Chapter 11

Bayesian Estimation and Inference

Objective: To study Bayesian analysis methods and procedures that are becoming very popular in building statistical models for real-world problems.

The Reverend Thomas Bayes

(*Source: http:en.wikipedia.org/wiki/Thomas_Bayes*)

The Reverend Thomas Bayes (1702–1761) was a Nonconformist minister. In the 1720s Bayes started working on the theory of probability. Even though he did not publish any of his works on mathematics during his lifetime, Bayes was elected a Fellow of the Royal Society in 1742. His famous work titled "Essay toward solving a problem in the doctrine of chances" was published in the *Philosophical Transactions of the Royal Society of London* in 1764, after his death. The paper was sent to the Royal Society by Richard Price, a friend of Bayes. Another mathematical publication on asymptotic series also appeared after his death.

11.1 INTRODUCTION

Bayesian procedures are becoming increasingly popular in building statistical models for real-world problems. In recent years, the Bayesian statistical methods have been increasingly used in scientific fields ranging from archaeology to computing. Bayesian inference is a method of analysis that combines information collected from experimental data with the knowledge one has prior to performing the experiment. Bayesian and classical (frequentist) methods take basically different outlooks toward statistical inference. In this approach to statistics, the uncertainties are expressed in terms of probabilities. In the Bayesian approach, we combine any new information that is available with the prior information we have, to form the basis for the statistical procedure. The classical approach to statistical inference that we have studied so far is based on the random sample alone. That is, if a probability distribution depends on a set of parameters θ, the classical approach makes inferences about θ solely on the basis of a sample $X_1, \ldots, X_n$. This approach to inference is based on the concept of a sampling distribution. To correctly interpret traditional inferential procedures, it is necessary to fully understand the notion of a sampling distribution. In this approach, we analyze only one set of sample values. However, we have to imagine what could happen if we drew a large number of random samples from the population. For example, consider a normal sample with known variance. We have seen that a 95% confidence interval for the population mean μ is given by the random interval $\left(\overline{X} - 1.96\sigma/\sqrt{n},\ \overline{X} + 1.96\sigma/\sqrt{n}\right)$. This means that when samples are repeatedly taken from the population, at least 95% of the random intervals contain the true mean μ. The classical inferential approach does not use any of the prior information we might have as a result of, say, our familiarity with the problem, or information from earlier studies. Scientists and engineers are faced with the problem that there is typically only a single data set, and they need to determine the value of the parameter at the time the data are taken. The basic question then is, "What is the best estimate of a parameter one can make from the data using one's prior information?" Statistical approaches that use prior knowledge, possibly subjective, in addition to the sample evidence to estimate the population parameters are known as Bayesian methods.

Bayesian statistics provides a natural method for updating uncertainty in the light of evidence. Data are still assumed to come from a distribution belonging to a known parametric family. However, the Bayesian outlook toward inference is founded on the subjective interpretation of probability. Subjective probability is a way of stating our belief in the validity of a random event. The following example will illustrate the idea. Suppose we are interested in the proportion of all undergraduate students at a particular university who take on out-of-campus jobs for at least 20 hours a week. Suppose we randomly select, say, 50 students from this university and obtain the proportion of

students who have out-of-campus jobs for at least 20 hours a week. Let us assume that the sample proportion is $30/50 = 0.6$. In a frequentist approach, all of the inferential procedures, such as point estimation, interval estimation, or hypothesis testing, are based on the sampling distribution.

That is, even though we are analyzing only one data set, it is necessary to have the knowledge of the mean, standard deviation, and shape of this sampling distribution of the proportion for the correct interpretation in classical inferential procedures. In the subjective interpretation of probability, the proportion of undergraduates who work on an out-of-campus job for at least 20 hours a week is assumed to be unknown and random. A probability distribution, called the prior, that represents our knowledge or belief about the location of this proportion before any data collected is used. For instance, the college placement office already may have an opinion on this proportion based on its earlier experience. The classical approach ignores this prior knowledge, whereas the Bayesian approach incorporates this knowledge with the current observed data to update the value of this proportion. That is, after the data are collected our opinion about the proportion may change. Using Bayes' rule, we will compute the posterior probability distribution for the proportion, based on our prior belief and evidence from the data. All of our inferences about the proportion are made by computing appropriate statistics of the posterior distribution.

The Bayesian approach seeks to optimally merge information from two sources: (1) knowledge that is known from theory or opinion formed at the beginning of the research in the form of a prior, and (2) information contained in the data in the form of likelihood functions. Basically, the prior distribution represents our initial belief, whereas the information in the data is expressed by the likelihood function. Combining prior distribution and likelihood function, we can obtain the posterior distribution. This expresses our revised uncertainty in light of the data. The main difference between the Bayesian approach and the classical approach is that in the Bayesian setting, the parameter is viewed as random variables, whereas the classical approach considers the parameter to be fixed but unknown. The parameter is random in the sense that we can assign to it a subjective probability distribution that describes our confidence about the actual value of the parameter.

Some of the reasons for Bayesian approaches are as follows: (1) Most Bayesian inferential conclusions are made conditional on the observed data. Unlike the traditional approach, one need not be concerned with data sets other than the one that is observed. There is no need to discuss sampling distributions using the Bayesian approach. Also, (2) from a Bayesian viewpoint, it is legitimate to talk about the probability that the proportion falls in a specific interval, say (0.2, 0.6), or the probability that a hypothesis is true. Too often, traditional inferential conclusions are misstated; for example, if a confidence interval computed from a sample for a parameter is (0.2, 0.6), it is common for the student to incorrectly state that the population parameter falls in the interval (0.2, 0.6) with probability at least 0.90. The Bayesian viewpoint provides a convenient model for implementing the scientific method. The prior probability distribution can be used to state initial beliefs about the population of interest, relevant sample data are collected, and the posterior probability distribution reflects one's new updated beliefs about the population parameter in light of the new data that were collected. All inferences about the parameter are made by computing appropriate summaries of the posterior probability distribution. Because of formidable theoretical and computational challenges, the Bayesian approach has found relatively limited use. Recent advances in Bayesian analysis combined with the

growing power of computers are making Bayesian methods practical and increasingly popular. The Markov chain Monte Carlo (MCMC) method described in Section 13.5 is one of the computationally intensive methods that is often useful in Bayesian estimation.

11.2 BAYESIAN POINT ESTIMATION

The cornerstone of Bayesian methodology is the Bayes theorem. It helps us to update our beliefs in the form of probability statements about the parameters after the sample has been taken. The conditional distribution of the parameters after observing the data is called the *posterior distribution* that integrates the prior and the sample information. Suppose we have two discrete random variables, X and Y. Then the joint probability function (pmf) can be written as $p(x, y) = p(x|y)p_Y(y)$, and the marginal probability density function of X is $p_X(x) = \sum_y p(x, y) = \sum_y p(x|y)p_Y(y)$. Then Bayes' rule for the conditional $p(y|x)$ is

$$p(y|x) = \frac{p(x, y)}{p_X(x)} = \frac{p(x|y)\ p_Y(y)}{p_X(x)} = \frac{p(x|y)\ p_Y(y)}{\sum_y p(x|y)\ p_Y(y)}.$$

The denominator in this expression is a fixed normalizing factor that ensures that the $\sum_y p(y|x) = 1$. If Y is continuous, the Bayes theorem can be stated as

$$p(y|x) = \frac{p(x|y)\ p_Y(y)}{\int p(x|y)\ p_Y(y)\ dy},$$

where the integral is over the range of values of y. These two equations are the Bayes formulas for random variables.

In Bayesian terminology, $p_Y(y)$ represents the probability statement of our *prior* belief, $p(x|y)$ is the probability of the data x given our prior beliefs, which is called the *likelihood*, and the updated probability $p(y|x)$ is the *posterior*. Because $p_X(x)$ (which is the likelihood accumulated over all possible prior values) is independent of y, we can express the posterior distribution as proportional ($\propto$) to [(likelihood) $\times$ (prior distribution)], that is,

$$p(y|x) \propto p(x|y)p(y).$$

We use the notation $f(x|\theta)$ to represent a probability distribution whose population parameter is considered to be a random variable. Now one of the problems is of finding a point estimate of the parameter θ (possibly a vector) for the population with distribution $f(x|\theta)$, given θ. Assume that $\pi(\theta)$ is the prior distribution of θ, which reflect the experimenter's prior belief about θ. We will not distinguish between the scalars and vectors, which will be clear based on the specific situation. Suppose that we have a random sample $X = (X_1, \ldots, X_n)$ of size n from $f(x|\theta)$. Then the posterior distribution can be written as

$$f(\theta|X_1, \ldots, X_n) = \frac{f(\theta, X_1, \ldots, X_n)}{f(X_1, \ldots, X_n)} = \frac{L(X_1, \ldots, X_n|\theta)\pi(\theta)}{f(X_1, \ldots, X_n)},$$

where $L(X_1, \ldots, X_n|\theta)$ is the likelihood function. Letting C represent all terms that do not involve θ (in this case, $C = 1/f(X_1, \ldots, X_n)$), we have

$$f(\theta\,|X_1, \ldots, X_n) = CL(X_1, \ldots, X_n\,|\theta)\pi(\theta),$$

For specific sample values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the foregoing equation can be written in a compact form as

$$f(\theta\,|x) \propto f(x\,|\theta)\pi(\theta), \quad \text{where} \quad x = (x_1, x_2, \ldots, x_n).$$

This can be expressed as

$$(\text{posterior distribution}) \propto (\text{prior distribution}) \times (\text{likelihood}).$$

The full result including the normalization can be written as

$$\text{(posterior distribution)} = [\text{(prior distribution)} \times \text{(likelihood)}] / \left[\sum(\text{prior} \times \text{likelihood})\right]$$

where the denominator is a fixed normalizing factor obtained by the likelihood accumulated over all possible prior values. We can now give a formal definition.

Definition 11.2.1 *The distribution of θ, given data $x_1, x_2, \ldots, x_n$, is called the* **posterior distribution**, *which is given by*

$$\pi(\theta\,|x) = \frac{f(x\,|\theta)\pi(\theta)}{g(x)}, \tag{11.1}$$

where $g\,(x)$ is the marginal distribution of X. The **Bayes estimate** *of the parameter θ is the posterior mean.*

The marginal distribution $g(x)$ can be calculated using the formula

$$g(x) = \begin{cases} \sum\limits_{\theta} f(x|\theta)\pi(\theta), & \text{in discrete case} \\ \int\limits_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta, & \text{in continuous case} \end{cases}$$

where $\pi(\theta)$ is the prior distribution of θ. Here, the marginal distribution $g\,(x)$ is also called the predictive distribution of X, because it represents our current predictions of the values of X taking into account both the uncertainty about the value of θ and the residual uncertainty about the random variable X when θ is known.

In a Bayesian setting, all the information about θ from the observed data and from the prior knowledge is contained in the posterior distribution, $\pi(\theta|x)$. In almost all practical cases, because we are combining our prior information with the information contained in the data, the posterior distribution provides a more refined estimation of θ than the prior. All inferences from Bayesian methods are based on the posterior probability distribution of the parameter θ. Using the explanation given later, we will take the *Bayes estimate* of a parameter as the posterior mean.

Furthermore, consider a Bayesian statistical inference problem where the parameter is a population proportion. In the Bernoulli trials, the population contains two types called "successes" and "failures." The proportion of successes in the population is denoted by θ. We take a random sample of size n from the population and observe s successes and f failures. The goal is to learn about the unknown proportion θ on the basis of these data.

In this situation, a model is represented by the population proportion θ. We do not know its value. In Chapter 5, we have seen that we could use the maximum likelihood estimator (MLE) for estimating θ, which did not use any prior knowledge we may have about θ. Note that the maximum likelihood estimate is broadly equivalent to finding the mode of the likelihood. In a Bayesian setting, we represent our beliefs about location of θ in terms of a prior probability distribution. We introduce proportion inference by using a discrete prior distribution for θ. We can construct a prior by specifying a list of possible values for the proportion θ, and then assigning probabilities to these values that reflect our knowledge about θ. Then the posterior probabilities can be computed using the Bayes theorem. The following example illustrates this concept.

Example 11.2.1

It is believed that cross-fertilized plants produce taller offspring than the self-fertilized plants. In order to obtain an estimate on the proportion θ of cross-fertilized plants that are taller, an experimenter observes a random sample of 15 pairs of plants that are exactly the same age. Each pair is grown in the same conditions with some cross-fertilized and the others self-fertilized. Based on previous experience, the experimenter believes that the following are possible values of θ and that the prior probability for each value of θ (prior weight) is $\pi(\theta)$.

θ:	0.80	0.82	0.84	0.86	0.88	0.90
$\pi(\theta)$:	0.13	0.15	0.22	0.25	0.15	0.10

From the experiment, it is observed that in 13 of 15 pairs, cross-fertilized is taller. Create a table with columns of the prior $\pi(\theta)$, likelihood of $L(X_1, X_2, \ldots, X_n|\theta)$ for different values of θ and for the given sample, prior times likelihood, and posterior probability of θ. Based on the posterior probabilities, what value of θ has the highest support? Also, find $E(\theta)$ based on the posterior probabilities.

Solution

The likelihood of obtaining 13 of 15 taller plants to the different prior values of π are given using the binomial pdf $\binom{15}{13}\theta^{13}(1-\theta)^2$. *For example, if the prior value of θ is 0.80, then the likelihood of θ given the sample is*

$$f(x|\theta) = \binom{15}{13}(0.8)^{13}(0.2)^2 = 0.2309.$$

Table 11.1

Prior values of θ	Prior $\pi(\theta)$	Likelihood of θ given sample	Prior times likelihood	Posterior probability of θ
0.80	0.13	0.2309	3.0017×10^{-2}	0.11029
0.82	0.15	0.2578	0.03867	0.14208
0.84	0.22	0.2787	6.1314×10^{-2}	0.22528
0.86	0.25	0.2897	7.2425×10^{-2}	0.2661
0.88	0.15	0.2870	0.4305	0.15817
0.90	0.10	0.2669	0.02669	0.098064
		Total	0.27217	$0.9998 \approx 1.0$

From Table 11.1 we obtain $\sum$(*prior* $\times$ *likelihood*) $= 0.27217$. *Hence, the normalized value corresponding to* $\theta = 0.80$ *is the posterior probability* $f(\theta|x)$, *which is equal to* $(0.030017/0.27217) = 0.11029$. *Now, we can obtain the table of posterior distribution of a proportion* π *using the discrete prior given in Table 11.1. When we substitute in Bayes' rule, the factor* $\binom{15}{13}$ *would be canceled. Hence, in the calculation of the likelihood function, we could have just used* $\theta^{13}(1-\theta)^2$ *instead of the full expression* $\binom{15}{13}\theta^{13}(1-\theta)^2$. *Thus, the Bayesian estimate of* θ *is*

$$\begin{aligned} E(\theta) &= (0.8)(0.11029) + (0.82)(0.14028) + (0.84)(0.22528) \\ &\quad + (0.86)(0.2661) + (0.88)(0.15817) + (0.9)(0.098065) \\ &= 0.84879 \approx 0.85. \end{aligned}$$

It may be noted that the MLE of θ *is* $13/15 = 0.867$.

In Example 11.2.1, the priors are called *informative priors*, because it favored certain values of θ; for example for the value $\theta = 0.86$, the prior value of $\pi(\theta)$ is 0.25, which is higher than all the rest of the values. If there was no information or no strong prior opinions, then we could select a *noninformative prior*, which would have assigned equal prior probability of 1/6 to each of the possible values of θ. A noninformative prior (also called a *flat* or *uniform prior*) provides little or no information. Based on the situation, noninformative priors may be quite disperse, may avoid only impossible values of the parameter, and oftentimes give results similar to those obtained by classical frequentist methods.

Example 11.2.2

Repeat the Example 11.2.1 using a noninformative prior, $\pi(\theta) = 1/6$, for each given value of θ.

Solution

Here $\pi(\theta) = \frac{1}{6}$ for each value of θ. See Table 11.2.

Table 11.2

Prior values of θ	Prior $\pi(\theta)$	Likelihood of θ given sample	Prior times likelihood	Posterior probability of θ
0.80	1/6	0.2309	3.8483×10^{-2}	0.14333
0.82	1/6	0.2578	4.2967×10^{-2}	0.16003
0.84	1/6	0.2787	0.04645	0.173
0.86	1/6	0.2897	4.8283×10^{-2}	0.17982
0.88	1/6	0.2870	4.7833×10^{-2}	0.17815
0.90	1/6	0.2669	4.4483×10^{-2}	0.16567
		Total	0.2685	1.0

The Bayesian estimate for the noninformative prior is

$$E(\theta) = (0.8)(0.14333) + (0.82)(0.16003) + (0.84)(0.173) + (0.86)(0.17982) + (0.88)(0.17815) + (0.9)(0.16567) = 0.85173.$$

It should be noted that because the choice of priors in Example 11.2.1 is only mildly informative, we do not see much difference in the values of Bayesian estimates. In general, it is difficult to construct an acceptable prior, because most often it has to be based on subjective experiences. Therefore, it is relatively easy to use a "noninformative" prior. For example, if we have no information on the values of proportion θ, then one type of standard "noninformative" prior is to take the proportion θ as one of the equally spaced values $0, 0.1, 0.2, \ldots, 0.9, 1$. We can assign for each value of θ the same probability, $\pi(\theta) = 1/11$. This prior is convenient and may work reasonably well when we do not have many data. It is fairly easy to construct a prior when there exists considerable prior information about the proportion of interest.

The posterior distribution gives us information regarding the likelihood of values of θ given sample data. Then the question is how to use this information to estimate θ. Instead of having explicit probabilities, the prior may be given through an assumed probability distribution. We illustrate the calculations involved to find the posterior distribution in the following example.

Example 11.2.3

Let X be a binomial random variable with parameters n and p. Assume that the prior distribution of p is uniform on [0,1]. Find the posterior distribution, $f(p|x)$.

Solution

Because X is binomial, the likelihood function is given by

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Because p is uniform on [0,1], $\pi(p) = 1,\ 0 \le p \le 1$.
Then the posterior distribution is given by

$$f(p|x) \propto f(x|p)\pi(p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n$$

which is the same as the likelihood.

This example illustrates that if the prior is noninformative (uniform), then the posterior is essentially the likelihood function. In the case where the prior and posterior are of the same functional form, we call it a *conjugate prior*. Bayesian inference becomes simpler when the prior density has the same functional form as the likelihood (which is the case for the conjugate prior) or when data are an independent sample from an exponential family (such as normal, Poisson, or binomial).

The following example demonstrates the method of finding posterior distribution for a continuous random variable.

Example 11.2.4

Suppose that X is a normal random variable with mean μ and variance σ^2, where σ^2 is known and μ is unknown. Suppose that μ behaves as a random variable whose probability distribution (prior) is $\pi(\mu)$ and is also normally distributed with mean μ_p and variance σ_p^2, both assumed to be known or estimated. Find the posterior distribution $f(\mu|x)$.

Solution

Using the Bayes theorem, we have

$$\begin{aligned} f(\mu|x) &= \frac{f(x|\mu)\pi(\mu)}{\int f(x|\mu)\pi(\mu)d\mu} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\mu-\mu_p)^2/2\sigma_p^2}}{\int \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\mu-\mu_p)^2/2\sigma_p^2} d\mu} \\ &= \frac{1}{2\pi\sigma\sigma_p} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_p)^2}{2\sigma_p^2}\right]}. \end{aligned} \tag{11.2}$$

Consider the exponential term in (11.2), namely, $\frac{(x-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_p)^2}{2\sigma_p^2}$.

$$\begin{aligned}
\frac{(x-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_p)^2}{2\sigma_p^2} &= \frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_p)^2}{\sigma_p^2}\right] \\
&= \frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_p^2}\right)\mu^2 - 2\left(\frac{\mu_p}{\sigma_p^2}+\frac{x}{\sigma^2}\right)\mu + \left(\frac{x^2}{\sigma^2}+\frac{\mu_p^2}{\sigma_p^2}\right)\right] \\
&= \frac{1}{2}\left[\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\mu^2 - 2\left(\frac{\mu_p}{\sigma_p^2}+\frac{x}{\sigma^2}\right)\mu + \left(\frac{x^2}{\sigma^2}+\frac{\mu_p^2}{\sigma_p^2}\right)\right] \\
&= \frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\mu^2 - 2\frac{\sigma^2\sigma_p^2}{\sigma_p^2+\sigma^2}\left(\frac{\mu_p}{\sigma_p^2}+\frac{x}{\sigma^2}\right)\mu\right. \\
&\quad \left. + \frac{\sigma^2\sigma_p^2}{\sigma_p^2+\sigma^2}\left(\frac{x^2}{\sigma^2}+\frac{\mu_p^2}{\sigma_p^2}\right)\right] \\
&= \frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\mu^2 - 2\left(\frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p + \frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x\right)\mu\right. \\
&\quad \left. + \left(\frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p + \frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x\right)^2\right] \\
&\quad + \frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\frac{x^2}{\sigma^2}+\frac{\mu_p^2}{\sigma_p^2} - \left(\frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x + \frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p\right)^2\right. \\
&= \frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\mu - \left(\frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p + \frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x\right)\right]^2 + \tilde{K},
\end{aligned}$$

where

$$\tilde{K} = \frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\frac{x^2}{\sigma^2}+\frac{\mu_p^2}{\sigma_p^2} - \left(\frac{\sigma^2}{\sigma^2+\sigma_p^2}\mu_p + \frac{\sigma_p^2}{\sigma^2+\sigma_p^2}x\right)^2\right].$$

From the foregoing derivation, we obtain

$$f(\mu|x) = Ke^{-\frac{1}{2}\frac{\sigma_p^2+\sigma^2}{\sigma^2\sigma_p^2}\left[\mu-\left(\frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p+\frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x\right)\right]^2},$$

where K *does not contain* μ.

This implies that the posterior density $f(\mu|x)$ *is the pdf of normal random variable with mean*

$$\left(\frac{\sigma^2}{\sigma_p^2+\sigma^2}\mu_p + \frac{\sigma_p^2}{\sigma_p^2+\sigma^2}x\right)$$

and variance

$$\frac{\sigma^2\sigma_p^2}{\sigma_p^2+\sigma^2}.$$

If we let $\tau_p = \frac{1}{\sigma_p^2}$ *and* $\tau = \frac{1}{\sigma^2}$, *then the posterior density can be rewritten as the pdf of normal random variable with mean* $\frac{1}{\tau_p+\tau}\left(\tau_p\mu_p + \tau x\right)$ *and variance* $\frac{1}{\tau_p+\tau}$.
As an example, suppose that $\mu_p = 100$, $\sigma_p = 15$, *and* $\sigma = 10$, $x = 115$. *Then* $f(\mu\,|x)$ *is the pdf of a normal random variable with*

$$Mean = \frac{100}{100+225}(100) + \frac{225}{100+225}(115) = 110.4$$

and

$$Variance = \frac{(100)(225)}{100+225} = 69.2.$$

■

11.2.1 Criteria for Finding the Bayesian Estimate

In the Bayesian approach to parameter estimation, we use both the prior and observations. This leads to an estimation strategy based on the posterior distribution. How do we know that the estimate thus obtained is "good"? To assess the quality of likely estimators, we use a loss function $L(\theta, a)$ that measures the loss incurred by using a as an estimate of θ. Here θ is the parameter being estimated (in real-world problems it is not known), and a is the estimate of θ. Then the "optimal" or "best" estimate $a = \hat{\theta}$ is chosen so as to minimize the expected loss $E[L(\theta, \hat{\theta})]$, where the expectation is taken over θ with respect to the posterior distribution $f(\theta\,|x)$. Here we mention two types of commonly used loss functions: quadratic and absolute error loss functions and the resulting estimates.

(1) A *quadratic (*or *squared error) loss function* is of the form $L(\theta, a) = (a-\theta)^2$. In this case,

$$\begin{aligned} E\left[L(\theta, a)\right] &= \int L(\theta, a)\, f(\theta|x_1, \ldots, x_n) d\theta \\ &= \int (a-\theta)^2 f(\theta|x_1, \ldots, x_n) d\theta. \end{aligned}$$

Differentiating with respect to a and equating to zero, we obtain

$$2\int (a-\theta)\, f\,(\theta\,|x_1, \ldots, x_n)\, d\theta = 0$$

This implies

$$a = \int \theta f\,(\theta\,|x_1, \ldots, x_n)\, d\theta.$$

This is the *posterior mean* (expected value) of θ, $E\,(\theta\,|x_1, \ldots, x_n)$. Hence the quadratic loss function is minimized by taking the estimate of θ, that is, $\hat{\theta}$, to be the posterior mean. In previous examples in this section, we used this value as the estimate $\hat{\theta}$. Note that what the quadratic loss function displays

is that if the estimate $\hat{\theta}$ and the true parameter θ are close to each other, the loss we expect is very small. Likewise, if the difference is larger, the expected loss in estimating θ with $\hat{\theta}$ is going to be large.

(2) An *absolute error loss function* is of the form $L(\theta, a) = |a - \theta|$. In this case,

$$\begin{aligned} E[L(\theta, a)] &= \int L(\theta, a) f(\theta | x_1, \ldots, x_n) d\theta \\ &= \int_{\theta=-\infty}^{a} (a - \theta) f(\theta | x_1, \ldots, x_n) d\theta \\ &\quad + \int_{\theta=a}^{\infty} (\theta - a) f(\theta | x_1, \ldots, x_n) d\theta \end{aligned}$$

Differentiating with respect to a and equating to zero, we obtain

$$\int_{\theta=-\infty}^{a} f(\theta | x_1, \ldots, x_n) \, d\theta - \int_{\theta=a}^{\infty} f(\theta | x_1, \ldots, x_n) \, d\theta = 0$$

The minimum loss is attained when the values of both integrals are equal to $\frac{1}{2}$. This can be achieved by taking $\hat{\theta}$ to be the *posterior median*.

The following can be considered as a general Bayesian procedure for point parameter estimation.

BAYESIAN PARAMETER ESTIMATION PROCEDURE

1. Consider the unknown parameter θ as a random variable.
2. Use a probability distribution(prior) to describe the uncertainty about the unknown parameter.
3. Update the parameter distribution using the Bayes theorem:

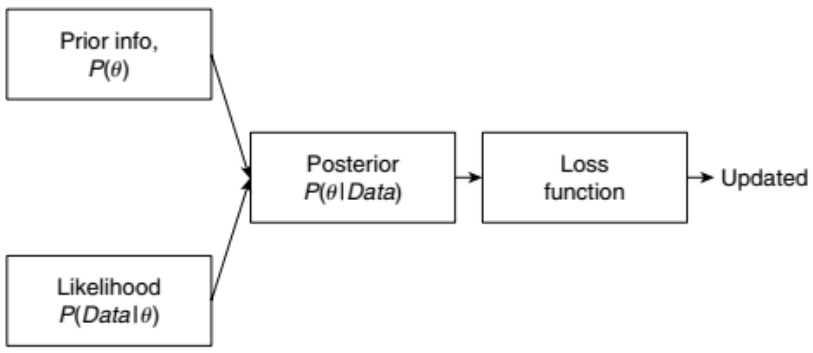
   $$P(\theta|Data) \propto P(\theta)P(Data|\theta),$$

   that is,

   $$(\text{posterior of } \theta) \propto (\textit{prior of } \theta).(\text{likelihood}).$$

4. The Bayes estimator of θ is set to be the expected value of the posterior distribution $P(\theta|Data)$ under quadratic loss function.
5. The Bayes estimator of θ is set to be posterior median under absolute error loss function.

From the procedure of Bayesian estimation, it is clear that a bad choice of prior may result in a bad estimate. Generally, if the priors are based on a previous and trustworthy sample, Bayesian estimation methods are desirable. A schematic figure of steps involved in the Bayesian estimate is given in Figure 11.1.



■ **FIGURE 11.1** Bayesian estimation procedure.

In this chapter, we use only the quadratic loss function unless it is explicitly stated otherwise. We also mention that this loss function is very popular because of its analytic tractability. We now derive Bayesian point estimates for some specific distributions.

Whereas uniform priors are useful in the noninformative situations, the beta family of distributions is one of the commonly taken informative priors. Distributions in the beta family take values in the interval $(0, 1)$. Recall that if $X \sim beta(\alpha, \beta)$, then the pdf of X is given by

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & 0 \le x < 1 \\ 0, & \text{otherwise, } \alpha > 0, \beta > 0. \end{cases}$$

The beta pdf can be written as

$$f(x) = Cx^{\alpha-1} (1-x)^{\beta-1} \propto x^{\alpha-1} (1-x)^{\beta-1},$$

where $C = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$. We also know that

$$E(X) = \frac{\alpha}{\alpha+\beta}, \quad \text{and} \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2((\alpha+\beta+1)}.$$

Example 11.2.5

Let $X_1, \ldots, X_n$ be a sample from geometric distribution with parameter $p, 0 \le p \le 1$. Assume that the prior distribution of p is beta with $\alpha = 4$, and $\beta = 4$.

(a) Find the posterior distribution of p.

(b) Find the Bayes estimate under quadratic loss function.

Solution

(a) *Because p is $Beta(4, 4)$, the prior density is*

$$\frac{\Gamma(8)}{\Gamma(4)\Gamma(4)} p^3 (1-p)^3 = 140 p^3 (1-p)^3.$$

Because the r.v.'s $X_i's$ have geometric distribution with parameter p, the likelihood is given by

$$L(X_1, \ldots, X_n \,|\theta) = \prod_{i=1}^{n} p\,(1-p)^{x_i - 1} = p^n (1-p)^{\sum_{i=1}^{n} x_i - n}.$$

The product of the likelihood function and the prior is given by

$$p^n (1-p)^{\sum_{i=1}^{n} x_i - n} \left[140 p^3 (1-p)^3\right] = 140 p^{n+3} (1-p)^{\sum_{i=1}^{n} x_i - n + 3}.$$

Because, (posterior of p) $\propto$ (prior of p) . (likelihood), rewriting the normalizing constant in the denominator of Equation (11.1) as C, and letting $C_1 = 140C$, the posterior distribution (because $\alpha - 1 = n + 3$, and $\beta - 1 = \sum_{i=1}^{n} x_i - n + 3$) is $Beta\left(n + 4, \sum_{i=1}^{n} x_i - n + 4\right)$.

(b) *Recall that for a $Beta(\alpha, \beta)$ random variable, the mean is $[\alpha/(\alpha + \beta)]$. Because the Bayes estimate is the posterior mean, the mean of $Beta\left(n + 4, \sum_{i=1}^{n} x_i - n + 4\right)$ is*

$$\frac{n+4}{\left[\sum_{i=1}^{n} x_i - n + 4\right] + (n+4)} = \frac{n+4}{\sum_{i=1}^{n} x_i + 8}$$

Note that for large n, the Bayes estimate is approximately $n/\sum_{i=1}^{n} x_i$, which is the MLE of p.
In general, for a Bernoulli random variable with unknown probability of success p in [0,1], the usual conjugate prior is the beta distribution, where the parameters of the beta distribution are chosen to reflect any prior information that we have.
We will follow the idea of the previous example in a binomial experiment of tossing a coin.

■

Example 11.2.6

Suppose we are flipping a biased coin, where the probability of heads p could be any value between 0 and 1. Given a sequence of toss samples $x_1, x_2, \ldots, x_n$, we want to estimate $P(H) = p$. We may have two sources of information: our prior belief, which we will express as a beta distribution, and the data, which could come from counts of heads x in $n = 20$ independent flips of the coin, say $x = 13$. Suppose that in six prior tosses, we observed three heads and three tails, which lead us to believe that the value of p is near 0.5. Obtain the posterior distribution of p.

Solution

Here our prior belief or assumption can be written in terms of beta distribution as

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} p^{\alpha - 1} (1-p)^{\beta - 1}$$

where $\alpha = 4$ and $\beta = 4$. That is (noting $\Gamma(n) = (n-1)!$)

$$\pi(p) = \frac{7!}{(3!)(3!)} p^3 (1-p)^3.$$

Hence, $\pi(p) \propto p^3(1-p)^3$. *Because the mean of a beta distribution is* $\alpha/(\alpha+\beta)$ *and the variance is* $\alpha\beta/\left((\alpha+\beta)^2(\alpha+\beta+1)\right)$, *for the prior,*

$$Mean(p) = \frac{4}{4+4} = 0.5,$$

and

$$Var(p) = \frac{(4)(4)}{(4+4)^2(4+4+1)} = 0.028.$$

Let X *denote the number of heads in 20 flips of this coin. Then* X *has a binomial distribution, and the pmf is given by*

$$f(x|p) = \binom{20}{x} p^x(1-p)^{20-x}, \quad x = 0, 1, \ldots, 20.$$

This we can write as

$$f(x|p) \propto p^x(1-p)^{20-x}.$$

In the 20 flips we have observed 13 heads. Then fix $x = 13$, *and we are interested in the likelihood, which is the relative value of the function at different values of* p:

$$f(13|p, 20) \propto p^{13}(1-p)^7.$$

The posterior probability of p, *given* $x = 13$, *is*

$$\begin{aligned}\pi(p|x=13) &\propto f(x|p)\pi(p)\\ &= \left(p^{13}(1-p)^{20-13}\right)p^3(1-p)^3\\ &= p^{16}(1-p)^{10}.\end{aligned}$$

Thus, the posterior is a beta distribution with $\alpha = 17$ *and* $\beta = 11$. *Consequently, we can now obtain the mean and variance of* p *as*

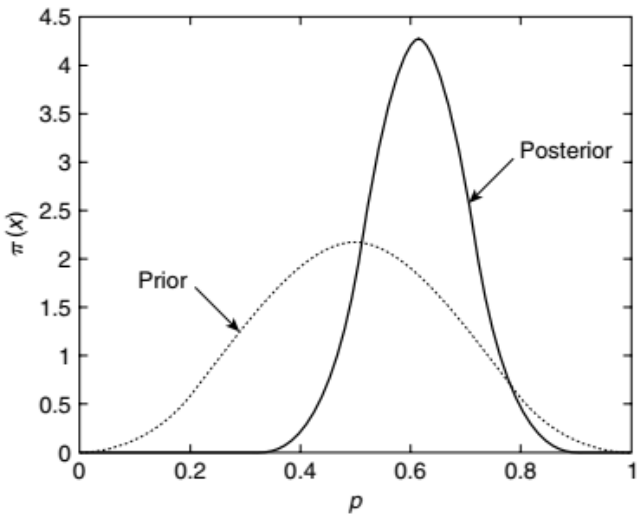
$$Mean(p) = \frac{17}{17+11} = 0.607$$

and

$$Var(p) = \frac{(17)(11)}{(17+11)^2(17+11+1)} = 0.008.$$

Note that the prior was beta distribution with mean 0.5 and variance 0.028. Figure 11.2 gives the prior and posterior densities.

Note that if we had ignored the prior and just took the point estimation, then the MLE of p *is* $MLE(p) = \hat{p} = \frac{13}{20} = 0.65$. *Compare this with the Bayesian estimate of* $p = 0.607$. *Because* $Beta(1, 1)$ *is the* $Uniform\,[0, 1]$,



■ **FIGURE 11.2** Prior and posterior distributions for the proportions.

the method of the previous example can be used for noninformative priors. The method could also be used in many applications. For example, suppose p represents the proportion of infected individuals in a population, and x is the number of infected individuals in a sample of size n. Then with a noninformative prior, we can show that the posterior of p is $Beta(x+1,\ n-x+1)$. This type of setting can be used for estimating the true proportion of infected individuals in the population.

■

Example 11.2.7

Suppose for the past million days we have been predicting whether the sun will rise the next morning or not. Each evening we say that the sun will rise the next morning ($\hat{R}$), and we were right (R) all these days. Suppose on the 10^6 evenings we predicted that the sun will rise on the next day. What is the probability that the sun will rise the next day?

Solution

The problem can be cast in the following table form.

1	2	...	10^6	10^6+1
$\hat{R}$	$\hat{R}$	...	$\hat{R}$	$\hat{R}$
R	R	...	R	?

$P(R|\hat{R}) = 1$ *if we use the frequency method of estimation (for example the MLE). Let us now consider the Bayes method. Suppose the prior is uniform on [0,1]. That is,*

$$\pi(p) = \begin{cases} 1, & \text{if } 0 \le p \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose we predict n times and we succeed x times. Then

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The joint pdf is given by

$$\begin{aligned} f(x, p) &= f(x|p)\pi(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n; \quad 0 \le p \le 1. \end{aligned}$$

By the Bayes theorem, the posterior pdf $\pi(p|x)$ is

$$\begin{aligned} \pi(p|x) &= \frac{f(x|p)\pi(p)}{\int_0^1 f(x|p)\pi(p)dp} \\ &= K(n, x)p^x(1-p)^{n-x}, \quad 0 \le p \le 1, \quad 0 \le x \le n, \end{aligned}$$

which is a beta probability distribution. Recall that the beta density is given by

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

and $E(Y) = \frac{\alpha}{\alpha+\beta}$. Thus,

$$E[\pi(p|x)] = \frac{x+1}{(x+1)+(n-x)+1} = \frac{x+1}{n+2}.$$

In our example, $x = 10^6$, $n = 10^6$, which implies that the posterior mean is given by

$$\hat{p}_\beta = \frac{10^6+1}{10^6+2} \approx 1.$$

Example 11.2.8

Let $X_1, X_2, \ldots, X_n$ be $N(\mu, \sigma^2)$ random variables with prior $\pi(\mu)$ having $N(\mu_0, \sigma_0^2)$ distribution with known σ^2.

(a) Obtain the posterior distribution of μ.

(b) Suppose it is known from past experience that the weight loss for a particular combination of diet and exercise program (if followed for a month) is normally distributed with mean 10 lb and standard deviation of 2 lb. A random sample of five persons who went through this program for a month produced the following weight loss in pounds:

14 8 11 7 11

What is the point estimate of the mean, μ? Assume $\sigma^2 = 4$.

Solution

(a) *Because* $\pi(\mu) \sim N(\mu_0, \sigma_0^2)$, $\pi(\mu) \propto \exp\left[(\mu - \mu_0)^2/\sigma_0^2\right]$ *and we omit the terms that do not depend on* μ. *We have from the data* $x = (x_1, \ldots, x_n)$, *the likelihood function,*

$$L(x_1, \ldots, x_n|\mu) = f(x|\mu) \propto \prod_{i=1}^{n} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{-\sum_{i=1}^{n}\left[(x_i - \mu)^2/2\sigma^2\right]\right\},$$

where μ *is determined by the posterior distribution. The product of the likelihood function and the prior gives the posterior, which is obtained (after some algebra) as follows:*

$$f(\mu|x) \propto \pi(\mu) f(x|\mu) \propto \exp\left[-(\mu - \mu_1)^2/2\sigma_1^2\right]$$

where

$$\mu_1 = \frac{\frac{n}{\sigma^2}\overline{x} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

and

$$\sigma_1^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}.$$

Thus, the posterior distribution of μ *is* $N(\mu_1, \sigma_1^2)$.

(b) *Note that the sample mean* $\overline{x} = 10.2$ lb, *and sample standard deviation* $s = 2.77$ lb. *Now from part (a), the posterior distribution of* μ *is normal with mean*

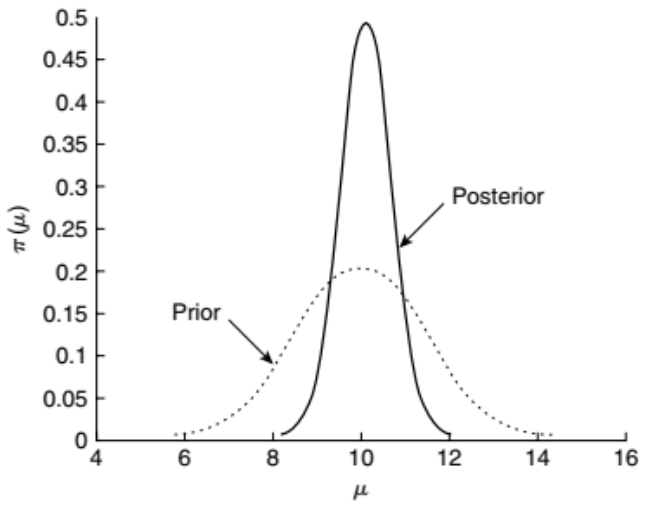
$$\mu_1 = \frac{\frac{n}{\sigma^2}\overline{x} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{5}{2^2}(10.2) + \frac{1}{2^2}(10)}{\frac{5}{2^2} + \frac{1}{2^2}} = 10.167$$

and variance

$$\sigma_1^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{1}{\frac{5}{2^2} + \frac{1}{2^2}} = 0.66667.$$

Thus, the point estimate of μ is the posterior mean, 10.167. Figure 11.3 represents the prior and posterior densities of μ.

Sometimes, the inverse of variance in the normal distribution is called the *precision* of the normal distribution and denoted by $\tau = 1/\sigma^2$. Also note that in part (a) of the previous example, if the prior variance $\sigma_0^2 \to \infty$, then the prior flattens out, $\pi(\mu) \propto c$, a constant. This basically amounts to saying that prior information on μ decreases, that is, all μ are equally probable. This corresponds to a noninformative prior. Also, in this case as $\sigma_0^2 \to \infty$, $\sigma_1^2 \to \frac{\sigma^2}{n}$ and $\mu_1 \to \overline{x}$. Hence, in the limit



■ **FIGURE 11.3** Prior and posterior densities of μ.

(i.e., for noninformative priors), the posterior $f(\mu|x)$ will have an $N(\overline{x}, \sigma^2/n)$ distribution, which is exactly the same inference as in classical statistics.

In Bayesian inference problems, one of the questions is, which will have relatively more influence, prior or likelihood? As we observe a large amount of data, it can be shown that the posterior distribution is almost exclusively determined by the data. That is, asymptotically, observed data will have a larger influence compared to the choice of prior, and thus the prior will be irrelevant. Hence, we can make the following general observations. If the prior is noninformative and we have a large data set, then we can expect that the likelihood will have greater influence. Whereas, if we have a small data set and an informative prior, then the prior will have a larger influence on the updated posterior distribution. Bayesian estimators are more complicated to compute than calculating the maximum likelihood estimates in simple cases. However, in complex settings Bayesian statistics are often relatively easier to compute.

One of the problems in using Bayesian analysis is choosing an appropriate prior. There are no specific rules available for this purpose. For instance, the following priors are commonly used in the literature. If data are in $[0,1]$, we could use uniform or beta distribution. If the data are in $[0, \infty)$, normal (with nonnegative and relatively large μ), gamma, or log-normal distributions are used. If the data are in $(-\infty, \infty)$, normal or t-distributions are commonly used.

EXERCISES 11.2

11.2.1. Suppose in a casino, two kinds of dice are used, one kind of which 98% are fair, and 2% are loaded such that five comes up 60% of the time and the rest of the numbers are equally probable. We pick a die at random and roll it three times. We get three consecutive fives. What is the probability that the die is loaded?

11.2.2. It is believed that cross-fertilized plants produce taller offspring than self-fertilized plants. In order to obtain an estimate on the proportion θ of cross-fertilized plants that are taller, an experimenter observes a random sample of 15 pairs of plants exactly the same age, with each pair grown in the same conditions with one cross-fertilized and the other self-fertilized. Based on previous experience, the experimenter believes that the following are possible values of π and prior probabilities for each value (prior weight), $\pi(\theta)$:

θ:	0.80	0.82	0.84	0.86	0.88	0.90
$\pi(\theta)$:	0.03	0.40	0.22	0.15	0.15	0.05

From the experiment, it is observed that in 13 of 15 pairs, the cross-fertilized is taller.

(a) Create a table with columns for prior, likelihood of θ given sample, prior times likelihood, and posterior probability of θ. Based on the posterior probabilities, what value of θ has the highest support? Also, find $E(\theta)$ based on the posterior probabilities.

(b) Redo part (a) with a completely noninformative prior, that is, take the prior for the proportion θ as one of the equally spaced values $0, 0.1, 0.2, \ldots, 0.9, 1$. Also assign for each value of θ the same probability, $\pi(\theta) = 1/11$.

(c) Calculate the MLE of θ and compare it with the Bayesian estimate.

11.2.3. Consider the problem of estimating p in a binomial distribution. Let X be number of successes in a sample of size n.

(a) Let the prior distribution of p be given by $Beta(3,1)$, that is

$$\pi(p) = \begin{cases} 3p^2, & 0 < p < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the posterior distribution of p.

$$\left[\text{Hint}: f(x|p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise.} \end{cases} \right]$$

(b) Let the prior distribution of p be given by $Beta(a,b)$ (that is, $\pi(p) \propto p^{a-1}(1-p)^{b-1}$. Find the posterior distribution of p.

11.2.4. A biased coin is tossed n times. Let x_i be 1 if the ith toss is heads and 0 if it is tails. Assume a noninformative prior, $p(\theta) = 1,\ 0 \le \theta \le 1$. Let t be the number of heads obtained. Show that the posterior distribution of θ is $Beta(t+1, n-t+1)$.

11.2.5. Let $X_1, X_2, \ldots, X_n$ be exponential random variables with parameter λ. Let the prior $\pi(\lambda)$ be exponentially distributed with parameter μ, which is a fixed and known constant.

(a) Show that the posterior distribution of λ is $Gamma\left(1 + \sum_{i=1}^{n} x_i,\ n+1\right)$.

(b) Obtain the Bayes estimate of λ.

11.2.6. Let $X_1, X_2, \ldots, X_n$ be Poisson random variables with parameter λ. Assume that λ has a $Gamma(\alpha, \beta)$ prior.

(a) Compute the posterior distribution of λ.
(b) Obtain the Bayes estimate of λ.
(c) Compare the MLE of λ with the Bayes estimate of λ.
(d) Which of the two estimates is better? Why?

11.2.7. Let $X_1, X_2, \ldots, X_n$ be Poisson random variables with parameter λ. Assume that λ has an exponential distribution with $\theta = 1$ prior.
(a) Compute the posterior distribution of λ.
(b) Show that the Bayes estimate of λ is $Gamma\left(\left(\sum_{i=1}^{n} x_i + 1\right), (n+1)\right)$.

11.2.8. It is known that a certain disease has affected 10% of a population. In a random sample of 50 patients typical of the disease group who are exposed to a new treatment, we observe that 12 patients were hospitalized in a year. Let μ be the rate of population that need hospitalization. Assume that

$$\mu \sim Gamma\,(0.1, 2) \quad \text{and} \quad f(x|\mu) \sim Poi\,(50\mu)\,.$$

Given that 0.24 is an observation from $f(x|\mu)$, find the Bayesian estimator of μ (that is, obtain $E(\mu|x)$).

11.2.9. Let $X_1, \ldots, X_n$ be an $N(\mu, 2)$ random sample with prior $\pi(\mu)$ having $N(0, \sigma^2)$ distribution with known σ^2. Obtain the posterior distribution of μ.

11.2.10. Let $X_1, \ldots, X_n$ be an $N(\mu, 1)$ random sample with prior $\pi(\mu)$ having the pdf $\left[1/\pi\left(1+\mu^2\right)\right]$. Show that the posterior

$$\pi(\mu\,|x) \propto \exp\left\{-\frac{n\,(\mu - \overline{x})^2}{2}\right\} \times \frac{1}{1+\mu^2}.$$

11.3 BAYESIAN CONFIDENCE INTERVAL OR CREDIBLE INTERVALS

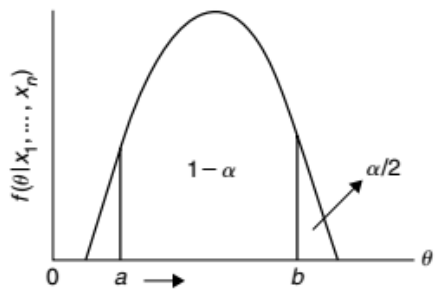
In this section, we want to study the question, "Can we construct an interval where we are confident that the interval contains the unknown true value of θ?" We have seen how in many situations it may be preferable to use an interval estimate instead of a point estimate for a population parameter θ. Such intervals in classical statistics were called confidence intervals. We can extend the concept of interval estimation to a Bayesian setting. The Bayesian analog of a confidence interval is called a credible interval and is defined as follows.

Definition 11.3.1 *A* $100(1-\alpha)\%$ **credible interval** *for* θ *is an interval* (a, b) *such that*

$$P(a \le \theta \le b\,|x_1, \ldots, x_n\,) \ge (1-\alpha)\ 100\%$$

Here α *is given small positive number between 0 and 1, and* $x_1, \ldots, x_n$ *are the sample values.*

Note that we read this definition backwards, that is, we are at least $(1-\alpha)\ 100\%$ confident that the true value of θ is between a and b, given the sampled information.



■ **FIGURE 11.4** Credible interval for θ.

Because the conditional distribution of θ given $X_1, \ldots, X_n$ is actually a probability distribution, it makes sense to talk about the probability that θ is in the interval (a, b). Once we have observed data, the credible interval is fixed while θ is random. This is in contrast to the classical confidence interval where the interval is random but θ is a fixed parameter. In the classical case, we would say, "In the long run, $100(1-\alpha)\%$ of all such intervals will contain the true parameter θ." In the Bayesian approach, we would say, "The probability is at least $100(1-\alpha)\%$ that θ lies within the specified interval (a, b)."

As in the classical case, it would be desirable to minimize the length of the credible interval. This entails choosing only those points with highest values in the density of $f(\theta|x_1, \ldots, x_n)$, as shown in Figure 11.4.

Definition 11.3.1 can be rephrased as follows using the posterior distribution of θ.

Definition 11.3.2 *A* $100(1-\alpha)\%$ **credible interval** *for* θ *is an interval* (a, b) *such that*

1. $\int_a^b f(\theta|x_1, \ldots, x_n)\, d\theta \geq 1-\alpha$, *if* θ *is continuous, and the posterior pdf of* θ *is* $f(\theta|x_1, \ldots, x_n)$.
2. $\sum_a^b f(\theta|x_1, \ldots, x_n) \geq 1-\alpha$, *if* θ *is discrete.*

We will now give some examples for computing credible intervals.

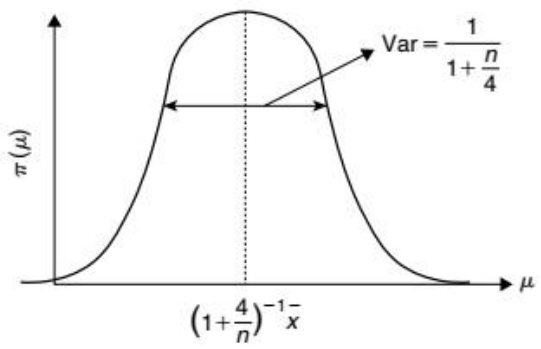
Example 11.3.1

Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ with $\sigma^2 = 4$. Suppose the prior pdf of μ is $N(0, 1)$, that is, $\pi(\mu) \sim N(0, 1)$. Find a 95% credible interval for μ.

Solution

We have seen from Example 11.2.8 that the posterior distribution of μ *given* $x_1, \ldots, x_n$ *is normally distributed with*

$$Mean = \frac{1}{1+\frac{4}{n}}\bar{x}$$



■ **FIGURE 11.5** Posterior distribution of μ.

and

$$Variance = \frac{1}{1+\frac{n}{4}}.$$

Figure 11.3 represents the posterior distribution of μ.

To find the 95% credible interval for μ, we have to find two numbers a and b such that

$$P(a \leq X \leq b) = 0.95$$

where

$$X \sim N\left(\mu = \frac{\overline{x}}{1+\frac{4}{n}}, \sigma^2 = \frac{1}{1+\frac{n}{4}}\right).$$

We choose a to be $-b$ (b is positive). Using z-scores, we get (X is continuous),

$$P\left(-z_{\alpha/2} < \frac{\mu - \frac{1}{1+\frac{4}{n}}\overline{x}}{\sqrt{\frac{1}{1+\frac{n}{4}}}} < z_{\alpha/2}\right) = 1-\alpha$$

which can be rearranged as

$$P\left(\frac{1}{1+\frac{4}{n}}\overline{x} - \frac{1}{\sqrt{1+\frac{n}{4}}}z_{\alpha/2} < \mu < \frac{1}{1+\frac{4}{n}}\overline{x} + \frac{1}{\sqrt{1+\frac{n}{4}}}z_{\alpha/2}\right) = 1-\alpha.$$

Thus, a 95% credible interval for μ is

$$\left(\frac{1}{1+\frac{4}{n}}\overline{x} - \frac{1}{\sqrt{1+\frac{n}{4}}}z_{\alpha/2}, \frac{1}{1+\frac{4}{n}}\overline{x} + \frac{1}{\sqrt{1+\frac{n}{4}}}z_{\alpha/2}\right).$$

For convenience, we summarize this procedure in the following steps.

BAYESIAN CREDIBLE INTERVAL PROCEDURE

1. Consider θ as a random variable with prior pdf (or pmf) $\pi(\theta)$.
2. Update the prior distribution $\pi(\theta)$ using the Bayes theorem. That is find the posterior distribution of θ by the formula

$$\pi(\theta\,|data\,) = \begin{cases} \dfrac{f(data|\theta)\pi(\theta)}{\int f(data|\theta)\pi(\theta)d\theta}, & \text{if continuous} \\ \dfrac{f(data|\theta)\pi(\theta)}{\sum f(data|\theta)\pi(\theta)}, & \text{if discrete.} \end{cases}$$

3. Find two numbers a and b such that

$$\int_a^b \pi(\theta\,|data)d\theta \geq 1-\alpha, \quad \text{if continuous}$$

$$\sum_{\theta=a}^{b} \pi(\theta\,|data) \geq 1-\alpha, \quad \text{if discrete.}$$

Note: The numbers a and b are found such that

$$\int_{-\infty}^{a} \pi(\theta\,|data)d\theta = \alpha/2, \quad \text{if continuous}$$

$$\sum_{\theta\leq a} \pi(\theta\,|data) = \alpha/2, \quad \text{if discrete}$$

and

$$\int_b^{\infty} \pi(\theta\,|data)d\theta = \alpha/2, \quad \text{if continuous}$$

$$\sum_{\theta\geq b} \pi(\theta\,|data) = \alpha/2, \quad \text{if discrete.}$$

4. The $(1-\alpha)100\%$ credible interval for θ is the interval (a, b).

In the discrete case, an easy way of finding a credible interval of smallest length is to arrange the values of θ from most likely to least likely (that is, in the order of the magnitude of the posterior probabilities), and then put values of θ into the interval until the cumulative posterior probability of the set exceeds $(1-\alpha)100\%$. Such an interval is called a highest posterior density (HPD) interval. It can be shown that the HPD interval always exists, and it is unique, so long as for all intervals of probability $(1-\alpha)$, the posterior density is never uniform in any interval of values of θ.

Example 11.3.2

For the data of Example 11.2.1, find a 90% credible interval for θ.

Solution

Arranging the values of θ from most likely to least likely, we have Table 11.3. Looking at the "cumulative probability" column, we see that the probability that θ is in the set $\{0.86, 0.84, 0.88, 0.82, 0.80\}$ is 0.90192. So this set is a 90% probability (or credible) interval for θ.

Table 11.3

Prior values of θ	Posterior probability of θ	Cumulative probability
0.86	0.2661	0.2661
0.84	0.22528	0.49138
0.88	0.15817	0.64955
0.82	0.14208	0.79163
0.80	0.11029	0.90192
0.90	9.8064×10^{-2}	0.99984

EXERCISES 11.3

11.3.1. **(a)** Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ with $\sigma^2 = 9$. Suppose the prior pdf of μ is $N(0, 1)$; that is $\pi(\mu) \sim N(0, 1)$. Find a 95% credible interval for μ.

(b) The following is a set of random data from a normal distribution with variance 9.

0.92	1.05	5.53	3.64	−4.47	−2.60	0.71	−3.66	1.38	3.87
7.42	1.76	0.01	2.69	1.54	3.97	1.34	−1.63	−1.24	−4.78

Using the results of part (a), compute a 95% credible interval for μ, interpret its meaning, and state any assumptions you have made.

11.3.2. Suppose that a person believes that his last year's weight was normally distributed with mean of 165 lb and standard deviation of 5 lb. That is, the prior pdf of μ is $N(165, 25)$, or $\pi(\mu) \sim N(165, 25)$. He expects his current weight X is normally distributed with mean μ and standard deviation 7 lb. Following are 10 random measurements (in pounds) from this year.

176	165	180	172	175
179	166	177	184	183

Find a 95% credible interval for μ.

11.3.3. It is known that a certain disease affects 10% of a population. In a random sample of 50 patients in the disease group who are exposed to a new treatment, we observe that 12 patients were hospitalized in a year. Let μ be the population rate that needs hospitalization in a year. Assume μ has a $Gamma\,(0.1, 2)$ prior. Let $\mu \sim Gamma\,(0.1, 2)$ and $f\,(x|\,\mu\,) \sim Poi\,(50\mu)$. Given that $x = 0.24$ is an observation of X, find 95% credible internal for μ. Obtain a Bayesian credible interval for μ. (If X is the number of patients admitted in a year, assume $X \sim Poi\,(50\mu)$, the Poisson approximation of the binomial.) How can we improve on this estimate?

11.3.4. For an upcoming congressional election, suppose we want to estimate the amount of support for a particular candidate in a district. By previous experience and voter registration data, we can assume that the prior distribution of the proportion of support, p, is a beta distribution with $\alpha = 10$, and $\beta = 8$ (i.e., $\pi\,(p) \sim Beta\,(10, 8)$). We conducted a survey of 1000 randomly selected voters, of whom 600 support the candidate. Obtain a 95% credible interval for p. What will happen to the credible interval if we reduce the confidence interval? What will happen to the 95% credible interval if we increase the sample size?

11.3.5. It is recommended that the daily intake of sodium be 2400 mg per day. From a previous study on a particular ethnic group, the prior distribution of sodium intake is believed to be normal with mean 2700 mg and standard deviation 250 mg. If a recent survey for this group resulted in a mean of 3000 mg and standard deviation of 300 mg, obtain a 95% credible interval for the mean intake of sodium for this ethnic group.

11.3.6. Suppose we have a coin (not necessarily balanced) with p being the probability of heads. Assume a uniform prior for p. Suppose in 20 tosses of this coin, we obtained 12 heads. Obtain a 90% credible interval for p.

11.3.7. Suppose that in a particular telephone exchange, the number of calls received per minute has a Poisson distribution with parameter λ. Assume an exponential prior for λ with parameter 2. Suppose this exchange had received 25 calls in five minutes. Obtain a 95% credible interval for λ.

11.4 BAYESIAN HYPOTHESIS TESTING

The Bayesian approach to hypothesis testing for simple hypotheses is pretty straightforward. Deciding between two hypotheses for a given set of data x reduces to computing their posterior probabilities. If an explicit loss function is available, the Bayes rule is chosen to minimize the expected value of the loss function with respect to the posterior distribution. In the absence of a loss function, the probabilities of type I and type II errors are of little interest to the Bayesian.

In the classical hypothesis testing, we test a null hypothesis (denoted by H_0) against an alternative hypothesis (denoted by H_1 or H_a). The test procedure is based on controlling the two types of errors—type I and type II. The classical test procedures limit the type I error to α and minimize the type II error. If the type II error is unacceptably high, it is reduced by increasing the sample size.

In the Bayesian approach, the problem of deciding between the null and alternative is rather straightforward. Consider the problem of hypothesis testing with

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1 \tag{11.3}$$

where Θ_0, Θ_1 are subsets of the real line. Let $X_1, \ldots, X_n$ be the sample from a population with pdf $f_\theta(x)$.

In the Bayesian hypothesis testing approach we compute the following posterior probabilities:

$$\alpha_0 = P(\theta \in \Theta_0 | x_1, \ldots, x_n) \tag{11.4}$$

and

$$\alpha_1 = P(\theta \in \Theta_1 | x_1, \ldots, x_n). \tag{11.5}$$

If $\alpha_0 > \alpha_1$, we accept the null hypothesis, and if $\alpha_0 < \alpha_1$, we reject the null hypothesis. We now outline the Bayes hypothesis testing procedure for testing hypothesis (11.3).

Let $\pi(\theta)$ be the prior. Also,

$$\pi_0 = P(\theta \in \Theta_0)$$

and

$$\pi_1 = P(\theta \in \Theta_1)$$

Definition 11.4.1 *The ratio π_0/π_1 is called the* **prior odds ratio**. *The ratio α_0/α_1 (see Equations (11.4) and (11.5)) is called the* **posterior odds ratio**.

The posterior odds ratio is the ratio of the posterior probabilities, given the data, of the null and alternate hypotheses. The posterior odds ratio will be used in decision making for testing the hypotheses. We now compute α_0 and α_1 using the Bayes theorem. That is,

$$\begin{aligned}\alpha_0 &= P(\theta \in \Theta_0 | x_1, \ldots, x_n) \\ &= \begin{cases} \int_{\Theta_0} f(\theta | x_1, \ldots, x_n)\, d\theta, & \text{if continuous} \\ \sum_{\theta \in \Theta_0} f(\theta | x_1, \ldots, x_n), & \text{if discrete.} \end{cases}\end{aligned}$$

Similarly,

$$\begin{aligned}\alpha_1 &= P(\theta \in \Theta_1 | x_1, \ldots, x_n) \\ &= \begin{cases} \int_{\Theta_1} f(\theta | x_1, \ldots, x_n)\, d\theta, & \text{if continuous} \\ \sum_{\theta \in \Theta_1} f(\theta | x_1, \ldots, x_n), & \text{if discrete.} \end{cases}\end{aligned}$$

We reject H_0 if the odds ratio $(\alpha_0/\alpha_1) < 1$ and accept H_0 if $(\alpha_0/\alpha_1) > 1$.

This method of hypothesis testing is called Jeffreys' hypothesis testing criterion. It basically says that if the posterior odds ratio is greater than 1, we accept the null hypothesis; otherwise, we reject the null in favor of the alternative hypothesis.

Because we cannot determine the probability of a single value in the continuous variable case, it should be noted that for a simple null hypothesis of the form θ equals some specified value cannot be dealt with easily in the Bayesian framework. Hence, unlike the classical framework, here we mostly deal with the composite hypotheses for both null and alternative.

Example 11.4.1

A student taking a standardized test is classified as gifted if he or she scores at least 100 out of a possible score of 150. Otherwise the student is classified as not gifted. Suppose the prior distribution of the scores of all students is a normal with mean 100 and standard deviation 15. It is believed that scores will vary each time the student takes the test and that these scores can be modeled as a normal distribution with mean μ and variance 100. Suppose the student takes the test and scores 115. Test the hypothesis that the student can be classified as a gifted student.

Solution

The hypothesis testing problem can be phrased as

$$H_0 : \theta < 100 \; vs. \; H_a : \theta \geq 100.$$

Referring to the Example 11.2.8, we know that the posterior distribution $f(\theta|x)$ is a normal with mean 110.4 and variance 69.2. Because the prior is an $N(100, 225)$, we have $\pi_0 = P(\theta < 100) = 1/2$ and $\pi_1 = P(\theta \geq 100) = 1/2$.
We can now compute

$$\begin{aligned}\alpha_0 &= P\left(\theta < 100 \,|x = 115\right) \\ &= P\left(\frac{\theta - 110.4}{\sqrt{69.2}} < \frac{100 - 110.4}{\sqrt{69.2}}\right) \\ &= P\left(z \leq -\frac{10.4}{\sqrt{69.2}}\right) = 0.106\end{aligned}$$

and

$$\begin{aligned}\alpha_1 &= P\left(\theta \geq 100 \,|x = 115\right) \\ &= 1 - P\left(\theta < 100 \,|x = 115\right) \\ &= 1 - 0.106 = 0.894.\end{aligned}$$

Thus, $\alpha_0/\alpha_1 = (0.106/0.894) = 0.119 < 1$, and we reject H_0.

BAYESIAN HYPOTHESIS TESTING PROCEDURE

To test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are given sets:

1. Consider θ as a random variable with prior distribution $\pi(\theta)$.
2. Compute the posterior distribution $f(\theta|x_1, \ldots, x_n)$ of θ given $x_1, \ldots, x_n$, using Bayes' theorem.
3. Compute α_0 and α_1 using the following formulas:

$$\alpha_0 = P(\theta \in \Theta_0 | x_1, \ldots, x_n) = \begin{cases} \int_{\Theta_0} f(\theta|x_1, \ldots, x_n)\, d\theta, & \text{if continuous} \\ \sum_{\theta \in \Theta_0} f(\theta|x_1, \ldots, x_n), & \text{if discrete} \end{cases}$$

and

$$\alpha_1 = P(\theta \in \Theta_1 | x_1, \ldots, x_n) = \begin{cases} \int_{\Theta_1} f(\theta|x_1, \ldots, x_n)\, d\theta, & \text{if continuous} \\ \sum_{\theta \in \Theta_1} f(\theta|x_1, \ldots, x_n), & \text{if discrete.} \end{cases}$$

4. Reject H_0 if the posterior odds ratio, $\dfrac{\alpha_0}{\alpha_1} < 1$. Otherwise accept.

In the foregoing procedure, we assume that $P(\theta \in \Theta_0)$ and $P(\theta \in \Theta_1)$ are both greater than zero.

EXERCISES 11.4

11.4.1. The following is random data from a normal distribution with variance 9.

0.92	1.05	5.53	3.64	−4.47	−2.60	0.71	−3.66	1.38	3.87
7.42	1.76	0.01	2.69	1.54	3.97	1.34	−1.63	−1.24	−4.78

(a) Test the hypothesis, $H_0 : \mu \leq 0$ vs. $H_a : \mu > 0$. Assume that the prior is $N(0, 4)$, so that $\mu \leq 0$ and $\mu > 0$ are equally probable.

(b) Compare your decision with classical hypothesis testing, with $\alpha = 0.05$.

11.4.2. **(a)** For the data of Exercise 11.3.2, using the Bayesian method, test the hypothesis $H_0 : \mu \leq 170$ vs. $H_a : \mu > 170$.

(b) Compare your decision with classical hypothesis testing, with $\alpha = 0.05$.

11.4.3. It is known that a certain disease affects 10% of a population. Of a random sample of 50 patients in the disease group who are exposed to a new treatment, we observe that 12 patients were hospitalized in a year. Let μ be the population rate that needs hospitalization in a year. Assume μ has a $Gamma(0.1, 2)$ prior. Let $\mu \sim Gamma(0.1, 2)$ and $f(x|\mu) \sim$

$Poi(50\mu)$. Given that $x = 0.24$ is an observation of X, test the hypothesis $H_0 : p \leq 0.10$ vs. $H_a : p > 0.10$. (If X is the number of patients admitted in a year, assume $X \sim Poi\,(50\mu)$, the Poisson approximation of the binomial.)

11.4.4. For an upcoming congressional election, suppose we want to estimate the amount of support for a particular candidate in a district. By previous experience and voter registration data, we can assume that the prior distribution, the proportion of support, p, is a beta distribution with $\alpha = 10$, and $\beta = 8$ (i.e., $\pi\,(p) \sim Beta\,(10, 8)$). We conducted a survey of 1000 randomly selected voters, of whom 600 support the candidate. Test the hypothesis $H_0 : p \geq 0.60$ vs. $H_a : p < 0.60$.

11.4.5. For the data of Exercise 11.3.5, test the hypothesis $H_0 : \mu \leq 2400$ mg vs. $H_a : \mu > 2400$ mg for this ethnic group.

11.4.6. Suppose we have a coin (not necessarily balanced) with p being the probability of heads. Assume a uniform prior for p. Suppose in 20 tosses of this coin, we obtained 12 heads. Test the hypothesis $H_0 : p \geq 0.50$ vs. $H_a : p > 0.50$.

11.5 BAYESIAN DECISION THEORY

Bayesian methods in general are more concerned with problems of decision making than with problems of inference. Decision theory, as the name implies, is concerned with the problem of making decisions. Statistical decision theory is concerned with optimal decision making under uncertainty or when statistical knowledge is available only on some of the uncertainties involved in the decision problem. Uncertainty could be about the true value related to the decision, or, uncertainty could be about the actual state of the nature. Abraham Wald (1902–1950) laid the foundation for statistical decision theory. Original works on the decision theory emerged out of game theory considerations. Many books and articles have been written on the various aspects of decision theory. The Bayesian approach to the decision theory was introduced by Leonard Jimmie Savage in 1954. In this section, we introduce the general idea of decision theory. We basically deal with analytical procedures for the decision-making process. This will involve selection of an optimum decision from a choice of courses of action among two or more alternatives. The Bayesian decision theory quantifies the trade-offs between different decisions using costs and probabilities that accompany such decisions.

Consider, as an example, a company deciding whether or not to market a new brand of toothpaste with a whitening agent. Clearly many factors will affect the decision (for example, the proportion of people who are likely to switch to the new brand, and the likelihood of other competing companies introducing similar toothpastes). These factors are generally unknown, but estimates can be obtained from statistical investigations.

The classical statistical approach relies exclusively on the data obtained from these statistical investigations, ignoring other relevant information such as the company's past experiences in marketing similar products. Statistical decision theory tries to combine other relevant information with the sample information to arrive at the optimal decision. Therefore, a Bayesian setting seems to be more appropriate for decision theory.

One piece of relevant information that decision theory considers is the possible consequences of the decisions. Often these consequences can be quantified. That is, the loss or gain of each decision can be expressed as a number (called the *loss* or *utility*). A loss or utility to a decision maker is the effect of the interaction of two factors: (1) the decision or action selected by the decision maker; and (2) the event or state of the world that actually occurs. Classical statistics does not explicitly use a loss function or a utility (payoff) function.

A second source of information that decision theory utilizes is the prior information. Prior information could be based on past experiences of similar situations or on expert opinion. We can follow the procedure explained next as a guideline for decision making.

GENERAL DECISION THEORY PROCEDURE

1. Identify the objectives of the decision-making process.
2. Identify the set of actions and set of possible events (states of nature).
3. Assign probabilities to the occurrence of each possible state of nature (prior). If more observations are available, calculate the posterior probabilities to the occurrence of each possible state of nature.
4. For each possible event, assign a numerical value to the anticipated payoff (or loss) of each course of action.
5. Compute the expected value of the payoffs (utility or loss function). This could be done by either using the prior probabilities if there are no observations, or using the posterior probabilities.
6. Select the optimum decision among the available alternative courses of action that maximizes the expected value of the payoffs.

We now consider an example to illustrate the idea of statistical decision making.

Example 11.5.1

Suppose you own a small stall at a flea market that is open only on weekends. If the weather is good, you make a profit of $200, and if it is bad, you close your stall and you make no (zero) profit. However, you have the option of buying, from an insurance company, weather insurance that costs $75. The company pays you $210 if the weather is bad. Suppose you believe that the probability of good weather on a particular weekend is p. Compute the expected gain if you insure and if you do not. What is the best course of action? Arrive at a decision.

Solution

From the information in the problem, we can obtain the utility gain or profit table shown in Table 11.4, based on our decision to insure or not insure. Suppose that we model the state of weather as good or bad by means of a random variable defined as follows.

$$\theta = \begin{cases} 1, & \text{if the weather is good} \\ 0, & \text{if the weather is bad.} \end{cases}$$

Table 11.4

	Weather	
Parameter Space → Decision Space ↓D	Good (θ_1)	Bad (θ_2)
Insurance (I)(d1)	\$125 (200–75)	\$135 (210–75)
No Insurance (NI)(d2)	\$200	\$0

Suppose for our example we believe that during a particular weekend $P(\theta = 1) = p$, *and* $P(\theta = 0) = 1 - p$. *This can be considered as prior information. The different values of* θ *are called* states of nature. *We assign (perhaps subjectively) a probability structure for the states of nature defined by a prior distribution* $\pi(\theta)$. *Now we can compute the expected gain when we insure and when we do not.*
Using the values in the table,

$$\text{Expected gain given we insure} = (125)\,p + (135)\,(1-p)$$
$$= 135 - 10p$$

$$\text{Expected gain when do not insure} = (200)\,p + (0)\,(1-p)$$
$$= 200p$$

Hence, insurance is preferable if

$$135 - 10p > 200p$$

or

$$p < \frac{135}{210} = 0.643.$$

That is, we should take the insurance if we believe the probability of good weather is less than 0.643.

In general the states of the nature are represented by $\theta_1, \ldots, \theta_n$ and the possible decisions (actions) are represented by $d_1, \ldots, d_m$. Let $U\left(d_j, \theta_i\right)$ represent the net gain when the true states of nature is θ_i and the decision d_j is made. Then we can construct the general utility table shown in Table 11.5.

In Bayesian decision theory, we assume a probability distribution on the states of nature called the prior distribution. Using this probability distribution, we can find the decision that maximizes the expected utility. That is, let the states of nature be initially modeled by a random variable θ with probability function $\pi(\theta)$ such that $P(\theta = \theta_i) = \pi(\theta_i),\ \ i = 1, \ldots, n$. Let U denote the utility. Then the expected utility for decision d_j is given by

$$E\left(U \left| d_j \right.\right) = \sum_{i=1}^{n} U\left(d_j, \theta_i\right) \pi\left(\theta_i\right).$$

Table 11.5

		States of nature					
		θ_1	θ_2	$\cdots$	θ_i	$\cdots$	θ_n
Decision States	d_1	$U(d_1, \theta_1)$	$U(d_1, \theta_2)$		$U(d_1, \theta_i)$		$U(d_1, \theta_n)$
	d_2						
	$\cdot$						
	d_j				$U(d_j, \theta_i)$		
	$\cdot$						
	d_m	$U(d_1, \theta_1)$					$U(d_m, \theta_n)$

The optimal decision, called the Bayes decision, denoted by d^*, is that which maximizes the expected utility. That is, d^* satisfies the following equation:

$$\max_{d_j} \sum_{i=1}^{n} U(d_j, \theta_i)\, \pi(\theta_i) = \sum_{i=1}^{n} U(d^*, \theta_i)\, \pi(\theta_i).$$

This procedure is called the *Bayes decision procedure* with respect to the assumed or given prior $\pi(\theta_i)$, $i = 1, 2, \ldots, n$.

PROCEDURE TO FIND OPTIMAL DECISION

1. For each decision d_i, compute $\sum_{i-1}^{n} U(d_j, \theta_i)\, \pi(\theta_i)$
2. Find a decision d^* from the decision space that maximizes the sum in step 1. This is the Bayes decision.

In determining the Bayes decision, we have assumed a prior distribution $\pi(\theta)$ for the states of nature $\{\theta_i\}$. Naturally the question arises: Can there be information or observations that will help us to determine $\pi(\theta)$?

Definition 11.5.1 *Observations that can aid us in determining the relative likelihoods of the possible states of nature are called* **observables**.

We remark that observables enable us to refine and update our initial prior $\pi(\theta)$. The updated prior is the conditional distribution $\pi(\theta|\text{observables})$, which clearly depends on the observables as well as the initial prior $\pi(\theta)$. The updated prior is also called the posterior.

For example, to determine the nature of weather we may hear the weather forecast (80% chance of rain), in which case we may assume $P(G) = 0.2$, and $P(B) = 0.8$. However, the weather forecast is not perfect. Let $\hat{G}$ and $\hat{B}$ denote the meteorologist's prediction. We may like to know $P(G|\hat{G})$ and $P(G|\hat{B})$. That is, what is the probability of the weather being good when the meteorologist predicts

the weather will be good, and what is the probability that the weather is good when the meteorologist predicts the weather will be bad?

It may be noted that there is no direct cause-effect relation in $G|\hat{G}$. That is, the prediction of the weather forecast does not influence the weather. If a probability distribution depends on a set of parameters θ, the classical approach estimates θ on the basis of an observed sample $X_1, \ldots, X_n$. The samples $X_1, \ldots, X_n$ are the observables. Thus, observables are used to estimate the parameters, that is, we want the distribution of θ given $X_1, \ldots, X_n$ or $p(\theta|X_1, \ldots, X_n)$. In our weather situation, the observable is the weather forecast, whereas the parameter is one of the weather conditions, good or bad. In $P(\hat{G}|G)$ we are asking, "Given that the weather is good, what is the probability that the weather forecast is correct?" We can imagine that meteorological conditions such as the barometric pressure determine the weather (that is, $G = f(m_1, \ldots, m_k)$, m_i = meterological factor), and in this sense we can consider that G is a parameter. We thus want $P(G|\hat{G})$.

To compute the posterior $P(G|\hat{G})$, we use the Bayes theorem (which needs a prior distribution, $P(G)$). That is,

$$P(G|\hat{G}) = \frac{P(\hat{G}|G)P(G)}{P(\hat{G}|G)P(G) + P(\hat{G}|B)P(B)}.$$

Similarly, we can compute $P(B|\hat{B})$.

Coming back to our weather situation, if $P(G)$ is known and $P(\hat{G}|G)$, $P(\hat{B}|B)$ are known, we could obtain the required posterior distributions $P(G|\hat{G})$ and $P(B|\hat{B})$. We can now use this distribution to calculate the expected utilities and choose the decision that maximizes the expected utility.

We now consider an example.

Example 11.5.2

Let us initially assume $P(\theta = 1) = P(\theta = 0) = \frac{1}{2}$. That is,

$$P\,(good\ weather) = P\,(bad\ weather) = \frac{1}{2}.$$

Suppose we have the following record on the meteorologist's predictions. The meteorologist predicts good weather $(\hat{G})$, given the weather is good, $\frac{2}{3}$ of the time, that is, $P(\hat{G}|G) = 2/3$, and predicts bad weather, given the weather is bad, $3/4$ of the time, that is, $P(\hat{B}|B) = 3/4$. Thus, given that the meteorologist predicts good weather, what is the probability that the weather will turn out to be good, and given the meteorologist predicts bad weather, what is the probability that the weather will turn out to be bad?

Solution

To compute the true probabilities, we use the Bayes theorem.
We are given $P(\hat{G}|G) = \frac{2}{3}$ *and* $P(\hat{B}|B) = \frac{3}{4}$, *which imply* $P(\hat{B}|G) = \frac{1}{3}$ *and* $P(\hat{G}|B) = \frac{1}{4}$. *Using the Bayes theorem, we obtain the likelihood of* G *as*

$$P(G|\hat{G}) = \frac{P(\hat{G}|G)P(G)}{P(\hat{G}|G)P(G) + P(\hat{G}|B)P(B)}$$
$$= \frac{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)}{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2}\right)} = \frac{8}{11}$$

and the likelihood of B *is*

$$P(B|\hat{B}) = \frac{P(\hat{B}|B)P(B)}{P(\hat{B}|B)P(B) + P(\hat{B}|G)P(G)}$$
$$= \frac{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)}{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)} = \frac{9}{13}.$$

Thus, we have the following updated prior depending upon the meteorologist's prediction. The updated prior when the meteorologist predicts good weather is

$$\pi(G) = P(G|\hat{G}) = \frac{8}{11};\ \pi(B) = 1 - \pi(G) = \frac{3}{11}.$$

Thus, the updated $\pi(G)$ *is actually* $\pi_{\hat{G}}(G)$*. Similarly, the updated prior when the meteorologist predicts bad weather (that is,* $\pi_{\hat{B}}(G)$*) is*

$$\pi(G) = P(G|\hat{B}) = \frac{4}{13};\ \pi(B) = P(B|\hat{B}) = \frac{9}{13}.$$

That is, if the meteorologist predicts good weather, he will be right about 72.7% of the time, and if he predicts bad weather, he will be right about 69.2% of the time.

■

Example 11.5.3

Consider Example 11.5.2, with the additional information that the meteorologist has predicted that the weather will be good on a given weekend. Referring to the utility table (Table 11.5) given in Example 11.5.1, we ask, what should be our decision—to insure or not to insure—in light of this prediction?

Solution

From Example 11.5.2, we know that the updated prior, given that the meteorologist predicts good weather, is

$$\pi(G) = P(G|\hat{G}) = \frac{8}{11} \text{ and } \pi(B) = P(B|\hat{G}) = \frac{3}{11}.$$

Using the foregoing prior and the utility table in Example 11.5.2, we can compute the following expected gains:

$$\text{Expected gain if we insure} = (125)\pi(G) + (135)\pi(B)$$
$$= (125)\frac{8}{11} + (135)\frac{3}{11} = 127.73.$$

and

$$\text{Expected gain if we do not insure} = (200)\frac{8}{11} = 145.45.$$

Therefore our decision, given that the meteorologist predicts good weather, is not to insure.

■

EXERCISES 11.5

11.5.1. Suppose that we will receive \$25 if we get two consecutive heads (H) on two flips of a balanced coin. If only one head appears, we will get \$10. On the other hand, if there is no heads, we will lose \$15. If monetary return is the only concern, should we play this game? Why?

11.5.2. In the previous problem, suppose we suspect the coin is not balanced. We feel that $P(H)$ is only 0.4. In our last 10 observations, we counted three heads and seven tails. Should we play the game? Defend your answer.

11.5.3. The owner of a small structural engineering firm in Tampa wants to open a new branch office in Orlando. The single most influential factor is the projected state of the economy for the next 4 years. If the economy keeps expanding or at least does not take a turn for the worse, the owner expects an annual profit of \$300,000 by opening the new office. If the economy experiences a downward trend, then the owner forecasts an annual loss of \$200,000. If he just continues to operate his business in Tampa, he expects a \$50,000 annual profit. Suppose a government forecast indicates that there is a 70% chance of economic expansion or status quo in the next 4 years and there is a 30% chance that the economy will show a decline. What is the optimal decision in this problem? Did you make any assumption in obtaining this optimal decision?

11.5.4. In Exercise 11.5.3, suppose the owner decides to look at the accuracy of past forecasts by the government. Suppose his study indicates that a forecast of economic expansion came true only 2/3 of the time, whereas an economic downturn came true 4/5 of the time. Now based on this new evidence, what is the optimal option for the owner?

11.5.5. Consider the weather Example 11.5.1, discussed earlier. The meteorologist's prediction record over the past 15 days is as follows:

Weather person's prediction	G	B	B	G	G	G	B	G	G	B	B	G	B	G	G
How the weather turned out to be	B	B	B	G	G	B	B	G	B	G	B	G	G	G	G

(a) Assuming a uniform distribution for the states of nature, obtain an updated prior (posterior) based on the meteorologist's record.
(b) Obtain the Bayes decision.

11.5.6. A coin (not necessarily fair) will be tossed once, and you have to predict the outcome. If you predict the outcome correctly you win \$1000. Otherwise, you lose \$5.
(a) What are the states of nature? What is the decision space? Write the utility table.
(b) Suppose that you believe that the probability of heads is 2/3. What is your price for the states of nature? Find the expected gains.
(c) Suppose that you are allowed to toss the coin twice and you find that the first toss results in heads and the second in tails. What are the observables?
(d) Assume the situation in (c). The coin is going to be tossed again and you have to predict the outcome. What is your updated prior?
(e) What are your expected gains, and what is your decision for the situation in (d)?

11.5.7. We are given the following utility table:

States of nature			
	θ_1	θ_2	θ_3
d_1	0	10	4
d_2	−2	5	1

Determine the Bayes decision assuming a uniform prior for the states of nature.

11.5.8. Suppose that we have an observable X that can take only two values, X_1 and X_2, for the situation in Exercise 11.5.7. The distribution of X depends on the states of nature and is as follows:

	θ_1	θ_2	θ_3
X_1	0.1	0.5	0.6
X_2	0.9	0.5	0.4

That is, $P(X = x_1|\theta_1) = 0.1$ or $P(X = x_2|\theta_3) = 0.4$, and so forth.
Suppose you observe X_1; what is the updated prior? What is the Bayes decision?

11.5.9. A large lot has p% defectives and you have to predict p. If you predict p correctly you gain \$$g$, and if the prediction is wrong, you lose \$$l$. It is known that the possible values of p are $p_1, p_2, \ldots, p_k$.
(a) Set up a utility table.
(b) Suppose you assume a uniform prior for p. That is $\pi(p_i) = \frac{1}{k}, i = 1, 2, \ldots, k$. Find an expression for the Bayes decision.
(c) Suppose you have an observable X such that $P(X = x_1|p_i) = a_i, i = 1, 2 \ldots, k$ and $P(X = x_1|p_i) = 1 - a_i, i = 1, 2, \ldots, k$. Find the updated prior for p. What is the Bayes decision in this case?

11.6 CHAPTER SUMMARY

In this chapter we introduced the basic philosophy, definitions, and methods of performing statistical analysis in a Bayesian setting. The treatment of unknown parameters as if they are random variables provides a feedback mechanism to update our original beliefs about the parameter(s). The posterior distribution of the parameter(s) represents our revised belief and is calculated by combining data and prior knowledge. We also saw a brief explanation of Bayesian decision theory. It should be noted that there are various other aspects of Bayesian analysis, such as Bayesian regression, in which priors are used about the regression coefficients as well as about the error variance. It is beyond the scope of one chapter to deal with all aspects of Bayesian analysis. There are many publications on Bayesian statistics. We have also briefly studied some elements of decision theory, which has a natural base in the Bayesian approach.

We now list some of the key definitions introduced in this chapter:

- Posterior distribution
- Quadratic loss function
- Absolute error loss function
- $100\,(1-\alpha)\,\%$ credible interval
- Prior odds ratio
- Posterior odds ratio
- Observable

In this chapter, we have also learned the following important concepts and procedures:

- Bayesian parameter estimation procedure
- Bayesian credible interval procedure
- General decision theory procedure
- Procedure to find optimal decision

11.7 COMPUTER EXAMPLES

A very popular software (and it is free) for the Bayesian computation is WinBUGS, which can be obtained from http://www.mrc-bsu.cam.ac.uk/bugs/. Computing posterior probability for proportions using the steps we learned in Section 11.2 can be performed using Minitab. Refer to the book, *Bayesian Computation Using Minitab*, by Jim Albert (Wadsworth, 1996).

PROJECTS FOR CHAPTER 11

11A. Predicting Future Observations

Suppose we want to predict the value of future observations based on the prior and observed data. In addition to the posterior distribution $f(\theta|x)$, in Bayesian statistics we are interested in the marginal density of the observations (note that because both θ and x are random, it makes sense to speak about their joint, marginal, and conditional densities). Using the Bayes theorem, we have seen that $g\,(x)$ is

the marginal density function of data at $x = (x_1, \ldots, x_n)$ (for the continuous case) to be

$$g(x) = \int f(x|\theta)\,\pi(\theta)\,d\theta$$

where $f(x|\theta)\,\pi(\theta)$ is the joint density of x and θ. This also can be written as

$$g(x) = E[f(x|\theta)],$$

the expected density of observations with respect to the prior distribution $\pi(\theta)$. With the help of $g(x)$, we can predict observations.

We are more interested in the density of future observations y, given present data x. However, because we have already updated the value of θ using the posterior density, this should be reflected in our prediction:

$$\begin{aligned} f(y|x) &= \int f(y, \theta|x)\,d\theta \\ &= \int f(y|\theta, x) \cdot \pi(\theta|x)\,d\theta \\ &= \int f(y|\theta)\,\pi(\theta|x)\,d\theta, \end{aligned}$$

if y and x are conditionally independent given θ. Conditional independence is achieved, for example, when $x = (x_1, \ldots, x_n)'$ and $y = (x_{n+1}, \ldots, x_{n+m})'$ both are samples from $f(x|\theta)$.

We see that the density of future observations is the expected density of observations with respect to posterior distribution. Consider two different priors for θ.

Uniform [0,2], (2) $N\left(1, \frac{1}{6}\right)$. Assume $f(x|\theta) \sim N(\theta, 1)$. Find the predictive distributions given the sample $X_1, X_2, \ldots, X_n$.