arbitrary function $w(\theta)$, namely

$$E(w(\theta) \mid s). \tag{7.3.1}$$

It would certainly be convenient if we could compute all these quantities exactly, but quite often we cannot. In fact, it is not really necessary that we evaluate (7.3.1) exactly. This is because we naturally expect any inference we make about the true value of the parameter to be subject (different data sets of the same size lead to different inferences) to sampling error. It is not necessary to carry out our computations to a much higher degree of precision than what sampling error contributes. For example, if the sampling error only allows us to know the value of a parameter to within only ± 0.1 units, then there is no point in computing an estimate to many more digits of accuracy.

In light of this, many of the computational problems associated with implementing Bayesian inference are effectively solved if we can sample from the posterior for θ. For when this is possible, we simply generate an i.i.d. sequence $\theta_1, \theta_2, \ldots, \theta_N$ from the posterior distribution of θ and estimate (7.3.1) by

$$\bar{w} = \frac{1}{N} \sum_{i=1}^{N} w(\theta_i).$$

We know then, from the strong law of large numbers (see Theorem 4.3.2), that $\bar{w} \overset{a.s.}{\to} E(w(\theta) \mid x)$ as $N \to \infty$.

Of course, for any given N, the value of $\bar{w}$ only approximates (7.3.1); we would like to know that we have chosen N large enough so that the approximation is appropriately accurate. When $E(w^2(\theta) \mid s) < \infty$, then the central limit theorem (see Theorem 4.4.3) tells us that

$$\frac{\bar{w} - E(w(\theta) \mid s)}{\sigma_w / \sqrt{N}} \overset{D}{\to} N(0, 1)$$

as $N \to \infty$, where $\sigma_w^2 = \mathrm{Var}(w(\theta) \mid s)$. In general, we do not know the value of σ_w^2, but we can estimate it by

$$s_w^2 = \frac{1}{N-1} \sum_{i=1}^{N} (w(\theta_i) - \bar{w})^2$$

when $w(\theta)$ is a quantitative variable, and by $s_w^2 = \bar{w}(1 - \bar{w})$ when $w = I_A$ for $A \subset \Omega$. As shown in Section 4.4.2, in either case, s_w^2 is a consistent estimate of σ_w^2. Then, by Corollary 4.4.4, we have that

$$\frac{\bar{w} - E(w(\theta) \mid s)}{s_w / \sqrt{N}} \overset{D}{\to} N(0, 1)$$

as $N \to \infty$.

From this result we know that

$$\bar{w} \pm 3 \frac{s_w}{\sqrt{N}}$$

is an approximate 100% confidence interval for $E(w(\theta) \mid s)$, so we can look at $3s_w/\sqrt{N}$ to determine whether or not N is large enough for the accuracy required.

One caution concerning this approach to assessing error is that $3s_w/\sqrt{N}$ is itself subject to error, as s_w is an estimate of σ_w, so this could be misleading. A common recommendation then is to monitor the value of $3s_w/\sqrt{N}$ for successively larger values of N and stop the sampling only when it is clear that the value of $3s_w/\sqrt{N}$ is small enough for the accuracy desired and appears to be declining appropriately. Even this approach, however, will not give a guaranteed bound on the accuracy of the computations, so it is necessary to be cautious.

It is also important to remember that application of these results requires that $\sigma_w^2 < \infty$. For a bounded w, this is always true, as any bounded random variable always has a finite variance. For an unbounded w, however, this must be checked — sometimes this is very difficult to do.

We consider an example where it is possible to exactly sample from the posterior.

EXAMPLE 7.3.1 *Location-Scale Normal*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution where $\mu \in R^1$ and $\sigma > 0$ are unknown, and we use the prior given in Example 7.1.4. The posterior distribution for (μ, σ^2) developed there is

$$\mu \mid \sigma^2, x_1, \ldots, x_n \sim N\left(\mu_x, (n + 1/\tau_0^2)^{-1}\sigma^2\right) \tag{7.3.2}$$

and

$$1/\sigma^2 \mid x_1, \ldots, x_n \sim \text{Gamma}(\alpha_0 + n/2, \beta_x), \tag{7.3.3}$$

where μ_x is given by (7.1.7) and β_x is given by (7.1.8).

Most statistical packages have built-in generators for gamma distributions and for the normal distribution. Accordingly, it is very easy to generate a sample $(\mu_1, \sigma_1^2), \ldots, (\mu_N, \sigma_N^2)$ from this posterior. We simply generate a value for $1/\sigma_i^2$ from the specified gamma distribution; then, given this value, we generate the value of μ_i from the specified normal distribution.

Suppose, then, that we want to derive the posterior distribution of the coefficient of variation $\psi = \sigma/\mu$. To do this we generate N values from the joint posterior of (μ, σ^2), using (7.3.2) and (7.3.3), and compute ψ for each of these. We then know immediately that $\psi_1, \ldots, \psi_N$ is a sample from the posterior distribution of ψ.

As a specific numerical example, suppose that we observed the following sample $(x_1, \ldots, x_{15})$.

11.6714	1.8957	2.1228	2.1286	1.0751
8.1631	1.8236	4.0362	6.8513	7.6461
1.9020	7.4899	4.9233	8.3223	7.9486

Here, $\bar{x} = 5.2$ and $s = 3.3$. Suppose further that the prior is specified by $\mu_0 = 4$, $\tau_0^2 = 2$, $\alpha_0 = 2$, and $\beta_0 = 1$.

From (7.1.7), we have

$$\mu_x = \left(15 + \frac{1}{2}\right)^{-1}\left(\frac{4}{2} + 15 \cdot 5.2\right) = 5.161,$$

and from (7.1.8),

$$\begin{aligned}\beta_x &= 1 + \frac{15}{2}(5.2)^2 + \frac{4^2}{2\cdot 2} + \frac{14}{2}(3.3)^2 - \frac{1}{2}\left(15+\frac{1}{2}\right)^{-1}\left(\frac{4}{2}+15\cdot 5.2\right)^2 \\ &= 77.578.\end{aligned}$$

Therefore, we generate

$$1/\sigma^2 \,|\, x_1, \ldots, x_n \sim \text{Gamma}(9.5, 77.578),$$

followed by

$$\mu \,|\, \sigma^2, x_1, \ldots, x_n \sim N(5.161, (15.5)^{-1}\sigma^2).$$

See Appendix B for some code that can be used to generate from this joint distribution.

In Figure 7.3.1, we have plotted a sample of $N = 200$ values of (μ, σ^2) from this joint posterior. In Figure 7.3.2, we have plotted a density histogram of the 200 values of ψ that arise from this sample.

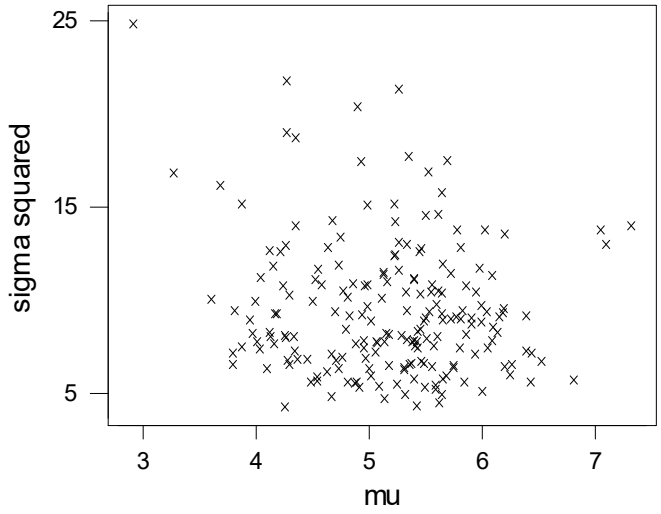

Figure 7.3.1: A sample of 200 values of (μ, σ^2) from the joint posterior in Example 7.3.1 when $n = 15, \bar{x} = 5.2, s = 3.3, \mu_0 = 4, \tau_0^2 = 2, \alpha_0 = 2$, and $\beta_0 = 1$.

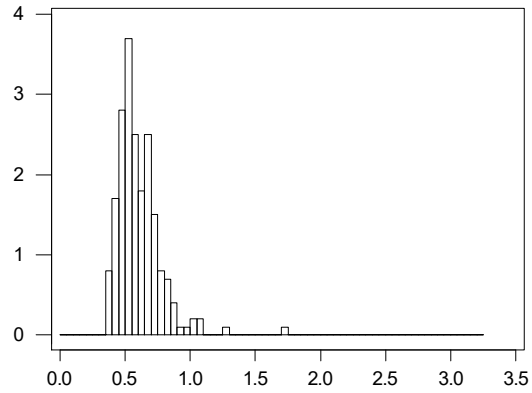

Figure 7.3.2: A density histogram of 200 values from the posterior distribution of ψ in Example 7.3.1.

A sample of 200 is not very large, so we next generated a sample of $N = 10^3$ values from the posterior distribution of ψ. A density histogram of these values is provided in Figure 7.3.3. In Figure 7.3.4, we have provided a density histogram based on a sample of $N = 10^4$ values. We can see from this that at $N = 10^3$, the basic shape of the distribution has been obtained, although the right tail is not being very accurately estimated. Things look better in the right tail for $N = 10^4$, but note there are still some extreme values quite disconnected from the main mass of values. As is characteristic of most distributions, we will need very large values of N to accurately estimate the tails. In any case, we have learned that this distribution is skewed to the right with a long right tail.

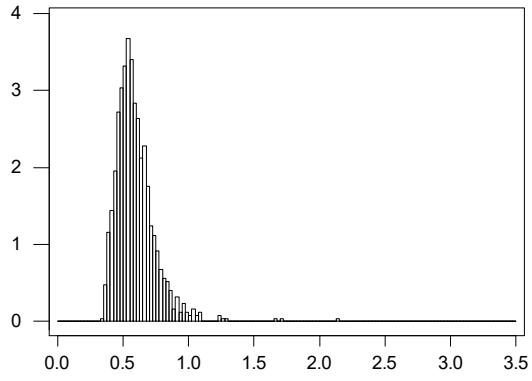

Figure 7.3.3: A density histogram of 1000 values from the posterior distribution of ψ in Example 7.3.1.

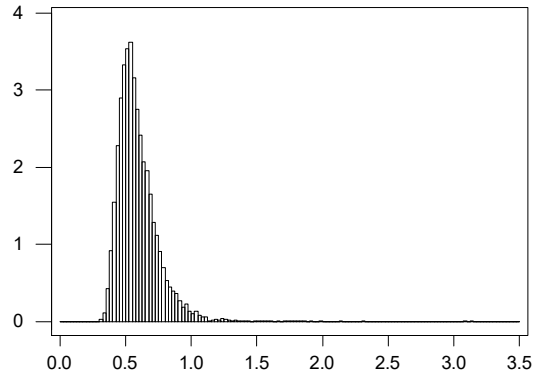

Figure 7.3.4: A density histogram of $N = 10^4$ values from the posterior distribution of ψ in Example 7.3.1.

Suppose we want to estimate

$$\Pi(\psi \leq 0.5 \,|\, x_1, \ldots, x_n) = E(I_{(-\infty,0.5)}(\psi) \,|\, x_1, \ldots, x_n).$$

Now $w = I_{(-\infty,0.5)}$ is bounded so its posterior variance exists. In the following table, we have recorded the estimates for each N together with the standard error based on each of the generated samples. We have included some code for computing these estimates and their standard errors in Appendix B. Based on the results from $N = 10^4$, it would appear that this posterior probability is in the interval $0.289 \pm 3\,(0.0045) = [0.2755, 0.3025]$.

N	Estimate of $\Pi(\psi \leq 0.5 \,\|\, x_1, \ldots, x_n)$	Standard Error
200	0.265	0.0312
10^3	0.271	0.0141
10^4	0.289	0.0045

This example also demonstrates an important point. It would be very easy for us to calculate the sample mean of the values of ψ generated from its posterior distribution and then consider this as an estimate of the posterior mean of ψ. But Problem 7.2.24 suggests (see Problem 7.3.15) that this mean will not exist. Accordingly, a Monte Carlo estimate of this quantity does not make any sense! So we must always check first that any expectation we want to estimate exists, before we proceed with some estimation procedure. ∎

When we cannot sample directly from the posterior, then the methods of the following section are needed.

7.3.3 Sampling from the Posterior Via Gibbs Sampling (Advanced)

Sampling from the posterior, as described in Section 7.3.2, is very effective, when it can be implemented. Unfortunately, it is often difficult or even impossible to do this directly, as we did in Example 7.3.1. There are, however, a number of algorithms that allow us to approximately sample from the posterior. One of these, known as *Gibbs sampling*, is applicable in many statistical contexts.

To describe this algorithm, suppose we want to generate samples from the joint distribution of $(Y_1, \ldots, Y_k) \in R^k$. Further suppose that we can generate from each of the full conditional distributions $Y_i \mid Y_{-i} = y_{-i}$, where

$$Y_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_k),$$

namely, we can generate from the conditional distribution of Y_i given the values of all the other coordinates. The Gibbs sampler then proceeds iteratively as follows.

1. Specify an initial value $(y_{1(0)}, \ldots, y_{k(0)})$ for $(Y_1, \ldots, Y_k)$.
2. For $N > 0$, generate $Y_{i(N)}$ from its conditional distribution given $(y_{1(N)}, \ldots, y_{i-1(N)}, y_{i+1(N-1)}, \ldots, y_{k(N-1)})$ for each $i = 1, \ldots, k$.

For example, if $k = 3$, we first specify $(y_{1(0)}, y_{2(0)}, y_{3(0)})$. Then we generate

$$\begin{aligned} Y_{1(1)} \mid Y_{2(0)} &= y_{2(0)},\ Y_{3(0)} = y_{3(0)} \\ Y_{2(1)} \mid Y_{1(1)} &= y_{1(1)},\ Y_{3(0)} = y_{3(0)} \\ Y_{3(1)} \mid Y_{1(1)} &= y_{1(1)},\ Y_{2(1)} = y_{2(1)} \end{aligned}$$

to obtain $(Y_{1(1)}, Y_{2(1)}, Y_{3(1)})$. Next we generate

$$\begin{aligned} Y_{1(2)} \mid Y_{2(1)} &= y_{2(1)},\ Y_{3(1)} = y_{3(1)} \\ Y_{2(2)} \mid Y_{1(2)} &= y_{1(2)},\ Y_{3(1)} = y_{3(1)} \\ Y_{3(2)} \mid Y_{1(2)} &= y_{1(2)},\ Y_{2(2)} = y_{2(2)} \end{aligned}$$

to obtain $(Y_{1(2)}, Y_{2(2)}, Y_{3(2)})$, etc. Note that we actually did not need to specify $Y_{1(0)}$, as it is never used.

It can then be shown (see Section 11.3) that, in fairly general circumstances, $(Y_{1(N)}, \ldots, Y_{k(N)})$ converges in distribution to the joint distribution of $(Y_1, \ldots, Y_k)$ as $N \to \infty$. So for large N, we have that the distribution of $(Y_{1(N)}, \ldots, Y_{k(N)})$ is approximately the same as the joint distribution of $(Y_1, \ldots, Y_k)$ from which we want to sample. So Gibbs sampling provides an approximate method for sampling from a distribution of interest.

Furthermore, and this is the result that is most relevant for simulations, it can be shown that, under conditions,

$$\bar{w} = \frac{1}{N} \sum_{i=1}^{N} w(Y_{1(i)}, \ldots, Y_{k(i)}) \overset{a.s.}{\to} E(w(Y_1, \ldots, Y_k)).$$

Estimation of the variance of $\bar{w}$ is different than in the i.i.d. case, where we used the sample variance, because now the $w(Y_{1(i)}, \ldots, Y_{k(i)})$ terms are not independent.

There are several approaches to estimating the variance of $\bar{w}$, but perhaps the most commonly used is the technique of *batching*. For this we divide the sequence

$$w(Y_{1(0)}, \ldots, Y_{k(0)}), \ldots, w(Y_{1(N)}, \ldots, Y_{k(N)})$$

into N/m nonoverlapping sequential batches of size m (assuming here that N is divisible by m), calculate the mean in each batch obtaining $\bar{w}_1, \ldots, \bar{w}_{N/m}$, and then estimate the variance of $\bar{w}$ by

$$\frac{s_b^2}{N/m}, \tag{7.3.4}$$

where s_b^2 is the sample variance obtained from the batch means, i.e.,

$$s_b^2 = \frac{1}{N/m - 1} \sum_{i=1}^{N/m} (\bar{w}_i - \bar{w})^2 .$$

It can be shown that $(Y_{1(i)}, \ldots, Y_{k(i)})$ and $(Y_{1(i+m)}, \ldots, Y_{k(i+m)})$ are approximately independent for m large enough. Accordingly, we choose the batch size m large enough so that the batch means are approximately independent, but not so large as to leave very few degrees of freedom for the estimation of the variance. Under ideal conditions, $\bar{w}_1, \ldots, \bar{w}_{N/m}$ is an i.i.d. sequence with sample mean

$$\bar{w} = \frac{1}{N/m} \sum_{i=1}^{N/m} \bar{w}_i,$$

and, as usual, we estimate the variance of $\bar{w}$ by (7.3.4).

Sometimes even Gibbs sampling cannot be directly implemented because we cannot obtain algorithms to generate from all the full conditionals. There are a variety of techniques for dealing with this, but in many statistical applications the technique of *latent variables* often works. For this, we search for some random variables, say $(V_1, \ldots, V_l)$, where each Y_i is a function of $(V_1, \ldots, V_l)$ and such that we can apply Gibbs sampling to the joint distribution of $(V_1, \ldots, V_l)$. We illustrate Gibbs sampling via latent variables in the following example.

EXAMPLE 7.3.2 *Location-Scale Student*
Suppose now that $(x_1, \ldots, x_n)$ is a sample from a distribution that is of the form $X = \mu + \sigma Z$, where $Z \sim t(\lambda)$ (see Section 4.6.2 and Problem 4.6.14). If $\lambda > 2$, then μ is the mean and $\sigma(\lambda/(\lambda-2))^{1/2}$ is the standard deviation of the distribution (see Problem 4.6.16). Note that $\lambda = \infty$ corresponds to normal variation, while $\lambda = 1$ corresponds to Cauchy variation.

We will fix λ at some specified value to reflect the fact that we are interested in modeling situations in which the variable under consideration has a distribution with longer tails than the normal distribution. Typically, this manifests itself in a histogram of the data with a roughly symmetric shape but exhibiting a few extreme values out in the tails, so a $t(\lambda)$ distribution might be appropriate.

Suppose we place the prior on (μ, σ^2), given by $\mu \mid \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2)$ and $1/\sigma^2 \sim \text{Gamma}(\alpha_0, \beta_0)$. The likelihood function is given by

$$\left(\frac{1}{\sigma^2}\right)^{n/2} \prod_{i=1}^{n} \left[1 + \frac{1}{\lambda}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]^{-(\lambda+1)/2}, \tag{7.3.5}$$

hence the posterior density of $\left(\mu, 1/\sigma^2\right)$ is proportional to

$$\left(\frac{1}{\sigma^2}\right)^{n/2} \prod_{i=1}^{n} \left[1 + \frac{1}{\lambda}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]^{-(\lambda+1)/2} \times \left(\frac{1}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau_0^2\sigma^2}\left(\mu - \mu_0\right)^2\right)\left(\frac{1}{\sigma^2}\right)^{\alpha_0 - 1} \exp\left(-\frac{\beta_0}{\sigma^2}\right).$$

This distribution is not immediately recognizable, and it is not at all clear how to generate from it.

It is natural, then, to see if we can implement Gibbs sampling. To do this directly, we need an algorithm to generate from the posterior of μ given the value of σ^2, and an algorithm to generate from the posterior of σ^2 given μ. Unfortunately, neither of these conditional distributions is amenable to the techniques discussed in Section 2.10, so we cannot implement Gibbs sampling directly.

Recall, however, that when $V \sim \chi^2(\lambda) = \text{Gamma}(\lambda/2, 1/2)$ (see Problem 4.6.13) independent of $Y \sim N(\mu, \sigma^2)$, then (Problem 4.6.14)

$$Z = \frac{Y - \mu}{\sigma\sqrt{V/\lambda}} \sim t(\lambda).$$

Therefore, writing

$$X = \mu + \sigma Z = \mu + \sigma \frac{Y - \mu}{\sigma\sqrt{V/\lambda}} = \mu + \frac{Y - \mu}{\sqrt{V/\lambda}},$$

we have that $X \mid V = v \sim N(\mu, \sigma^2\lambda/v)$.

We now introduce the *n latent* or *hidden variables* $(V_1, \ldots, V_n)$, which are i.i.d. $\chi^2(\lambda)$ and suppose $X_i \mid V_i = v_i \sim N(\mu, \sigma^2\lambda/v_i)$. The V_i are considered latent because they are not really part of the problem formulation but have been added here for convenience (as we shall see). Then, noting that there is a factor $v_i^{1/2}$ associated with the density of $X_i \mid V_i = v_i$, the joint density of the values $(X_1, V_1), \ldots, (X_n, V_n)$ is proportional to

$$\left(\frac{1}{\sigma^2}\right)^{n/2} \prod_{i=1}^{n} \exp\left(-\frac{v_i}{2\sigma^2\lambda}\left(x_i - \mu\right)^2\right) v_i^{(\lambda/2)-(1/2)} \exp\left(-\frac{v_i}{2}\right).$$

From the above argument, the marginal joint density of $(X_1, \ldots, X_n)$ (after integrating out the v_i's) is proportional to (7.3.5), namely, a sample of n from the distribution

specified by $X = \mu + \sigma Z$, where $Z \sim t(\lambda)$. With the same prior structure as before, we have that the joint density of

$$(X_1, V_1), \ldots, (X_n, V_n), \mu, 1/\sigma^2$$

is proportional to

$$\left(\frac{1}{\sigma^2}\right)^{n/2} \prod_{i=1}^{n} \exp\left(-\frac{v_i}{2\sigma^2\lambda}(x_i-\mu)^2\right) v_i^{(\lambda/2)-(1/2)} \exp\left(-\frac{v_i}{2}\right) \times \left(\frac{1}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau_0^2\sigma^2}(\mu-\mu_0)^2\right)\left(\frac{1}{\sigma^2}\right)^{\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma^2}\right). \quad (7.3.6)$$

In (7.3.6), treat $x_1, \ldots, x_n$ as constants (we observed these values) and consider the conditional distributions of each of the variables $V_1, \ldots, V_n, \mu, 1/\sigma^2$ given all the other variables. From (7.3.6), we have that the full conditional density of μ is proportional to

$$\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}(x_i-\mu)^2 + \frac{1}{\tau_0^2}(\mu-\mu_0)^2\right)\right\},$$

which is proportional to

$$\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}\right)+\frac{1}{\tau_0^2}\right]\mu^2 + \frac{2}{2\sigma^2}\left[\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}x_i\right)+\frac{\mu_0}{\tau_0^2}\right]\mu\right\}.$$

From this, we immediately deduce that

$$\mu \mid x_1, \ldots, x_n, v_1, \ldots, v_n, \sigma^2 \sim N\left(r(v_1, \ldots, v_n)\left[\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}x_i\right)+\frac{\mu_0}{\tau_0^2}\right], r(v_1, \ldots, v_n)\sigma^2\right),$$

where

$$r(v_1, \ldots, v_n) = \left[\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}\right)+\frac{1}{\tau_0^2}\right]^{-1}.$$

From (7.3.6), we have that the conditional density of $1/\sigma^2$ is proportional to

$$\left(\frac{1}{\sigma^2}\right)^{(n/2)+\alpha_0-(1/2)} \exp\left\{-\left(\begin{array}{c}\sum_{i=1}^{n}\frac{v_i}{\lambda}(x_i-\mu)^2 \\ +\frac{1}{\tau_0^2}(\mu-\mu_0)^2+2\beta_0\end{array}\right)\frac{1}{2\sigma^2}\right\},$$

and we immediately deduce that

$$\frac{1}{\sigma^2} \mid x_1, \ldots, x_n, v_1, \ldots, v_n, \mu \sim \text{Gamma}\left(\frac{n}{2}+\alpha_0+\frac{1}{2}, \frac{1}{2}\left(\sum_{i=1}^{n}\frac{v_i}{\lambda}(x_i-\mu)^2 + \frac{1}{\tau_0^2}(\mu-\mu_0)^2 + 2\beta_0\right)\right).$$

Finally, the conditional density of V_i is proportional to

$$v_i^{(\lambda/2)-(1/2)} \exp\left\{-\left[\frac{(x_i-\mu)^2}{2\sigma^2\lambda}+\frac{1}{2}\right]v_i\right\},$$

and it is immediate that

$$V_i \,|x_1,\ldots,x_n,\ v_1,\ldots,v_{i-1},v_{i+1},\ldots,v_n,\mu,\sigma^2 \sim \text{Gamma}\left(\frac{\lambda}{2}+\frac{1}{2},\frac{1}{2}\left(\frac{(x_i-\mu)^2}{\sigma^2\lambda}+1\right)\right).$$

We can now easily generate from all these distributions and implement a Gibbs sampling algorithm. As we are not interested in the values of $V_1,\ldots,V_n$, we simply discard these as we iterate.

Let us now consider a specific computation using the same data and prior as in Example 7.3.1. The analysis of Example 7.3.1 assumed that the data were coming from a normal distribution, but now we are going to assume that the data are a sample from a $\mu+\sigma t(3)$ distribution, i.e., $\lambda=3$. We again consider approximating the posterior distribution of the coefficient of variation $\psi=\sigma/\mu$.

We carry out the Gibbs sampling iteration in the order $v_1,\ldots,v_n,\mu,1/\sigma^2$. This implies that we need starting values only for μ and σ^2 (the full conditionals of the v_i do not depend on the other v_j). We take the starting value of μ to be $\bar{x}=5.2$ and the starting value of σ to be $s=3.3$. For each generated value of (μ,σ^2), we calculate ψ to obtain the sequence $\psi_1,\psi_2,\ldots,\psi_N$.

The values $\psi_1,\psi_2,\ldots,\psi_N$ are not i.i.d. from the posterior of ψ. The best we can say is that

$$\psi_m \xrightarrow{D} \psi \sim \omega(\cdot\,|\,x_1,\ldots,x_n)$$

as $m\to\infty$, where $\omega(\cdot\,|\,x_1,\ldots,x_n)$ is the posterior density of ψ. Also, values sufficiently far apart in the sequence, will be like i.i.d. values from $\omega(\cdot\,|\,x_1,\ldots,x_n)$. Thus, one approach is to determine an appropriate value m and then extract $\psi_m,\psi_{2m},\psi_{3m},\ldots$ as an approximate i.i.d. sequence from the posterior. Often it is difficult to determine an appropriate value for m, however.

In any case, it is known that, under fairly weak conditions,

$$\bar{w}=\frac{1}{N}\sum_{i=1}^{N} w(\psi_i) \xrightarrow{a.s.} E(w(\psi)\,|\,x_1,\ldots,x_n)$$

as $N\to\infty$. So we can use the whole sequence $\psi_1,\psi_2,\ldots,\psi_N$ and record a density histogram for ψ, just as we did in Example 7.3.1. The value of the density histogram between two cut points will converge almost surely to the correct value as $N\to\infty$. However, we will have to take N larger when using the Gibbs sampling algorithm than with i.i.d. sampling, to achieve the same accuracy. For many examples, the effect of the deviation of the sequence from being i.i.d. is very small, so N will not have to be much larger. We always need to be cautious, however, and the general recommendation is to

compute estimates for successively higher values of N, only stopping when the results seem to have stabilized.

In Figure 7.3.5, we have plotted the density histogram of the ψ values that resulted from 10^4 iterations of the Gibbs sampler. In this case, plotting the density histogram of ψ based upon $N = 5 \times 10^4$ and $N = 8 \times 10^4$ resulted in only minor deviations from this plot. Note that this density looks very similar to that plotted in Example 7.3.1, but it is not quite so peaked and it has a shorter right tail.

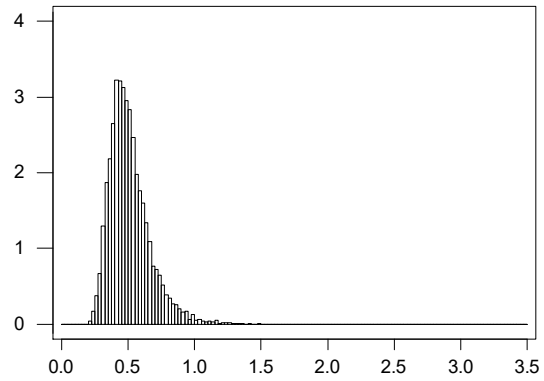

Figure 7.3.5: A density histogram of $N = 10^4$ values of ψ generated sequentially via Gibbs sampling in Example 7.3.2.

We can also estimate $\Pi(\psi \leq 0.5 \,|\, x_1, \ldots, x_n)$, just as we did in Example 7.3.1, by recording the proportion of values in the sequence that are smaller than 0.5, i.e., $w(\psi) = I_A(\psi)$, where $A = \{\theta : \psi \leq 0.5\}$. In this case, we obtained the estimate 0.5441, which is quite different from the value obtained in Example 7.3.1. So using a $t(3)$ distribution to describe the variation in the response has made a big difference in the results.

Of course, we must also quantify how accurate we believe our estimate is. Using a batch size of $m = 10$, we obtained the standard error of the estimate 0.5441 to be 0.00639. When we took the batch size to be $m = 20$, the standard error of the mean is 0.00659; with a batch size of $m = 40$, the standard error of the mean is 0.00668. So we feel quite confident that we are assessing the error in the estimate appropriately. Again, under conditions, we have that $\bar{w}$ is asymptotically normal so that in this case we can assert that the interval $0.5441 \pm 3(0.0066) = [0.5243, 0.5639]$ contains the true value of $\Pi(\psi \leq 0.5 \,|\, x_1, \ldots, x_n)$ with virtual certainty.

See Appendix B for some code that was used to implement the Gibbs sampling algorithm described here. ∎

It is fair to say that the introduction of Gibbs sampling has resulted in a revolution in statistical applications due to the wide variety of previously intractable problems that it successfully handles. There are a number of modifications and closely related

algorithms. We refer the interested reader to Chapter 11, where the general theory of what is called Markov chain Monte Carlo (MCMC) is discussed.

Summary of Section 7.3

- Implementation of Bayesian inference often requires the evaluation of complicated integrals or sums.
- If, however, we can sample from the posterior of the parameter, this will often lead to sufficiently accurate approximations to these integrals or sums via Monte Carlo.
- It is often difficult to sample exactly from a posterior distribution of interest. In such circumstances, Gibbs sampling can prove to be an effective method for generating an approximate sample from this distribution.

EXERCISES

7.3.1 Suppose we have the following sample from an $N(\mu, 2)$ distribution, where μ is unknown.

2.6	4.2	3.1	5.2	3.7	3.8	5.6	1.8	5.3	4.0
3.0	4.0	4.1	3.2	2.2	3.4	4.5	2.9	4.7	5.2

If the prior on μ is Uniform(2, 6), determine an approximate 0.95-credible interval for μ based on the large sample results described in Section 7.3.1.

7.3.2 Determine the form of the approximate 0.95-credible interval of Section 7.3.1, for the Bernoulli model with a Beta(α, β) prior, discussed in Example 7.2.2.

7.3.3 Determine the form of the approximate 0.95-credible intervals of Section 7.3.1, for the location-normal model with an $N(\mu_0, \tau_0^2)$ prior, discussed in Example 7.2.3.

7.3.4 Suppose that $X \sim$ Uniform$[0, 1/\theta]$ and $\theta \sim$ Exponential(1). Derive a crude Monte Carlo algorithm, based on generating from a gamma distribution, to generate a value from the conditional distribution $\theta \mid X = x$. Generalize this to a sample of n from the Uniform$[0, 1/\theta]$ distribution. When will this algortithm be inefficient in the sense that we need a lot of computation to generate a single value?

7.3.5 Suppose that $X \sim N(\theta, 1)$ and $\theta \sim$ Uniform$[0, 1]$. Derive a crude Monte Carlo algorithm, based on generating from a normal distribution, to generate from the conditional distribution $\theta \mid X = x$. Generalize this to a sample of n from the $N(\theta, 1)$ distribution. When will this algortithm be inefficient in the sense that we need a lot of computation to generate a single value?

7.3.6 Suppose that $X \sim 0.5N(\theta, 1) + 0.5N(\theta, 2)$ and $\theta \sim$ Uniform$[0, 1]$. Derive a crude Monte Carlo algorithm, based on generating from a mixure of normal distributions, to generate from the conditional distribution $\theta \mid X = x$. Generalize this to a sample of $n = 2$ from the $0.5N(\theta, 1) + 0.5N(\theta, 2)$ distribution.

COMPUTER EXERCISES

7.3.7 In the context of Example 7.3.1, construct a density histogram of the posterior distribution of $\psi = \mu + \sigma z_{0.25}$, i.e., the population first quartile, using $N = 5 \times 10^3$ and $N = 10^4$, and compare the results. Estimate the posterior mean of this distribution and assess the error in your approximation. (Hint: Modify the program in Appendix B.)

7.3.8 Suppose that a manufacturer takes a random sample of manufactured items and tests each item as to whether it is defective or not. The responses are felt to be i.i.d. Bernoulli(θ), where θ is the probability that the item is defective. The manufacturer places a Beta(0.5, 10) distribution on θ. If a sample of $n = 100$ items is taken and 5 defectives are observed, then, using a Monte Carlo sample with $N = 1000$, estimate the posterior probability that $\theta < 0.1$ and assess the error in your estimate.

7.3.9 Suppose that lifelengths (in years) of a manufactured item are known to follow an Exponential(λ) distribution, where $\lambda > 0$ is unknown and for the prior we take $\lambda \sim$ Gamma(10, 2). Suppose that the lifelengths 4.3, 6.2, 8.4, 3.1, 6.0, 5.5, and 7.8 were observed.

(a) Using a Monte Carlo sample of size $N = 10^3$, approximate the posterior probability that $\lambda \in [3, 6]$ and assess the error of your estimate.

(b) Using a Monte Carlo sample of size $N = 10^3$, approximate the posterior probability function of $\lfloor 1/\lambda \rfloor$ ($\lfloor x \rfloor$ equals the greatest integer less than or equal to x).

(c) Using a Monte Carlo sample of size $N = 10^3$, approximate the posterior expectation of $\lfloor 1/\lambda \rfloor$ and assess the error in your approximation.

7.3.10 Generate a sample of $n = 10$ from a Pareto(2) distribution. Now pretend you only know that you have a sample from a Pareto(α) distribution, where $\alpha > 0$ is unknown, and place a Gamma(2, 1) prior on α. Using a Monte Carlo sample of size $N = 10^4$, approximate the posterior expectation of $1/(\alpha + 1)$ based on the observed sample, and assess the accuracy of your approximation by quoting an interval that contains the exact value with virtual certainty. (Hint: Problem 2.10.15.)

PROBLEMS

7.3.11 Suppose $X_1, \ldots, X_n$ is a sample from the model $\{f_\theta : \theta \in \Omega\}$ and all the regularity conditions of Section 6.5 apply. Assume that the prior $\pi(\theta)$ is a continuous function of θ and that the posterior mode $\hat{\theta}(X_1, \ldots, X_n) \stackrel{a.s.}{\to} \theta$ when $X_1, \ldots, X_n$ is a sample from f_θ (the latter assumption holds under very general conditions).

(a) Using the fact that, if $Y_n \stackrel{a.s.}{\to} Y$ and g is a continuous function, then $g(Y_n) \stackrel{a.s.}{\to} g(Y)$, prove that

$$-\frac{1}{n}\frac{\partial^2 \ln(L(\theta \mid x_1, \ldots, x_n)\pi(\theta))}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \stackrel{a.s.}{\to} I(\theta)$$

when $X_1, \ldots, X_n$ is a sample from f_θ.

(b) Explain to what extent the large sample approximate methods of Section 7.3.1 depend on the prior if the assumptions just described apply.

7.3.12 In Exercise 7.3.10, explain why the interval you constructed to contain the posterior mean of $1/(\alpha+1)$ with virtual certainty may or may not contain the true value of $1/(\alpha+1)$.

7.3.13 Suppose that (X, Y) is distributed Bivariate Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Determine a Gibbs sampling algorithm to generate from this distribution. Assume that you have an algorithm for generating from univariate normal distributions. Is this the best way to sample from this distribution? (Hint: Problem 2.8.27.)

7.3.14 Suppose that the joint density of (X, Y) is given by $f_{X,Y}(x, y) = 8xy$ for $0 < x < y < 1$. Fully describe a Gibbs sampling algorithm for this distribution. In particular, indicate how you would generate all random variables. Can you design an algorithm to generate exactly from this distribution?

7.3.15 In Example 7.3.1, prove that the posterior mean of $\psi = \sigma/\mu$ does not exist. (Hint: Use Problem 7.2.24 and the theorem of total expectation to split the integral into two parts, where one part has value ∞ and the other part has value $-\infty$.)

7.3.16 *(Importance sampling based on the prior)* Suppose we have an algorithm to generate from the prior.

(a) Indicate how you could use this to approximate a posterior expectation using importance sampling (see Problem 4.5.21).

(b) What do you suppose is the major weakness is of this approach?

COMPUTER PROBLEMS

7.3.17 In the context of Example 7.3.2, construct a density histogram of the posterior distribution of $\psi = \mu + \sigma z_{0.25}$, i.e., the population first quartile, using $N = 10^4$. Estimate the posterior mean of this distribution and assess the error in your approximation.

7.4 Choosing Priors

The issue of selecting a prior for a problem is an important one. Of course, the idea is that we choose a prior to reflect our *a priori* beliefs about the true value of θ. Because this will typically vary from statistician to statistician, this is often criticized as being too subjective for scientific studies. It should be remembered, however, that the sampling model $\{f_\theta : \theta \in \Omega\}$ is also a subjective choice by the statistician. These choices are guided by the statistician's judgment. What then justifies one choice of a statistical model or prior over another?

In effect, when statisticians choose a prior and a model, they are prescribing a joint distribution for (θ, s). The only way to assess whether or not an appropriate choice was made is to check whether the observed s is reasonable given this choice. If s is surprising, when compared to the distribution prescribed by the model and prior, then we have evidence against the statistician's choices. Methods designed to assess this are called model-checking procedures, and are discussed in Chapter 9. At this point, however, we should recognize the subjectivity that enters into statistical analyses, but take some comfort that we have a methodology for checking whether or not the choices made by the statistician make sense.

Often a statistician will consider a particular family $\{\pi_\lambda : \lambda \in \Lambda\}$ of priors for a problem and try to select a suitable prior $\pi_{\lambda_0} \in \{\pi_\lambda : \lambda \in \Lambda\}$. In such a context the parameter λ is called a *hyperparameter*. Note that this family could be the set of all possible priors, so there is no restriction in this formulation. We now discuss some commonly used families $\{\pi_\lambda : \lambda \in \Lambda\}$ and methods for selecting $\lambda_0 \in \Lambda$.

7.4.1 Conjugate Priors

Depending on the sampling model, the family may be conjugate.

Definition 7.4.1 The family of priors $\{\pi_\lambda : \lambda \in \Lambda\}$ for the parameter θ of the model $\{f_\theta : \theta \in \Omega\}$ is *conjugate*, if for all data $s \in S$ and all $\lambda \in \Lambda$ the posterior $\pi_\lambda(\cdot \mid s) \in \{\pi_\lambda : \lambda \in \Lambda\}$.

Conjugacy is usually a great convenience as we start with some choice $\lambda_0 \in \Lambda$ for the prior, and then we find the relevant $\lambda_s \in \Lambda$ for the posterior, often without much computation. While conjugacy can be criticized as a mere mathematical convenience, it has to be acknowledged that many conjugate families offer sufficient variety to allow for the expression of a wide spectrum of prior beliefs.

EXAMPLE 7.4.1 *Conjugate Families*
In Example 7.1.1, we have effectively shown that the family of all Beta distributions is conjugate for sampling from the Bernoulli model. In Example 7.1.2, it is shown that the family of normal priors is conjugate for sampling from the location normal model. In Example 7.1.3, it is shown that the family of Dirichlet distributions is conjugate for Multinomial models. In Example 7.1.4, it is shown that the family of priors specified there is conjugate for sampling from the location-scale normal model. ∎

Of course, using a conjugate family does not tell us how to select λ_0. Perhaps the most justifiable approach is to use *prior elicitation*.

7.4.2 Elicitation

Elicitation involves explicitly using the statistician's beliefs about the true value of θ to select a prior in $\{\pi_\lambda : \lambda \in \Lambda\}$ that reflects these beliefs. Typically, these involve the statistician asking questions of himself, or of experts in the application area, in such a way that the answers specify a prior from the family.

EXAMPLE 7.4.2 *Location Normal*
Suppose we are sampling from an $N(\mu, \sigma_0^2)$ distribution with μ unknown and σ_0^2 known, and we restrict attention to the family $\{N(\mu_0, \tau_0^2) : \mu_0 \in R^1, \tau_0^2 > 0\}$ of priors for μ. So here, $\lambda = (\mu_0, \tau_0^2)$ and there are two degrees of freedom in this family. Thus, specifying two independent characteristics specifies a prior.

Accordingly, we could ask an expert to specify two quantiles of his or her prior distribution for μ (see Exercise 7.4.10), as this specifies a prior in the family. For example, we might ask an expert to specify a number μ_0 such that the true value of μ was as likely to be greater than as less than μ_0, so that μ_0 is the median of the prior.

We might also ask the expert to specify a value v_0 such that there is 99% certainty that the true value of μ is less than v_0. This of course is the 0.99-quantile of their prior.

Alternatively, we could ask the expert to specify the center μ_0 of their prior distribution and for a constant τ_0 such that $\mu_0 \pm 3\tau_0$ contains the true value of μ with virtual certainty. Clearly, in this case, μ_0 is the prior mean and τ_0 is the prior standard deviation. ∎

Elicitation is an important part of any Bayesian statistical analysis. If the experts used are truly knowledgeable about the application, then it seems intuitively clear that we will improve a statistical analysis by including such prior information.

The process of elicitation can be somewhat involved, however, for complicated problems. Furthermore, there are various considerations that need to be taken into account involving, prejudices and flaws in the way we reason about probability outside of a mathematical formulation. See Garthwaite, Kadane and O'Hagan (2005), "Statistical methods for eliciting probability distributions", *Journal of the American Statistical Association* (Vol. 100, No. 470, pp. 680–700), for a deeper discussion of these issues.

7.4.3 Empirical Bayes

When the choice of λ_0 is based on the data s, these methods are referred to as *empirical Bayesian methods*. Logically, such methods would seem to violate a basic principle of inference, namely, the principle of conditional probability. For when we compute the posterior distribution of θ using a prior based on s, in general this is no longer the conditional distribution of θ given the data. While this is certainly an important concern, in many problems the application of empirical Bayes leads to inferences with satisfying properties.

For example, one empirical Bayesian method is to compute the prior predictive $m_\lambda(s)$ for the data s, and then base the choice of λ on these values. Note that the prior predictive is like a likelihood function for λ (as it is the density or probability function for the observed s), and so the methods of Chapter 6 apply for inference about λ. For example, we could select the value of λ_s that maximizes $m_\lambda(s)$. The required computations can be extensive, as λ is typically multidimensional. We illustrate with a simple example.

EXAMPLE 7.4.3 *Bernoulli*
Suppose we have a sample $x_1, \ldots, x_n$ from a Bernoulli(θ) distribution and we contemplate putting a Beta(λ, λ) prior on θ for some $\lambda > 0$. So the prior is symmetric about 1/2 and the spread in this distribution is controlled by λ. Since the prior mean is 1/2 and the prior variance is $\lambda^2/[(2\lambda+1)(2\lambda)^2] = 1/4(2\lambda+1) \to 0$ as $\lambda \to \infty$, we see that choosing λ large leads to a very precise prior. Then we have that

$$\begin{aligned} m_\lambda(x_1, \ldots, x_n) &= \frac{\Gamma(2\lambda)}{\Gamma^2(\lambda)} \int_0^1 \theta^{n\bar{x}+\lambda-1}(1-\theta)^{n(1-\bar{x})+\lambda-1}\, d\theta \\ &= \frac{\Gamma(2\lambda)}{\Gamma^2(\lambda)} \frac{\Gamma(n\bar{x}+\lambda)\Gamma(n(1-\bar{x})+\lambda)}{\Gamma(n+2\lambda)}. \end{aligned}$$

It is difficult to find the value of λ that maximizes this, but for real data we can tabulate and plot $m_\lambda(x_1, \ldots, x_n)$ to obtain this value. More advanced computational methods can also be used.

For example, suppose that $n = 20$ and we obtained $n\bar{x} = 5$ as the number of 1's observed. In Figure 7.4.1 we have plotted the graph of $m_\lambda(x_1, \ldots, x_n)$ as a function of λ. We can see from this that the maximum occurs near $\lambda = 2$. More precisely, from a tabulation we determine that $\lambda = 2.3$ is close to the maximum. Accordingly, we use the Beta$(5 + 2.3, 15 + 2.3) =$ Beta$(7.3, 17.3)$ distribution for inferences about θ.

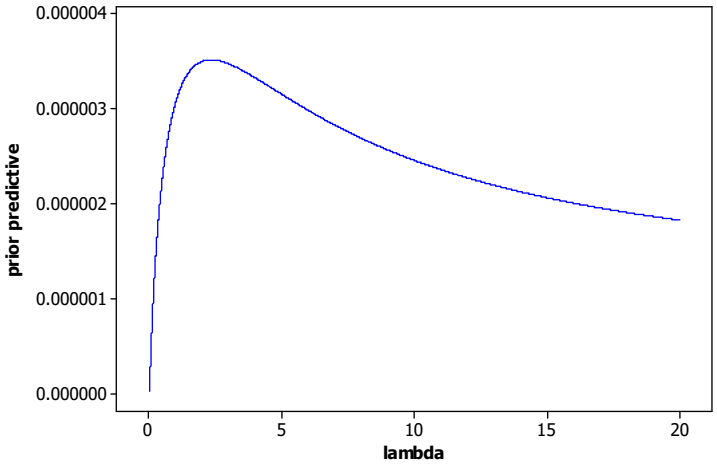

Figure 7.4.1: Plot of $m_\lambda(x_1, \ldots, x_n)$ in Example 7.4.3.

∎

There are many issues concerning empirical Bayes methods. This represents an active area of statistical research.

7.4.4 Hierarchical Bayes

An alternative to choosing a prior for θ in $\{\pi_\lambda : \lambda \in \Lambda\}$ consists of putting yet another prior distribution ω, called a *hyperprior*, on λ. This approach is commonly called *hierarchical Bayes*. The prior for θ basically becomes $\pi(\theta) = \int_\Lambda \pi_\lambda(\theta)\omega(\lambda)\,d\lambda$, so we have in effect integrated out the hyperparameter. The problem then is how to choose the prior ω. In essence, we have simply replaced the problem of choosing the prior on θ with choosing the hyperprior on λ. It is common, in applications using hierarchical Bayes, that default choices are made for ω, although we could also make use of elicitation techniques. We will discuss this further in Section 7.4.5.

So in this situation, the posterior density of θ is equal to

$$\pi(\theta \mid s) = \frac{f_\theta(s)\int_\Lambda \pi_\lambda(\theta)\omega(\lambda)\,d\lambda}{m(s)} = \int_\Lambda \frac{f_\theta(s)\pi_\lambda(\theta)}{m_\lambda(s)}\frac{m_\lambda(s)\omega(\lambda)}{m(s)}\,d\lambda,$$

where $m(s) = \int_\Lambda \int_\Omega f_\theta(s)\pi_\lambda(\theta)\omega(\lambda)\, d\theta\, d\lambda = \int_\Lambda m_\lambda(s)\omega(\lambda)\, d\lambda$ and, for fixed λ, $m_\lambda(s) = \int f_\theta(s)\pi_\lambda(\theta)\, d\theta$ (assuming λ is continuous with prior density given by ω). Note that the posterior density of λ is $m_\lambda(s)\omega(\lambda)/m(s)$ while $f_\theta(s)\pi_\lambda(\theta)/m_\lambda(s)$ is the posterior density of θ given λ.

Therefore, we can use $\pi(\theta \mid s)$ for inferences about the model parameter θ (e.g., estimation, credible regions, and hypothesis assessment) and $m_\lambda(s)\omega(\lambda)/m(s)$ for inferences about λ. Typically, however, we are not interested in λ and in fact it doesn't really make sense to talk about the "true" value of λ. The true value of θ corresponds to the distribution that actually produced the observed data s, at least when the model is correct, while we are not thinking of λ as being generated from ω. This also implies another distinction between θ and λ. For θ is part of the likelihood function based on how the data was generated, while λ is not.

EXAMPLE 7.4.4 *Location-Scale Normal*
Suppose the situation is as is discussed in Example 7.1.4. In that case, both μ and σ^2 are part of the likelihood function and so are model parameters, while $\mu_0, \tau_0^2, \alpha_0$, and β_0 are not, and so they are hyperparameters. To complete this specification as a hierarchical model, we need to specify a prior $\omega(\mu_0, \tau_0^2, \alpha_0, \beta_0)$, a task we leave to a higher-level course. ∎

7.4.5 Improper Priors and Noninformativity

One approach to choosing a prior, and to stop the chain of priors in a hierarchical Bayes approach, is to prescribe a *noninformative prior* based on ignorance. Such a prior is also referred to as a *default prior* or *reference prior*. The motivation is to specify a prior that puts as little information into the analysis as possible and in some sense characterizes ignorance. Surprisingly, in many contexts, statisticians have been led to choose noninformative priors that are *improper*, i.e., $\int_\Omega \pi(\theta)\, d\theta = \infty$, so they do not correspond to probability distributions.

The idea here is to give a rule such that, if a statistician has no prior beliefs about the value of a parameter or hyperparameter, then a prior is prescribed that reflects this. In the hierarchical Bayes approach, one continues up the chain until the statistician declares ignorance, and a default prior completes the specification.

Unfortunately, just how ignorance is to be expressed turns out to be a rather subtle issue. In many cases, the default priors turn out to be *improper*, i.e., the integral or sum of the prior over the whole parameter space equals ∞, e.g., $\int_\Omega \pi(\theta)\, d\theta = \infty$, so the prior is not a probability distribution. The interpretation of an improper prior is not at all clear, and their use is somewhat controversial. Of course, (s, θ) no longer has a joint probability distribution when we are using improper priors, and we cannot use the principle of conditional probability to justify basing our inferences on the posterior.

There have been numerous difficulties associated with the use of improper priors, which is perhaps not surprising. In particular, it is important to note that there is no reason in general for the posterior of θ to exist as a proper probability distribution when π is improper. If an improper prior is being used, then we should always check to make sure the posterior is proper, as inferences will not make sense if we are using an improper posterior.

When using an improper prior π, it is completely equivalent to instead use the prior $c\pi$ for any $c > 0$, for the posterior under π is proper if and only if the posterior under $c\pi$ is proper; then the posteriors are identical (see Exercise 7.4.6).

The following example illustrates the use of an improper prior.

EXAMPLE 7.4.5 *Location Normal Model with an Improper Prior*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \Omega = R^1$ is unknown and σ_0^2 is known. Many arguments for default priors in this context lead to the choice $\pi(\mu) = 1$, which is clearly improper.

Proceeding as in Example 7.1.2, namely, pretending that this π is a proper probability density, we get that the posterior density of μ is proportional to

$$\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x} - \mu)^2\right).$$

This immediately implies that the posterior distribution of μ is $N(\bar{x}, \sigma_0^2/n)$. Note that this is the same as the limiting posterior obtained in Example 7.1.2 as $\tau_0 \to \infty$, although the point of view is quite different. ∎

One commonly used method of selecting a default prior is to use, when it is available, the prior given by $I^{1/2}(\theta)$ when $\theta \in R^1$ (and by $(\det I(\theta))^{1/2}$ in the multidimensional case), where I is the Fisher information for the statistical model as defined in Section 6.5. This is referred to as *Jeffreys' prior*. Note that Jeffreys' prior is dependent on the model.

Jeffreys' prior has an important invariance property. From Challenge 6.5.19, we have that, under some regularity conditions, if we make a 1–1 transformation of the real-valued parameter θ via $\psi = \Psi(\theta)$, then the Fisher information of ψ is given by

$$I\left(\Psi^{-1}(\psi)\right)\left(\left(\Psi^{-1}\right)'(\psi)\right)^2.$$

Therefore, the default Jeffreys' prior for ψ is

$$I^{1/2}\left(\Psi^{-1}(\psi)\right)\left|\left(\Psi^{-1}\right)'(\psi)\right|. \tag{7.4.1}$$

Now we see that, if we had started with the default prior $I^{1/2}(\theta)$ for θ and made the change of variable to ψ, then this prior transforms to (7.4.1) by Theorems 2.6.2 and 2.6.3. A similar result can be obtained when θ is multidimensional.

Jeffreys' prior often turns out to be improper, as the next example illustrates.

EXAMPLE 7.4.6 *Location Normal (Example 7.4.5 continued)*
In this case, Jeffreys' prior is given by $\pi(\theta) = \sqrt{n}/\sigma_0$, which gives the same posterior as in Example 7.4.5. Note that Jeffreys' prior is effectively a constant and hence the prior of Example 7.4.5 is equivalent to Jeffreys' prior. ∎

Research into rules for determining noninformative priors and the consequences of using such priors is an active area in statistics. While the impropriety seems counterintuitive, their usage often produces inferences with good properties.

Summary of Section 7.4

- To implement Bayesian inference, the statistician must choose a prior as well as the sampling model for the data.
- These choices must be checked if the inferences obtained are supposed to have practical validity. This topic is discussed in Chapter 9.
- Various techniques have been devised to allow for automatic selection of a prior. These include empirical Bayes methods, hierarchical Bayes, and the use of noninformative priors to express ignorance.
- Noninformative priors are often improper. We must always check that an improper prior leads to a proper posterior.

EXERCISES

7.4.1 Prove that the family $\{\text{Gamma}(\alpha, \beta) : \alpha > 0, \beta > 0\}$ is a conjugate family of priors with respect to sampling from the model given by Pareto(λ) distributions with $\lambda > 0$.

7.4.2 Prove that the family $\{\pi_{\alpha,\beta}(\theta) : \alpha > 1, \beta > 0\}$ of priors given by

$$\pi_{\alpha,\beta}(\theta) = \frac{\theta^{-\alpha} I_{[\beta,\infty)}(\theta)}{(\alpha - 1)\beta^{\alpha-1}}$$

is a conjugate family of priors with respect to sampling from the model given by the Uniform$[0, \theta]$ distributions with $\theta > 0$.

7.4.3 Suppose that the statistical model is given by

	$p_\theta(1)$	$p_\theta(2)$	$p_\theta(3)$	$p_\theta(4)$
$\theta = a$	1/3	1/6	1/3	1/6
$\theta = b$	1/2	1/4	1/8	1/8

and that we consider the family of priors given by

	$\pi_\tau(a)$	$\pi_\tau(b)$
$\tau = 1$	1/2	1/2
$\tau = 2$	1/3	2/3

and we observe the sample $x_1 = 1, x_2 = 1, x_2 = 3$.

(a) If we use the maximum value of the prior predictive for the data to determine the value of τ, and hence the prior, which prior is selected here?

(b) Determine the posterior of θ based on the selected prior.

7.4.4 For the situation described in Exercise 7.4.3, put a uniform prior on the hyperparameter τ and determine the posterior of θ. (Hint: Theorem of total probability.)

7.4.5 For the model for proportions described in Example 7.1.1, determine the prior predictive density. If $n = 10$ and $n\bar{x} = 7$, which of the priors given by $(\alpha, \beta) = (1, 1)$ or $(\alpha, \beta) = (5, 5)$ would the prior predictive criterion select for further inferences about θ?

7.4.6 Prove that when using an improper prior π, the posterior under π is proper if and only if the posterior under $c\pi$ is proper for $c > 0$, and then the posteriors are identical.

7.4.7 Determine Jeffreys' prior for the Bernoulli(θ) model and determine the posterior distribution of θ based on this prior.

7.4.8 Suppose we are sampling from a Uniform$[0, \theta], \theta > 0$ model and we want to use the improper prior $\pi(\theta) \equiv 1$.

(a) Does the posterior exist in this context?

(b) Does Jeffreys' prior exist in this context?

7.4.9 Suppose a student wants to put a prior on the mean grade out of 100 that their class will obtain on the next statistics exam. The student feels that a normal prior centered at 66 is appropriate and that the interval $(40, 92)$ should contain 99% of the marks. Fully identify the prior.

7.4.10 A lab has conducted many measurements in the past on water samples from a particular source to determine the existence of a certain contaminant. From their records, it was determined that 50% of the samples had contamination less than 5.3 parts per million, while 95% had contamination less than 7.3 parts per million. If a normal prior is going to be used for a future analysis, what prior do these data determine?

7.4.11 Suppose that a manufacturer wants to construct a 0.95-credible interval for the mean lifetime θ of an item sold by the company. A consulting engineer is 99% certain that the mean lifetime is less than 50 months. If the prior on θ is an Exponential(λ), then determine λ based on this information.

7.4.12 Suppose the prior on a model parameter μ is taken to be $N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are hyperparameters. The statistician is able to elicit a value for μ_0 but feels unable to do this for σ_0^2. Accordingly, the statistician puts a hyperprior on σ_0^2 given by $1/\sigma_0^2 \sim$ Gamma$(\alpha_0, 1)$ for some value of α_0. Determine the prior on μ. (Hint: Write $\mu = \mu_0 + \sigma_0 z$, where $z \sim N(0, 1)$.)

COMPUTER EXERCISES

7.4.13 Consider the situation discussed in Exercise 7.4.5.

(a) If we observe $n = 10, n\bar{x} = 7$, and we are using a symmetric prior, i.e., $\alpha = \beta$, plot the prior predictive as a function of α in the range $(0, 20)$ (you will need a statistical package that provides evaluations of the gamma function for this). Does this graph clearly select a value for α?

(b) If we observe $n = 10, n\bar{x} = 9$, plot the prior predictive as a function of α in the range $(0, 20)$. Compare this plot with that in part (a).

7.4.14 Reproduce the plot given in Example 7.4.3 and verify that the maximum occurs near $\lambda = 2.3$.

PROBLEMS

7.4.15 Show that a distribution in the family $\{N(\mu_0, \tau_0^2) : \mu_0 \in R^1, \tau_0^2 > 0\}$ is completely determined once we specify two quantiles of the distribution.

7.4.16 (*Scale normal model*) Consider the family of $N(\mu_0, \sigma^2)$ distributions, where μ_0 is known and $\sigma^2 > 0$ is unknown. Determine Jeffreys' prior for this model.

7.4.17 Suppose that for the location-scale normal model described in Example 7.1.4, we use the prior formed by the Jeffreys' prior for the location model (just a constant) times the Jeffreys' prior for the scale normal model. Determine the posterior distribution of (μ, σ^2).

7.4.18 Consider the location normal model described in Example 7.1.2.

(a) Determine the prior predictive density m. (Hint: Write down the joint density of the sample and μ. Use (7.1.2) to integrate out μ and do not worry about getting m into a recognizable form.)

(b) How would you generate a value $(X_1, \ldots, X_n)$ from this distribution?

(c) Are $X_1, \ldots, X_n$ mutually independent? Justify your answer. (Hint: Write $X_i = \mu + \sigma_0 Z_i$, $\mu = \mu_0 + \tau_0 Z$, where $Z, Z_1, \ldots, Z_n$ are i.i.d. $N(0, 1)$.)

7.4.19 Consider Example 7.3.2, but this time use the prior $\pi(\mu, \sigma^2) = 1/\sigma^2$. Develop the Gibbs sampling algorithm for this situation. (Hint: Simply adjust each full conditional in Example 7.3.2 appropriately.)

COMPUTER PROBLEMS

7.4.20 Use the formulation described in Problem 7.4.17 and the data in the following table

2.6	4.2	3.1	5.2	3.7	3.8	5.6	1.8	5.3	4.0
3.0	4.0	4.1	3.2	2.2	3.4	4.5	2.9	4.7	5.2

generate a sample of size $N = 10^4$ from the posterior. Plot a density histogram estimate of the posterior density of $\psi = \sigma/\mu$ based on this sample.

CHALLENGES

7.4.21 When $\theta = (\theta_1, \theta_2)$, the Fisher information matrix $I(\theta_1, \theta_2)$ is defined in Problem 6.5.15. The Jeffreys' prior is then defined as $(\det I(\theta_1, \theta_2))^{1/2}$. Determine Jeffreys' prior for the location-scale normal model and compare this with the prior used in Problem 7.4.17.

DISCUSSION TOPICS

7.4.22 Using empirical Bayes methods to determine a prior violates the Bayesian principle that all unknowns should be assigned probability distributions. Comment on this. Is the hierarchical Bayesian approach a solution to this problem?

7.5 Further Proofs (Advanced)

Derivation of the Posterior Distribution for the Location-Scale Normal Model

In Example 7.1.4, the likelihood function is given by

$$L(\theta \mid x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2}s^2\right).$$

The prior on (μ, σ^2) is given by $\mu \mid \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2)$ and $1/\sigma^2 \sim \text{Gamma}(\alpha_0, \beta_0)$, where μ_0, τ_0^2, α_0 and β_0 are fixed and known.

The posterior density of (μ, σ^{-2}) is then proportional to the likelihood times the joint prior. Therefore, retaining only those parts of the likelihood and the prior that depend on μ and σ^2, the joint posterior density is proportional to

$$\begin{aligned}
& \left\{\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2}s^2\right)\right\} \times \\
& \left\{\left(\frac{1}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau_0^2\sigma^2}(\mu-\mu_0)^2\right)\right\} \left\{\left(\frac{1}{\sigma^2}\right)^{\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma^2}\right)\right\} \\
= \; & \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{1}{2\tau_0^2\sigma^2}(\mu-\mu_0)^2\right) \left(\frac{1}{\sigma^2}\right)^{\alpha_0+n/2-1/2} \times \\
& \exp\left(-\left[\beta_0 + \frac{n-1}{2}s^2\right]\frac{1}{\sigma^2}\right) \\
= \; & \exp\left(-\frac{1}{2\sigma^2}\left[\left(n+\frac{1}{\tau_0^2}\right)\mu^2 - 2\left(n\bar{x}+\frac{\mu_0}{\tau_0^2}\right)\mu\right]\right) \times \\
& \left(\frac{1}{\sigma^2}\right)^{\alpha_0+n/2-1/2} \exp\left(-\left[\beta_0 + \frac{n}{2}\bar{x}^2 + \frac{1}{2\tau_0^2}\mu_0^2 + \frac{n-1}{2}s^2\right]\frac{1}{\sigma^2}\right) \\
= \; & \left(\frac{1}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\left(n+\frac{1}{\tau_0^2}\right)\left[\mu - \left(n+\frac{1}{\tau_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+n\bar{x}\right)\right]^2\right) \times \\
& \left(\frac{1}{\sigma^2}\right)^{\alpha_0+n/2-1} \exp\left(-\left[\begin{array}{c} \beta_0 + \frac{n}{2}\bar{x}^2 + \frac{1}{2\tau_0^2}\mu_0^2 + \frac{n-1}{2}s^2 \\ -\frac{1}{2}\left(n+\frac{1}{\tau_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+n\bar{x}\right)^2 \end{array}\right]\frac{1}{\sigma^2}\right)
\end{aligned}$$

From this, we deduce that the posterior distribution of (μ, σ^2) is given by

$$\mu \mid \sigma^2, x \sim N\left(\mu_x, \left(n+\frac{1}{\tau_0^2}\right)^{-1}\sigma^2\right)$$

and

$$\frac{1}{\sigma^2} \mid x \sim \text{Gamma}(\alpha_0 + n/2, \beta_x),$$

where

$$\mu_x = \left(n + \frac{1}{\tau_0^2}\right)^{-1} \left(\frac{\mu_0}{\tau_0^2} + n\bar{x}\right)$$

and

$$\begin{aligned} \beta_x &= \beta_0 + \frac{n}{2}\bar{x}^2 + \frac{\mu_0^2}{2\tau_0^2} + \frac{n-1}{2}s^2 - \frac{1}{2}\left(n + \frac{1}{\tau_0^2}\right)^{-1} \left(\frac{\mu_0}{\tau_0^2} + n\bar{x}\right)^2 \\ &= \beta_0 + \frac{n-1}{2}s^2 + \frac{1}{2}\frac{n(\bar{x} - \mu_0)^2}{1 + n\tau_0^2}. \end{aligned}$$

Derivation of $J(\theta(\psi_0, \lambda))$ for the Location-Scale Normal

Here we have that

$$\psi = \psi(\mu, \sigma^{-2}) = \frac{\sigma}{\mu} = \frac{1}{\mu}\left(\frac{1}{\sigma^2}\right)^{-1/2}$$

and

$$\lambda = \lambda(\mu, \sigma^{-2}) = \frac{1}{\sigma^2}.$$

We have that

$$\begin{aligned} \left|\det\begin{pmatrix} \frac{\partial\psi}{\partial\mu} & \frac{\partial\psi}{\partial\left(\frac{1}{\sigma^2}\right)} \\ \frac{\partial\lambda}{\partial\mu} & \frac{\partial\lambda}{\partial\left(\frac{1}{\sigma^2}\right)} \end{pmatrix}\right| &= \left|\det\begin{pmatrix} -\mu^{-2}\left(\frac{1}{\sigma^2}\right)^{-1/2} & -\frac{1}{2}\mu^{-1}\left(\frac{1}{\sigma^2}\right)^{-3/2} \\ 0 & 1 \end{pmatrix}\right| \\ &= \left|\det\begin{pmatrix} -\psi^2\lambda^{1/2} & -\frac{1}{2}\psi\lambda^{-1} \\ 0 & 1 \end{pmatrix}\right| = \psi^2\lambda^{1/2}, \end{aligned}$$

and so

$$J(\theta(\psi_0, \lambda)) = (\psi^2\lambda^{1/2})^{-1}.$$