Chapter 10

Bayesian estimation and inference

Chapter outline

Objective

The objective of this chapter is to study the 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 (1702−61) 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.

Mathematical Statistics with Applications in R. https://doi.org/10.1016/B978-0-12-817815-7.00010-5

10.1 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 on 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 off-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 off-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 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 at an off-campus job for at least 20 hours a week is assumed to be unknown and random. A probability distribution, called the prior, represents our knowledge or belief about the location of this proportion before any collected data are 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 combines 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 a random variable, 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. Advances in Bayesian analysis combined with the growing power of computers are making Bayesian methods practical and increasingly popular. The Markov chain Monte Carlo method described in Section 13.5 is one of the computationally intensive methods that are often useful in Bayesian estimation.

10.2 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 mass function (pmf) can be written as $p(x, y) = p(x|y)p_Y(y)$, and the marginal probability mass 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 $\sum_y p(y|x) = 1$. 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 finding a point estimate of the parameter θ (possibly a vector) for the population with distribution $f(x|\theta)$, given θ. Since θ is assumed to be a random variable, we can talk of the distribution of θ. Assume that $\pi(\theta)$ is the prior distribution of θ, which reflects the experimenter's prior belief about θ. We will not distinguish between the scalars and the 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 of θ 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 10.2.1. *The distribution of* θ*, given data* $x_1, ..., x_n$*, is called the* **posterior distribution***, which is given by:*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)}, \tag{10.1}$$

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:

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

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 assign 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 10.2.1

It is believed that cross-fertilized plants produce taller offspring than self-fertilized plants. 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 under 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, the cross-fertilized plant 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 taller cross-fertilized plants in 15 pairs compared with the different prior values of π *is given using the binomial pmf* $\binom{15}{13}\theta^{13}(1-\theta)^2$. *For example, if the prior value of* θ *is 0.80, then the likelihood of* θ *given the sample is:*

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

From Table 10.1 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 10.1. 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* $13/15 = 0.867$.

TABLE 10.1 Summary of Prior and Posterior Probabilities.

Prior value of θ	Prior probability $\pi(\theta)$	Likelihood of θ given sample	Prior×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	$0.9998 \approx 1.0$

In Example 10.2.1, the priors are called *informative priors*, because they favored certain values of θ; for example, for the value $\theta = 0.86$, the prior value of π (θ) is 0.25, which is higher than all the rest of the values. If there were 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. In practice, a noninformative prior is used, when we do not have any prior information but still want to use the Bayes method. Thus, uniform prior amounts to random choice from possible values of the parameter. Based on the situation, noninformative priors may be quite dispersed, may avoid only impossible values of the parameter, and oftentimes give results similar to those obtained by classical frequentist methods.

EXAMPLE 10.2.2

Repeat Example 10.2.1 using a noninformative prior, $\pi(\theta) = 1/6$, for each given value of θ.

Solution

Here $\pi(\theta) = 1/6$ *for each value of* θ. *See Table 10.2.*

The Bayesian estimate for the noninformative prior is:

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

TABLE 10.2 Prior and Posterior Probabilities with Noninformative Prior.

Prior value of θ	Prior probability $\pi(\theta)$	Likelihood of θ given sample	Prior×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

It should be noted that because the choice of priors in Example 10.2.1 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, …, 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 10.2.3

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.

Note that in the previous example, the forms of the pmf in both $f(x|p)$ and $f(p|x)$ are the same; however, in $f(p|x)$, p is considered random and in $f(x|p)$, p is not random. This particular form of $f(p|x)$ is also called *beta-binomial distribution* for p with parameters $\alpha = x+1$ and $\beta = n-x+1$. 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 the 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 the data are an independent sample from an exponential family (such as normal, Poisson, or binomial). Bayesian priors act just like pseudo-observations added to the data.

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

EXAMPLE 10.2.4

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 which 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:

$$f(\mu|x) = \frac{f(x|\mu)\pi(\mu)}{\int f(x|\mu)\pi(\mu)\mathrm{d}\mu}$$

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}\frac{1}{\sqrt{2\pi}\sigma_p}e^{-\left(\mu-\mu_p\right)^2/2\sigma_p^2}}{\int \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}\frac{1}{\sqrt{2\pi}\sigma_p}e^{-\left(\mu-\mu_p\right)^2/2\sigma_p^2}d\mu} \tag{10.2}$$

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

Consider the exponential term in Eq. (10.2), namely, $\frac{(x-\mu)^2}{2\sigma^2}+\frac{\left(\mu-\mu_p\right)^2}{2\sigma_p^2}$.

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

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

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

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

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

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

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

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

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

where

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

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(μ|x) *is the pdf of a 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 a normal random variable with mean* $\frac{1}{\tau_p+\tau}(\tau_p\mu_p + \tau x)$ *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(μ|x) *is the pdf of a normal random variable with*

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

and

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

10.2.1 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 $a = \widehat{\theta}$ is chosen so as to minimize the expected loss $E\left[L\left(\theta, \widehat{\theta}\right)\right]$, 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,

$$E[L(\theta, a)] = \int L(\theta, a) f(\theta|x_1, \ldots, x_n) d\theta$$

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

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, $\widehat{\theta}$, to be the posterior mean. In previous examples in this section, we used this value as the estimate $\widehat{\theta}$. Note that what the quadratic loss function displays is that if the estimate $\widehat{\theta}$ 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 $\widehat{\theta}$ is going to be large.

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

$$E[L(\theta, a)] = \int L(\theta, a) f(\theta|x_1, \ldots, x_n) d\theta$$

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

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

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 $1/2$. This can be achieved by taking $\widehat{\theta}$ to be the *posterior median*.

There are other loss functions such as the all or nothing (or 0−1) loss function *given by:*

$$L(a, \theta) = 1 - \delta_{a\theta} = \begin{cases} 0, & \text{if } \theta = a \\ 1, & \text{otherwise} \end{cases}$$

where δ *is the Kronecker Delta function. This loss function is used mostly when values of* θ *are assumed to be discrete. In this case, it can be shown that expected loss is minimized when* $\widehat{\theta}$ *is the maximum of the posterior distribution, or the mode.*

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 the quadratic loss function.
5. The Bayes estimator of θ is set to be the posterior median under the 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 the steps involved in the Bayesian estimate is given in Fig. 10.1.

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:

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

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:

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

When using a beta prior, we will take the number of successes as $\alpha - 1$ and the number of failures as $\beta - 1$.

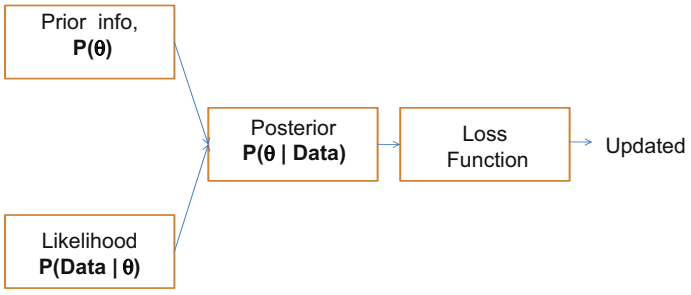


FIGURE 10.1 Bayesian estimation procedure.

EXAMPLE 10.2.5

Let $X_1, \ldots, X_n$ be a sample from a 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 the quadratic loss function.

Solution

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

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

Because the random variables X_i *have a 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:

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

Because (posterior of p*)* $\propto$ *(prior of* p*(.). (likelihood), rewriting the normalizing constant in the denominator of Eq. (10.1) 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*(α, β) *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 10.2.6

Suppose we are flipping a biased coin, for which the probability of heads p could be any value between 0 and 1. Given a sequence of toss samples, $x_1, \ldots, x_n$, 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 led 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* $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, *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 a beta distribution with mean 0.5 and variance 0.028. Fig. 10.2 gives the prior and posterior densities.

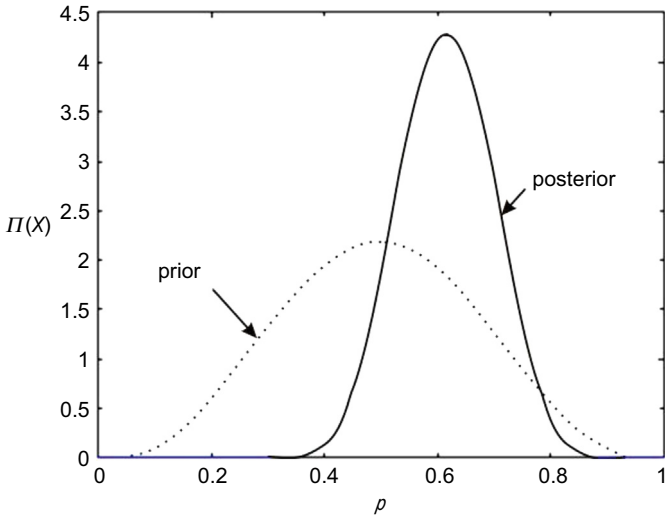


FIGURE 10.2 Prior and posterior distributions for the proportions.

Note that if we had ignored the prior and taken just the point estimation, then the MLE of p *would be* $\text{MLE(p)} = \widehat{\text{p}} = \frac{13}{20} = 0.65$. *Compare this with the Bayesian estimate of* $\text{p} = 0.607$. *Because Beta(1, 1) is the Uniform [0, 1], 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* $\text{Beta}(\text{x} + 1, \text{n} - \text{x} + 1)$. *This type of setting can be used for estimating the true proportion of infected individuals in the population.*

EXAMPLE 10.2.7

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 $(\widehat{R})$, and we were right (R) all these days. Suppose on the 10^6-th evening we predict 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	$10^6 + 1$
$\widehat{R}$	$\widehat{R}$	⋯	$\widehat{R}$	$\widehat{R}$
R	R	…	R	?

$\text{P}(\text{R}|\widehat{\text{R}}) = 1$ *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 $\pi(\text{p}|\text{x})$ *is:*

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

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

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

In our example, $\text{x} = 10^6$, $\text{n} = 10^6$, *which implies that the posterior mean is given by:*

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

EXAMPLE 10.2.8

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 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* $\pi(\mu) \sim N(\mu_0, \sigma_0^2), \pi(\mu) \propto \exp\left[(\mu - \mu_0)^2 \Big/ \sigma_0^2\right]$ *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:*

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

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 (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

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

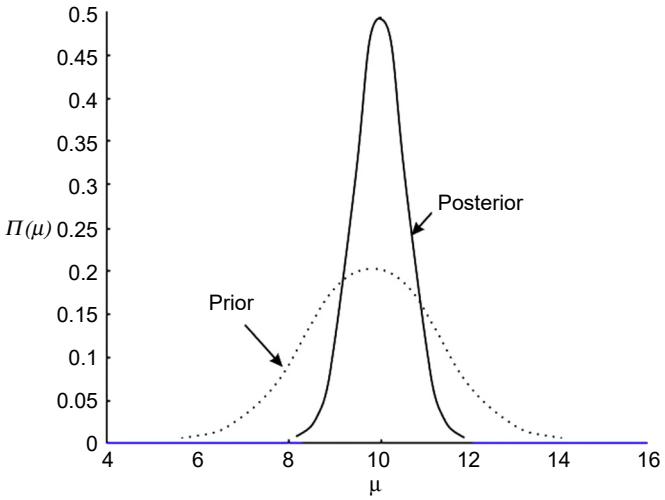


FIGURE 10.3 Prior and posterior densities of μ.

Thus, the point estimate of μ is the posterior mean, 10.167. Fig. 10.3 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 (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 $\mu_1 \to \bar{x}$. Hence, in the limit (i.e., for noninformative priors), the posterior $f(\mu|\mathrm{x})$ will have an $N(\bar{x}, \sigma^2/n)$ 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 with 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, ∞), 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. In Section 10.6, we will learn the empirical Bayes method for choosing priors based on the data itself.

Exercises 10.2

10.2.1. Suppose, in a casino, two kinds of dice are used: one kind (98%) is fair, and the other kind (2%) is loaded such that 5 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 5s. What is the probability that the die is loaded?

10.2.2. It is believed that cross-fertilized plants produce taller offspring than self-fertilized plants. 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 under 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 plant 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 (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, …, 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.

10.2.3. Consider the problem of estimating p in a binomial distribution. Let X be the 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.

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

(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.

10.2.4. 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*$(t + 1, n - t + 1)$.

10.2.5. 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 $Gamma\left(n + 1, \mu + \sum_{i=1}^{n} x_i\right)$.

(b) Obtain the Bayes estimate of λ.

10.2.6. Let $X_1, X_2, \ldots, X_n$ be Poisson random variables with parameter λ. Assume that λ has a *Gamma*(α, β) 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?

10.2.7. 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 and show that it is $Gamma\left(\left(\sum_{i=1}^{n} x_i + 1\right), (n + 1)\right)$.

(b) Find the Bayes estimate of λ.

10.2.8. 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 the population that needs 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 $E(\mu|x)$).

10.2.9. 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.

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

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

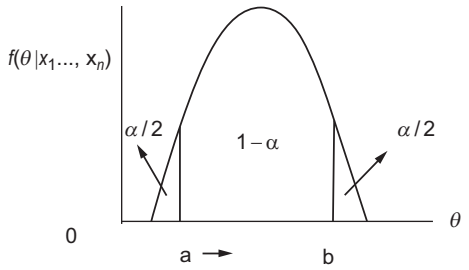


FIGURE 10.4 Credible interval for θ.

10.3 Bayesian confidence interval or credible interval

In this section, we want to study the question, "Can we construct an interval such that 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 analogue of a confidence interval is called a credible interval and is defined as follows.

Definition 10.3.1. *A* **$100(1-\alpha)\%$ *credible interval*** for θ is an interval *(a, b)* such that:

$$p(a \leq \theta \leq b | x_1, \ldots, x_n) \geq (1-\alpha).$$

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

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

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 $(1-\alpha)$ 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 posterior density of $f(\theta|x_1, \ldots, x_n)$, as shown in Fig. 10.4. This will be better especially if the density is not symmetric.

Definition 10.3.1 can be rephrased as follows using the posterior distribution of θ.

Definition 10.3.2. *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. $\sum^b f(\theta|x_1, \ldots, x_n) \geq 1-\alpha$, *if θ is discrete.*

We will now give some examples for computing credible intervals.

EXAMPLE 10.3.1

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 10.2.8 that the posterior distribution of μ given $x_1, \ldots, x_n$ is normally distributed with:

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

and

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

Fig. 10.5 *presents the posterior distribution of* μ.

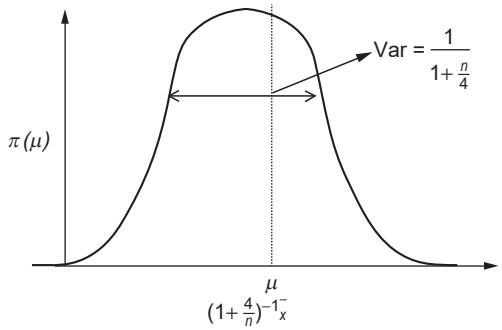


FIGURE 10.5 Posterior distribution of μ.

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 10.3.2

For the data of Example 10.2.1, find a 90% credible interval for θ.

Solution

Arranging the values of θ *from most likely to least likely, we have Table 10.3. 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 10.3 Posterior and Cumulative Probability.

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 10.3

10.3.1. **(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 (a), compute a 95% credible interval for μ, interpret its meaning, and state any assumptions you have made.

10.3.2. 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.

10.3.3. 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 the 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?

10.3.4. 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 $\mu = 10$, 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?

10.3.5. 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.

10.3.6. 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.

10.3.7. 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 5 minutes. Obtain a 95% credible interval for λ.

10.4 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 the alternative is rather straightforward. Consider the problem of hypothesis testing with:

$$H_0 : \theta \in \Theta_0 \; vs. \; H_1 : \theta \in \Theta_1 \tag{10.3}$$

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:

$$\alpha_0 = P(\theta \in \Theta_0 | x_1, \ldots, x_n) \tag{10.4}$$

and

$$\alpha_1 = P(\theta \in \Theta_1 | x_1, \ldots, x_n). \tag{10.5}$$

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 (10.3).

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

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

and

$$\pi_1 = P(\theta \in \Theta_1) = P(H_1)$$

Definition 10.4.1. *The ratio π_0/π_1 is called the* ***prior odds ratio.*** *The ratio α_0/α_1 (see Eqs. 10.4 and 10.5) is called the* ***posterior odds ratio.***

The posterior odds ratio is the ratio of the posterior probabilities, given the data, of the null and alternative 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,

$$\alpha_0 = p(\theta \in \Theta_0 | x_1, \ldots, x_n)$$

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

Similarly,

$$\alpha_1 = p(\theta \in \Theta_1 | x_1, \ldots, x_n)$$

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

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 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 10.4.1

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:

$$H_0: \theta < 100 \text{ vs. } H_a: \theta \geq 100.$$

Referring to Example 10.2.8, we know that the posterior distribution $f(\theta|x)$ *is a normal with mean 110.4 and variance 69.2. Because the prior is a* $N(100, 225)$*, we have* $\pi_0 = P(\theta < 100) = 1/2$ *and* $\pi_1 = P(\theta \geq 100) = 1/2$.

We can now compute:

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

and

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

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

EXAMPLE 10.4.1 BAYESIAN HYPOTHESIS TESTING PROCEDURE

To test H_0: $\theta \in \Theta_0$ versus 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 $f(\theta|x_1, ..., x_n)$ 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 $\frac{\alpha_0}{\alpha_1} < 1$. 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 10.4

10.4.1. The following are 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: $\mu \leq 0$ versus 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 $\alpha = 0.05$.

10.4.2. **(a)** For the data of Exercise 10.3.2, using the Bayesian method, test the hypothesis H_0: $\mu \leq 170$ versus H_a: $\mu > 170$.
(b) Compare your decision with classical hypothesis testing, with $\alpha = 0.05$.

10.4.3. 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: $p \leq 0.10$ versus 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.)

10.4.4. 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: $p \geq 0.60$ versus H_a: $p < 0.60$.

10.4.5. Using the data of Exercise 10.3.5, test the hypothesis H_0: $\mu \leq 2400$ mg versus H_a: $\mu > 2400$ mg for this ethnic group.

10.4.6. 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: $p \geq 0.50$ versus H_a: $p > 0.50$.

10.5 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 nature. Abraham Wald (1902−50) laid the foundation for statistical decision theory. Original works on 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 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 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 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 of 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.

There are many other decision criteria available in the literature. In this section, we consider only the expected utility or loss function approach. We now consider an example to illustrate the idea of statistical decision-making.

EXAMPLE 10.5.1

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 10.4, 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:

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

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,

$$\begin{aligned} \text{Expected gain given we insure} &= (125)p + (135)(1-p) \\ &= 135 - 10p \\ \text{Expected gain when do not insure} &= (200)p + (0)(1-p) \\ &= 200p \end{aligned}$$

TABLE 10.4 Weather Insurance.

	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

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 nature are represented by $\theta_1, \ldots, \theta_n$ and the possible decisions (actions) are represented by $d_1, \ldots, d_m$. Let $U(d_j, \theta_i)$ represent the net gain when the true state of nature is θ_i and the decision d_j is made. Then we can construct the general utility table shown in Table 10.5.

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:

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

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

TABLE 10.5 General Utility Table.

		States of nature					
		$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	…	$\boldsymbol{\theta_i}$	…	$\boldsymbol{\theta_n}$
	d_1	$U(d_1, \theta_1)$	$U(d_1, \theta_2)$		$U(d_1, \theta_i)$		$U(d_1, \theta_n)$
	d_2						
Decision	·						
states							
	d_j				$U(d_j, \theta_i)$		
	·						
	d_m	$U(d_1, \theta_1)$					$U(d_m, \theta_n)$

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 $\widehat{G}$ and $\widehat{B}$ denote the meteorologist's prediction. We may like to know $P\left(G\middle|\widehat{G}\right)$ and $P\left(G\middle|\widehat{B}\right)$. 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 will be good when the meteorologist predicts the weather will be bad?

It may be noted that there is no direct cause−effect relation in $G\middle|\widehat{G}$. 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 $P\left(\widehat{G}\middle|G\right)$ 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, $G = f(m_1, \ldots, m_k)$, $m_i =$ meteorological factor), and in this sense we can consider that G is a parameter. We thus want $P\left(G\middle|\widehat{G}\right)$.

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

$$P\left(G\middle|\widehat{G}\right) = \frac{P\left(\widehat{G}\middle|G\right)P(G)}{P\left(\widehat{G}\middle|G\right)P(G) + P\left(\widehat{G}\middle|B\right)P(B)}.$$

Similarly, we can compute $P\left(B\middle|\widehat{B}\right)$.

Coming back to our weather situation, if $P(G)$ is known and $P\left(\widehat{G}|G\right)$, $P\left(\widehat{B}|B\right)$ are known, we could obtain the required posterior distributions $P\left(G\middle|\widehat{G}\right)$ and $P\left(B|\widehat{B}\right)$. 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 10.5.2

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

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

Suppose we have the following record of the meteorologist's predictions. The meteorologist predicts good weather $\left(\widehat{G}\right)$, given the weather is good, 2/3 of the time, that is, $P\left(\widehat{G}\middle|G\right) = 2/3$, and predicts bad weather, given the weather is bad, 3/4 of the time, that is, $P\left(\widehat{B}|B\right) = 3/4$. 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 $P(\widehat{\mathrm{G}}|\mathrm{G}) = \frac{2}{3}$ and $\mathrm{P}(\widehat{\mathrm{B}}|\mathrm{B}) = \frac{3}{4}$, *which imply* $P\left(\widehat{\mathrm{B}}|\mathrm{G}\right) = \frac{1}{3}$ and $\mathrm{P}\left(\widehat{\mathrm{G}}\middle|\mathrm{B}\right) = \frac{1}{4}$. *Using the Bayes theorem, we obtain the likelihood of* G *as:*

$$P\left(G\middle|\widehat{G}\right) = \frac{P\left(\widehat{G}\middle|G\right)P(G)}{P\left(\widehat{G}\middle|G\right)P(G) + P\left(\widehat{G}\middle|B\right)P(B)}$$

$$= \frac{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)}{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2}\right)} = \frac{8}{11}$$

and the likelihood of B *is:*

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

$$= \frac{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)}{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)} = \frac{9}{13}.$$

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

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

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

$$\pi(G) = P\left(G\middle|\widehat{B}\right) = \frac{4}{13}; \pi(B) = P\left(B|\widehat{B}\right) = \frac{9}{13}.$$

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 10.5.3

Consider Example 10.5.2, 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 10.5) given in Example 10.5.1, we ask, what should be our decision—to insure or not to insure—in light of this prediction?

Solution

From Example 10.5.2, we know that the updated prior, given that the meteorologist predicts good weather, is:

$$\pi(G) = P\left(G\middle|\widehat{G}\right) = \frac{8}{11} \textit{ and } \pi(B) = P\left(B|\widehat{G}\right) = \frac{3}{11}.$$

Using the foregoing prior and the utility table in Example 10.5.2, 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 10.5

10.5.1. 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 are no heads, we will lose \$15. If monetary return is the only concern, should we play this game? Why?

10.5.2. 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.

10.5.3. 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?

10.5.4. In Exercise 10.5.3, 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?

10.5.5. Consider the weather problem in Example 10.5.1, 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.

10.5.6. 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)?

10.5.7. 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.

10.5.8. Suppose that we have an observable X that can take only two values, X_1 and X_2, for the situation in Exercise 10.5.7. 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?

10.5.9. 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?

10.6 Empirical Bayes estimates

Empirical Bayes methods are techniques for statistical inference in which the prior distribution is estimated from the data, instead of assuming a specific fixed prior distribution. Thus, the essential empirical Bayes task is to learn an appropriate prior distribution from ongoing statistical experience, rather than knowing it by assumption. The empirical Bayes model often provides superior estimates of parameters in comparison to the ordinary Bayes model.

The roots of empirical Bayes can be traced back to a work by von Mises in the 1940s; however, the first major work was developed by Herbert Robbins, who introduced the concept of empirical Bayes to estimate the parameter that behaves as a random variable in the pdf of a given set of data, $f(x|\theta)$. Robbins's framework is considered nonparametric empirical Bayes in which the prior distribution is completely unspecified. In this section, we will mostly study parametric empirical Bayes. In this approach, we specify a parametric family of distributions. Major works on (parametric) empirical Bayes were done by Efron and Morris in the 1970s.

What we have learned in the previous sections of this chapter we refer to as ordinary or standard Bayesian analysis of a given set of data, $X_1, \ldots, X_n$, that has been drawn from a population or follows the pdf, $f(x|\theta)$. Usually, once we are given the random sample of size n, we perform a goodness-of-fit test to identify the pdf that probabilistically characterizes the behavior of the given data. Sometimes, it is not possible to define the pdf of difficult data; then we proceed to analyze the subject data using nonparametric methods that we present in Chapter 12.

In an ordinary Bayes estimate, we assume that we have identified the pdf, $f(x|\theta)$, that characterizes the behavior of the given data. In such a situation Bayes theory assumes that the parameter θ in $f(x|\theta)$ behaves as a random variable rather than a fixed point estimate. Thus, we need the pdf of the parameter θ since we assumed its behavior as a random variable. We refer to the pdf of θ, say, $\pi(\theta)$, as the *prior* pdf of θ. In summary, to perform a standard Bayesian estimation, we need the following:

1. Identify through the goodness-of-fit methods the pdf of the given data, $f(x|\theta)$.
2. Assume a prior pdf that defines the random parameter θ, $\pi(\theta)$.
3. Assume a loss function, $L(\theta, \widehat{\theta})$, to be used in the analysis.

Thus, using the above information, we develop the posterior pdf and its expected value in the ordinary (standard) Bayesian estimate of the parameter θ, which was assumed to behave as a random variable. The major problem that we have in performing ordinary Bayesian analysis is that we must assume or guess the *prior* pdf. In analyzing real-world data from health sciences, business, and engineering, among others, we cannot assume a *prior* pdf. That is, if we obtain ordinary Bayesian results from an assumed *prior* pdf, and we check our result using a different *prior* pdf, the results will be different. Thus, an ordinary Bayesian estimate is very sensitive to the choice of the assumed prior pdf. To address the issue of not assuming the prior pdf, we will study some basic aspect of ***empirical base estimates***. There are several methods that have been introduced to empirically estimate the prior pdf of θ, $\pi(\theta)$, so that we do not have to assume or guess it as in the standard Bayes method. Recall that the Bayes theorem can be stated as follows:

$$p(\theta|y) = \frac{\pi(\theta)p(y|\theta)}{p(y)},$$

where $p(\theta|y)$ is the posterior pdf, and $\pi(\theta)$ and $p(y|\theta)$ are the prior and sampling pdf, respectively. If θ has a discrete distribution, then the marginal pdf of Y is given by:

$$p(y) = \sum\nolimits_{\theta} \pi(\theta)p(y|\theta)$$

where the sum is over all possible values of θ. In the continuous case of θ we have:

$$p(y) = \int \pi(\theta)p(y|\theta)d\theta.$$

In standard Bayes, in general we assume a prior pdf $\pi(\theta)$ depends on another parameter, η, that is, $\pi(\theta|\eta)$, where η is called a hyperparameter (η could be a vector). Since the prior of θ depends on another parameter η, the posterior density is:

$$p(\theta|y) \propto p(\eta)\pi(\theta|\eta)p(\mathrm{y}|\theta).$$

In general, it is difficult if not impossible to directly calculate $\pi(\theta|\eta)$. In the empirical Bayes, we consider:

$$p(\theta|y) \approx p(\theta|y, \widehat{\eta}(y)), \quad with \quad \widehat{\eta}(y) = \arg\max_{\eta} \pi(\theta|\eta).$$

Note that we use *argmax*; since $\pi(\theta|\eta)$ is a function of both θ and η, we are maximizing only with respect to η. This way, we could reduce the complexity of choosing the hyperparameter (prior) by replacing, in most of the cases, with the MLE of $\widehat{\eta}$ of η based on observed data y. This is why, in empirical Bayes, the choice of the prior itself is based on the observed data. The main difference between ordinary Bayes and empirical Bayes methods is that, in standard Bayes, the hyperparameter η is assumed to be known, that is, a hyperprior pdf has been placed as η. In contrast, in the empirical Bayes approach, the hyperparameter remains unknown. Thus, η needs to be estimated from the given data. This approach is not really Bayesian because we use the same data to identify the prior pdf. Empirical Bayes methods are often considered as a bridge between classical and Bayesian inference.

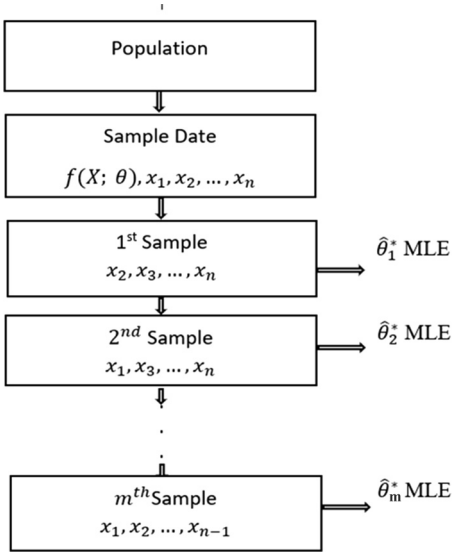
It is more common to use bootstrap methodology to estimate the prior in the empirical Bayes approach; in the present approach, we shall introduce two methods of resampling the given data to estimate (identify) the prior pdf. The resampling methods that we shall use are the jackknife and the bootstrap. These resampling methods are discussed in Chapter 13; however, we will give here a brief discussion.

10.6.1 Jackknife resampling

M.H. Quenouille, in 1949, introduced this resampling method, and in 1956, John Tukey refined the method and named it jackknife, after the Swiss jackknife, which has multiple useful tools. Given a random sample of size n, $X = (X_1, \ldots, X_n)$, the *jackknife* samples are computed by omitting one observation x_i at a time, that is,

$$x_i = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n).$$

The dimension of the jackknife sample x_i is $m = n - 1$, that is, n different jackknife samples, $\{x_{(i)}\}_{i=1, \ldots, n}$. The following diagram illustrates the process of jackknife resampling, and for each new sample we obtain the MLE of θ, that is, $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_m^*, m = n - 1$.

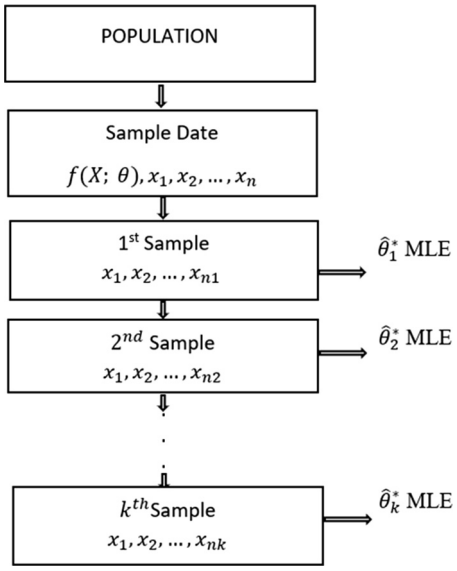


Now, our objective is to use this sequence of jackknife resampling of the MLE of θ, $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_m^*, m = n - 1$, to obtain if possible the pdf of these estimates and use it as our prior pdf, $\pi(\theta)$, and proceed to obtain the Bayesian estimate of θ, without guessing it.

10.6.2 Bootstrap resampling

Bradley Efron in 1979 introduced the bootstrap resampling method for estimating the sampling distribution of an estimator. Given a set of data n, using the subject method, we generate k samples with replacement from the given data with $k < n$. The pdf of the k samples will follow the original pdf of the n independent and identically distributed observations. Consider the observation $x_1, x_2, \ldots, x_n$; by bootstrapping we obtain different subsets of our original sample, that is, a subsample of size k. There are several uses of this method, but in our present study of empirical Bayes, we shall use bootstrapping

resampling to obtain an estimate of the prior pdf. For a given set of data $x_1,\ x_2,\ \ldots,\ x_n$, we will proceed if possible to identify the pdf, $f(x|\theta)$, that follows the observations or the population that it is drawn from. Through bootstrap resampling we will obtain a sequence of estimates, $\widehat{\theta}_1^*,\ \widehat{\theta}_2^*,\ \ldots,\ \widehat{\theta}_k^*$, and through goodness-of-fit methods, we proceed to obtain an estimate of the prior pdf, $\pi(\theta)$, if possible. The following diagram illustrates the process we follow to resample using the bootstrap method:



Thus, our objective is to use this sequence of estimates to obtain, through the goodness-of-fit method, if possible, the pdf that drives these estimated $\widehat{\theta}_1^*,\ \widehat{\theta}_2^*,\ \ldots,\ \widehat{\theta}_k^*$ and consider it as the prior pdf, $\pi(\theta)$, of the parameter θ, that is, $\pi\left(\widehat{\theta}^*\right)$. We then proceed to obtain the Bayesian estimate of θ without having to guess it.

10.6.3 Parametric, standard Bayes, empirical Bayes: Bootstrapping and jackknife

In statistics, when we are given a set of data, initially we characterize if the data were randomly collected and test and, if necessary, remove any outliers. Our next step is to perform parametric analysis, that is, through a goodness-of-fit test we try to identify, if possible, the pdf that probabilistically characterizes the behavior of the given data. If we cannot identify a well-defined pdf we must rely on nonparametric methods. The parametric analysis is the underlying pdf, say, $f(x|\theta)$. A better estimate than the parametric estimate, usually the MLE, is the Bayesian estimate. In the Bayesian estimate we ask for more information, such as the prior pdf; therefore we expect to get more about the estimate of the true parameter θ. In Bayesian analysis we proceed with standard and empirical Bayes methods to study θ. In the following Example 10.6.1 we shall use the data given in Table 10.6 that represent a certain phenomenon of interest to illustrate the parametric, standard, and empirical Bayesian estimate of the true parameter θ.

TABLE 10.6 The Data.

0.46	0.36	0.05	1.55	0.31	0.59	0.05	0.87	0.12	0.10	0.26	0.17	1.01	0.56	0.57
0.19	0.04	0.21	0.04	0.25	0.69	0.88	0.27	0.10	0.47	0.20	0.06	0.05	0.28	0.33
0.10	0.42	0.46	0.51	0.99	0.79	0.35	1.11	0.57	0.18	0.47	0.43	0.67	0.50	0.07
0.22	0.27	0.33	1.27	0.55	0.01	0.77	0.56	0.48	0.02	0.69	1.85	0.63	1.54	0.57
0.07	0.18	0	0.34	0.31	1.19	0.71	0.07	0.34	0.64	0.63	0.47	2.06	0.05	1.36

EXAMPLE 10.6.1

(a) Parametric analysis

We would like to find, if possible, the pdf, say, $f(x|\theta)$, that follows the data given in Table 10.6. We shall use the three commonly used goodness-of-fit tests in search of the pdf of the given data. These tests are:

1. Kolmogorov–Smirnov
2. Anderson–Darling
3. Cramer–von Mises criterion

More information about the definition and structure of these goodness-of-fit tests will be found in Chapter 11.

To obtain a visual idea of what type of pdf we are looking for, we structure the histogram of the given data as shown in Fig. 10.6 to obtain some idea of what type of pdf we are looking for.

The histogram suggests some sort of exponential decay of a pdf. Thus, we believe that exponential pdf is a good candidate and we begin to perform goodness-of-fit, using the three tests we mentioned. Table 10.7 shows the goodness-of-fit result for the exponential pdf:

Below is the SAS code for producing the goodness-of-fit results that are given in Table 10.7.

```
data random;
input var @@;
cards;

0.46  0.36  0.05  1.55  ·····

;

run;
proc univariate data = random;
var var ; histogram/exp; run;
```

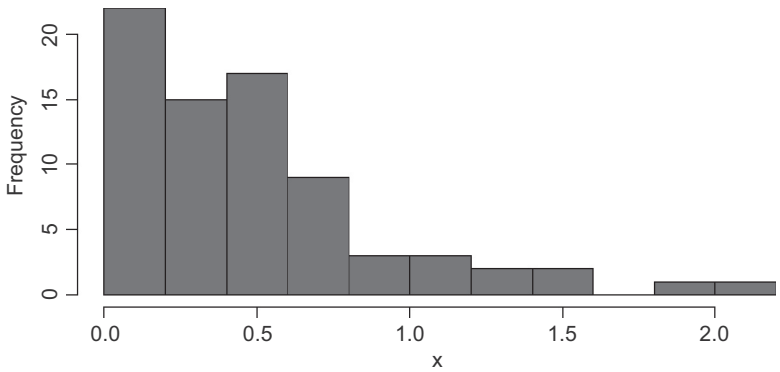


FIGURE 10.6 Histogram of the data.

TABLE 10.7 Goodness-of-Fit Result.

Goodness-of-fit tests for exponential distribution				
Test	**Statistic**		***p* value**	
Kolmogorov–Smirnov	*D*	0.09398415	Pr > *D*	>0.250
Cramer–von Mises	W-Sq	0.14110857	Pr > W-Sq	0.168
Anderson–Darling	A-Sq	0.71224681	Pr > A-Sq	>0.250

Thus, all three tests identify the one-parameter exponential pdf that fits the given data, that is,

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), & x \geq 0,\ \theta > 0 \\ 0, & \textit{elsewhere}. \end{cases}$$

We used the MLE of θ in performing the aforementioned tests, which is given by:

$$\widehat{\theta}_{MLE} = \overline{X} = \frac{1}{75}\sum_{i=1}^{75} x_i = 0.492.$$

The graph of the exponential pdf, $f(x|\theta = 0.492)$, is given in Fig. 10.7.

In addition to the MLE of the true parameter, we can obtain $100(1-\alpha)\%$ confidence limits for the parameter θ, which will be used to compare with standard and empirical Bayes estimates. The confidence limit is based on the χ^2 distribution. That is,

$$P\left[\frac{2n}{\overline{X}\left(\chi^2_{\frac{\alpha}{2},\ 2n}\right)} \leq \theta \leq \frac{2n}{\overline{X}\left(\chi^2_{1-\frac{\alpha}{2},\ 2n}\right)}\right] \geq 100(1-\alpha)\%.$$

For 90% and 95% confidence limits, we have:

$$P[\,0.41 \leq \theta \leq 0.58] \geq 0.90$$

and

$$P[\,0.40 \leq \theta \leq 0.60] \geq 0.95.$$

Thus, the confidence range of θ for the 90% and 95% confidence limits is 0.17 and 0.2, respectively.

(b) Standard Bayes estimate

Now that we have identified the pdf of the given data to be an exponential distribution, we shall assume that the parameter θ behaves as a random variable and we denote the pdf as $f(x|\theta)$. Here we will assume or guess that the pdf of θ, that is, the prior pdf $\pi(\theta)$, is given by the inverted gamma, that is,

$$\pi(\theta|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp\left(-\frac{\beta}{\theta}\right), & \theta > 0,\ \alpha,\beta > 0 \\ 0, & \textit{elsewhere}. \end{cases}$$

where α and β are hyperparameters.

Now, we have identified the pdf of the data, we have assumed the prior pdf, and, assuming a mean squared error loss function, we can obtain a Bayesian estimate of the true parameter θ.

The square error loss function is given by:

$$L\left(\theta,\ \widehat{\theta}\right) = c(\theta)\left(\theta - \widehat{\theta}\right)^2,$$

and assume that $c(\theta) = 1$. We choose the estimate of θ, $\widehat{\theta}$, so as to minimize the expected loss, $E\left[L\left(\theta,\ \widehat{\theta}\right)\right]$.

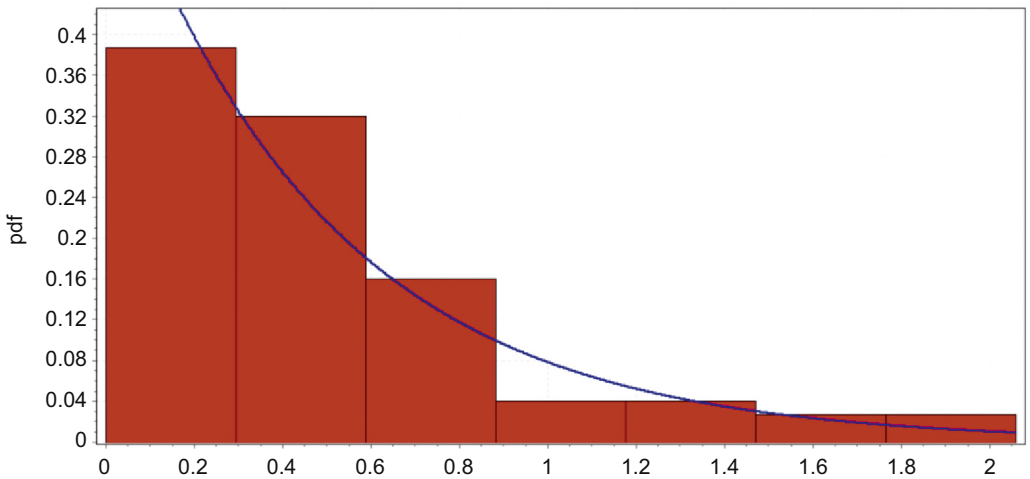


FIGURE 10.7 Probability density function (*pdf*) of $f(x;\theta = 0.492)$.

The posterior pdf is given by:

$$P(\theta|X) = \frac{\pi(\theta)\ L(\theta|X)}{m(x)}. \tag{10.1}$$

The likelihood function $L(\theta|X)$ and prior pdf, $\pi(\theta)$, for a single observation are given by:

$$\begin{aligned} L(\theta|X) &= \frac{1}{\theta}\, e^{\frac{-x}{\theta}}\, \frac{\beta^\alpha}{\Gamma\alpha}\, \theta^{-(\alpha+1)}\ e^{\frac{-\beta}{\theta}} \\ &= \frac{\beta^\alpha}{\Gamma\alpha}\, \frac{1}{\theta^{2+\alpha}}\, e^{\frac{-(\beta+x)}{\theta}}. \end{aligned}$$

The marginal pdf, $m(x)$, is given by:

$$m(x) = \int_0^\infty \frac{\beta^\alpha}{\Gamma\alpha}\, \frac{1}{\theta^{2+\alpha}}\, e^{\frac{-(\beta+x)}{\theta}}\, d\theta,$$

To compute the above integral, we make a transformation from θ to Y and assume $\theta = \frac{1}{y}$, then:

$$m(x) = \frac{\beta^\alpha}{\Gamma\alpha} \int_0^\infty y^{\alpha+2}\, e^{-y(\beta+x)}\, |J|\, dy,$$

where $|J| = \left|\frac{d\theta}{dy}\right|$ is the absolute value of the Jacobian of transformation. Simplifying $m(x)$, we have:

$$\begin{aligned} m(x) &= \frac{\beta^\alpha}{\Gamma\alpha} \int_0^\infty y^{\alpha+2}\, e^{-y(\beta+x)}\, y^{-2}\, dy \\ &\frac{\beta^\alpha}{\Gamma\alpha} \int_0^\infty y^{(\alpha+1)-1}\, e^{-y(\beta+x)}\ dy \\ &= \frac{\beta^\alpha}{\Gamma\alpha}\, \frac{\Gamma(\alpha+1)}{\Gamma(\beta+x)^{\alpha+1}}\ . \end{aligned}$$

Thus, the posterior pdf, Eq. (10.6.1), $P(\theta|X)$, is given by:

$$P(\theta|X) = \frac{(\beta+x)^{\alpha+1}}{\Gamma(\alpha+1)}\, \theta^{-(\alpha+2)}\, e^{\frac{-(\beta+x)}{\theta}}.$$

Note that for a single observation x, $P(\theta|X)$ is the same as an inverted gamma pdf with hyperparameters $\alpha+1$ and $\beta+x$. Thus, in general, $\underline{x} = (x_1, x_2, \ldots, x_n)$, the posterior pdf is given by:

$$P(\theta|X = \underline{x}) = \frac{\left(\beta + \sum_{i=1}^n x_i\right)^{\alpha+n}}{\Gamma(\alpha+n)}\, \theta^{-(\alpha+n)}\, e^{\frac{-\beta+\sum_{i=1}^n x_i}{\theta}}.$$

Using the R-output given below we have estimated α and β for the given data: $\widehat{\alpha} = 0.57$ and $\widehat{\beta} = 0.06$.

```
> library(fitdistrplus)
> ob = fitdist(1/ro, "gamma")
> ob
```

Fitting of the distribution "gamma" by maximum likelihood parameters:

	Estimate	Standard Error
Shape	0.57263897	0.07816841
Rate	0.05705247	0.01167617

Thus, for $n = 75$, $\widehat{\alpha} = 0.57$, and $\widehat{\beta} = 0.06$, the posterior pdf, $P(\theta|X)$ is given by:

$$P(\theta|X) = \frac{(36.9)^{75.57}}{\Gamma(75.57)}\, \theta^{-(75.57)}\, e^{\frac{-(36.9)}{\theta}},\ 0 < \theta.$$

Recall that the inverted gamma pdf is of the form:

$$\pi(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, \theta^{\alpha-1}\, exp\left(\frac{-\beta}{\theta}\right),\ \theta > 0,\ \alpha, \beta > 0\,,$$

which has mean $E[\theta] = \frac{\beta}{\alpha-1}$ and $var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 0$.

Thus, for a squared loss function, the expected value of the posterior pdf is the standard ordinary Bayesian estimate of θ. That is,

$$E(P(\theta|x)) = \frac{\widehat{\beta}}{\widehat{\alpha}-1} = \frac{36.9}{75.57-1} \approx 0.495.$$

Thus,

$$\widehat{\theta}_{OB} \approx 0.495.$$

Now, to calculate the Bayesian credible interval (upper and lower confidence limits), say, and a and b, we need to find:

$$P[a \leq \theta \leq b|\ x_1,\ x_2,\ \ldots,\ x_n] \geq (1-\alpha)100\%.$$

That is, we need to integrate:

$$\int_0^a \frac{(36.9)^{75.57}}{\Gamma(75.57)}\ \theta^{-(75.57)}\ e^{\frac{-(36.9)}{\theta}}\ d\theta = \frac{\alpha}{2},$$

and

$$\int_0^b \frac{(36.9)^{75.57}}{\Gamma(75.57)}\ \theta^{-(75.57)}\ e^{\frac{-(36.9)}{\theta}}\ d\theta = \frac{\alpha}{2}.$$

The following R-code gives us 95% and 90% confidence limits on the true θ.

```
> library(pscl)
> c (qigamma (0.025, alpha=75.57, beta=36.9), qigamma(0.975, alpha=75.57, beta=36.9))
```

[1] 0.3945061 0.6201909

> c (qigamma (0.05, alpha=75.57, beta=36.9), qigamma(0.95, alpha=75.57, beta=36.9))

[1] 0.4081207 0.5964911

That is, the Bayesian 95% and 90% confidence limits (credible interval) are:

$$P[0.4 \leq \theta \leq 0.62] \geq .95$$

and

$$P[0.41 \leq \theta \leq 0.6] \geq .90.$$

Thus, we have 95% and 90% confidence ranges of 0.22 and 0.19, respectively. While these ranges are slightly wider than the non-Bayesian range, we need not assume the sampling distribution.

(c) Empirical Bayes: Bootstrap

Here we will use bootstrap resampling to obtain the MLE of each of the samples of the true parameter θ that behave as random variables. We follow the resampling procedure of bootstrap that we have discussed and obtain 50 samples from the original data $n = 75$ that was given. Through goodness-of-fit we found the exponential pdf, $P(x|\theta)$. That is, for each of the 50 samples of size 75 we obtained 50 estimates of θ, $\widehat{\theta}^*_1,\ \widehat{\theta}^*_2,\ \ldots,\ \widehat{\theta}^*_{50}$, as shown in Table 10.8.

The R-code for obtaining the bootstrap of θ is:

```
> set.seed(100)
> N ← length (ro)
> nboots ← 50
> boot.result ← numeric (nboots)
> for (i in 1:nboots)
+{
+ boot.samp ← sample(ro, N, replace=TRUE)
+ boot.result [i] ← mean (boot.samp)
+}
```

To obtain, if possible, the pdf of θ, we started with a histogram to obtain a visual indication of a possible pdf. Given in Fig. 10.8 is the histogram of the 50 MLE of θ.

TABLE 10.8 Bootstrap Estimate of θ.

0.52	0.46	0.48	0.44	0.53	0.52	0.53	0.54	0.47	0.49
0.45	0.52	0.63	0.49	0.54	0.44	0.46	0.49	0.49	0.51
0.45	0.52	0.53	0.52	0.47	0.51	0.50	0.45	0.47	0.54
0.42	0.53	0.43	0.54	0.51	0.48	0.57	0.49	0.57	0.47
0.55	0.42	0.44	0.44	0.50	0.52	0.52	0.44	0.42	0.48

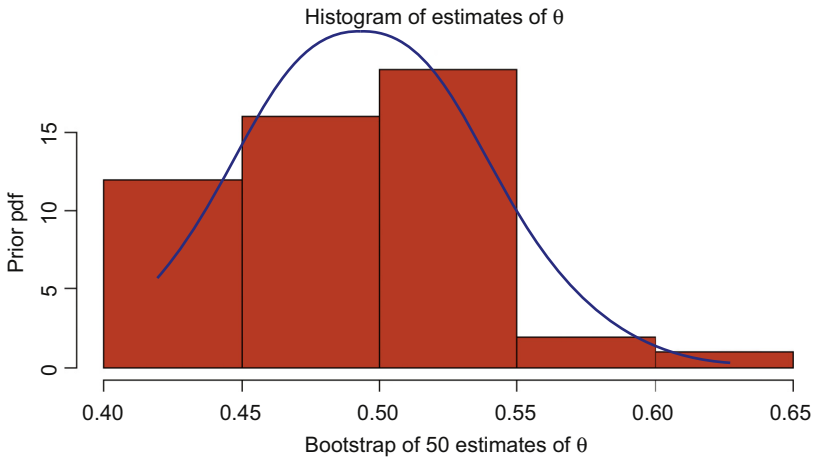


FIGURE 10.8 Histogram of the estimate of θ, $\widehat{\theta}^*_{50}$. *pdf*, probability density function.

The histogram indicates a gamma pdf for $\widehat{\theta}^*_{50}$. We performed a goodness-of-fit test to confirm that indeed the 50 bootstrap estimates of θ, $\widehat{\theta}^*_1, \widehat{\theta}^*_2, \ldots, \widehat{\theta}^*_{50}$, follow the gamma pdf. That is,

$$\pi(\theta;\alpha,\beta) = \pi(\theta^*|\alpha,\beta) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}\exp\left(-\dfrac{\theta}{\beta}\right), & \theta>0,\ \ \alpha,\beta>0 \\ 0, & \textit{elsewhere}. \end{cases}$$

The goodness-of-fit results using the three tests are given in Table 10.9.

All three tests confirm the gamma prior pdf for $\widehat{\theta}^*$, the estimate of θ, using bootstrap resampling. Through the goodness-of-fit testing we obtained the MLE of the hyperparameters α and β of the identified prior pdf, that is, $\widehat{\alpha} = 125.23$ and $\widehat{\beta} = \frac{1}{253.5} = 0.0039$. Thus,

$$\pi\left(\widehat{\theta}^*;\widehat{\alpha},\widehat{\beta}\right) = \frac{1}{\Gamma(125.23)\,(0.0039)^{125.23}}\,\widehat{\theta}^{*124.23}\,exp\left(\frac{-\widehat{\theta}^*}{0.0039}\right),\ \widehat{\theta}^* > 0.$$

TABLE 10.9 Goodness-of-Fit Results.

Goodness-of-fit tests for exponential distribution				
Test	**Statistic**		***p* value**	
Kolmogorov–Smirnov	*D*	0.10861215	Pr > *D*	0.146
Cramer–von Mises	W-Sq	0.07346459	Pr > W-Sq	>0.250
Anderson–Darling	A-Sq	0.49233552	Pr > A-Sq	0.220

The posterior pdf, $\pi\left(\widehat{\theta}^{*};\widehat{\alpha},\widehat{\beta}|\underline{X}\right)$, is given by:

$$\pi\left(\widehat{\theta}^{*};\widehat{\alpha},\widehat{\beta}|\underline{X}\right) = \frac{\left(\widehat{\beta}^{-1}+n\overline{X}\right)^{\widehat{\alpha}+n}}{\Gamma(\widehat{\alpha}+n)}\,\widehat{\theta}^{*\widehat{\alpha}+n-1}\,exp^{-}\widehat{\theta}^{*}\left(\widehat{\beta}^{-1}+n\overline{X}\right),$$

$$= \frac{(355)^{175.23}}{\Gamma(175.23)}\,\widehat{\theta}^{*174.23}\,exp^{-\widehat{\theta}^{*}(355)},\ \ \widehat{\theta}^{*}>0\,.$$

We know, under squared error loss function, the Bayes estimate of θ is the posterior mean. That is, for $\alpha = n+\widehat{\alpha}$ and $\beta = \widehat{\beta}^{-1}+n\overline{X} = 355$, we have:

$$E\left[\pi\left(\widehat{\theta}^{*};\,\widehat{\alpha},\widehat{\beta}|\underline{X}\right)\right] = \frac{n+\widehat{\alpha}}{\widehat{\beta}^{-1}+n\overline{X}} = \frac{50+125.23}{254.5+101.5} = 0.494.$$

We used the SAS code given below to obtain the necessary calculations.

```
data
input bootvar @@;
cards;

0.52  0.46  0.48  0.44  0.53  0.52  .....

;
run;
proc univariate data= boot;
var bootvar;
histogram/ gamma odstitle = " fitting gamma distribution on 50 bootstrap estimates"
VAXISLABEL= "prior"; inset n mean (5.3) std='Std Dev' (5.3)
skewness (5.3) kurtosis (5.3)
/ pos = ne header = 'Summary Statistics' ; run;
```

Thus, the Bayesian estimate of θ under bootstrap resampling to determine the prior pdf of θ is:

$$\widehat{\theta}^{*}_{empboot} = 0.494.$$

The analytical form of a $100(1-\alpha)\%$ credible interval for the true parameter θ, under the bootstrapping Bayesian estimate for the true θ, is given by:

$$\int_{0}^{a}\frac{(355)^{175.23}}{\Gamma(175.23)}\ \theta^{(174.23)}\,e^{-\theta(355)}d\theta = \frac{\alpha}{2},$$

and

$$\int_{b}^{\infty}\frac{(355)^{175.23}}{\Gamma(175.23)}\,\theta^{(174.23)}\,e^{-\theta(355)}d\theta = \frac{\alpha}{2}.$$

Using the following R-code we obtain (a, b) for 90% and 95% credible intervals using the bootstrap Bayes estimate of $\theta, \widehat{\theta}^{*}$:

```
> bootconf← rgamma(50, shape = 175.23, scale = .0028)
> library(EnvStats)
> eqgamma ( bootconf, p = 0.5, method = "mle", ci = TRUE, ci.type = "two-sided",
+ conf.level = 0.95, normal.approx.transform = "Kulkarni.powar", digits = 0)
```

Thus, the 90% credible interval for the true θ is [0.476, 0.495] and

$$P[0.476 \le \theta \le 0.495] \ge 90\%\,,$$

with a confidence range of 0.019. Similarly, the 95% credible interval for the true θ is [0.481, 0.502]; thus,

$$P[0.481 \le \theta \le 0.502] \ge 95\%$$

with a confidence range of 0.021. These are much smaller confidence ranges compared with parts (a) and (b).

(d) Empirical Bayes estimate: Jackknife

Here we shall use the jackknife resampling method on the data given in Table 10.6. Recall that we have identified the one-parameter exponential pdf with the MLE of $\widehat{\theta} = 0.492$, that is, $f(x; 0.492)$, the underlying pdf of the data.

Now, we will follow the jackknife resampling procedure and obtain 50 samples, and for each sample we will calculate the MLE of the true θ, that is, $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_{50}^*$. These jackknife MLE estimates of θ are given in Table 10.10.

The R-code for obtaining the jackknife estimates of θ is:

```
> set.seed(10)
> jack ← numeric(length (ro)-1)
> pseudo ← numeric(length (ro))
> for (i in 1:length(ro))
+{ for (j in 1:length(ro))
+ { if(j<i) jack[j] ← ro[j] else if(j>i) jack[j-1] ← ro[j]}
+ pseudo[i] ← length(ro)*mean(ro) − (length(ro)-1)*mean(jack)}
> samj ← sample(pseudo, 50, replace = T)
```

Now, we are interested in finding, if possible, the pdf of the 50 jackknife estimates through goodness-of-fit testing. Using the three commonly used goodness-of-fit tests, the results are given in Table 10.11.

All three tests at the level of significance $\alpha = 0.05$ identified the one-parameter exponential pdf fit, the jackknife estimates with the MLE of the hyperparameter being 0.46. That is, the estimated prior pdf of θ, $\pi(\theta; 0.46)$, is given by:

$$\pi\left(\widehat{\theta}^*; 0.46\right) = \frac{1}{0.46}\exp\left(-\frac{\widehat{\theta}^*}{0.46}\right), \widehat{\theta}^* \geq 0.$$

The histogram along with the graph of the estimated prior, $\pi(\theta; 0.46)$, is given in Fig. 10.9.

Fig. 10.9 shows the pdf of 50 jackknife estimates, which has been identified as the exponential pdf. Thus, we will use the exponential pdf as our prior to obtain the empirical Bayes estimate of the true θ.

The SAS codes for producing the above results are given below:

```
data
input jackvar @@;
cards;

0.57  0.10  0.46  0.77  0.05  0.04.....

;
```

TABLE 10.10 Jackknife Estimates of θ.

0.57	0.10	0.46	0.77	0.05	0.04	0.69	0.69	0.27	0.46
1.27	0.67	0.12	0.07	0.06	0.46	1.55	0.25	0.33	0.00
0.31	0.27	1.54	0.06	0.10	0.56	0.00	0.21	0.63	0.06
0.47	0.05	1.01	0.07	0.42	1.85	0.18	0.47	0.77	1.11
0.69	0.21	0.36	0.02	0.04	1.01	0.36	0.35	0.87	0.07

TABLE 10.11 Goodness-of-Fit Result.

Goodness-of-fit tests for exponential distribution				
Test	**Statistic**		***p* value**	
Kolmogorov–Smirnov	D	0.09045389	Pr > D	>0.500
Cramer–von Mises	W-Sq	0.08260139	Pr > W-Sq	>0.250
Anderson–Darling	A-Sq	0.51924948	Pr > A-Sq	>0.250

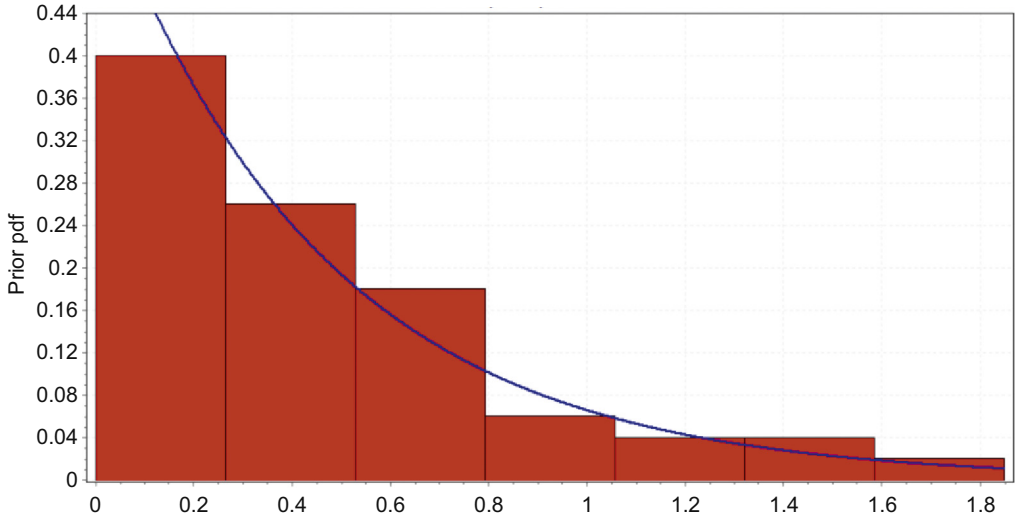


FIGURE 10.9 Histogram and prior probability density function (*pdf*) of $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_{50}^*$.

```
run;
proc univariate data= jack; var jackvar;
histogram/ exp odstitle = " fitting an Exponential distribution on 50 Jackknife estimates"
VAXISLABEL= "Prior PDF"; inset n mean (5.3) std='Std Dev' (5.3)
skewness (5.3) kurtosis (5.3)
/ pos = ne header = 'Summary Statistics' ; run;
```

The posterior pdf $\widehat{\theta}^*$ with hyperparameter β, and its MLE $\widehat{\beta}$, $\pi\left(\widehat{\theta}^*; \widehat{\beta} \,|x_1, x_2, \ldots, x_n\right)$, is given by:

$$\pi\left(\widehat{\theta}^*; \widehat{\beta}|\underline{X}\right) = \frac{\left(\widehat{\beta}^{-1} + n\overline{X}\right)^{n+1}}{\Gamma(n+1)} \widehat{\theta}^{*(n+1)-1} \, exp^{-}\widehat{\theta}^* \left(\widehat{\beta}^{-1} + n\overline{X}\right), \; 0 \leq \theta.$$

For $n = 50$ and $\widehat{\beta} = 0.46$, we have:

$$\pi\left(\widehat{\theta}^*; 0.49| \, 2.03\right) = \frac{(103.67)^{51}}{\Gamma(51)} \widehat{\theta}^{*50} \, exp^{-}\widehat{\theta}^* (103.67), \; 0 \leq \theta.$$

Thus, the posterior pdf of $\widehat{\theta}^*$ is the gamma pdf with shape parameter equal to 51 and scale parameter equal to $\frac{1}{103.67}$. We know the Bayes estimate under the square error loss function is the posterior mean.

Thus, the mean or expected value of the gamma pdf is:

$$E\left[\pi\left(\widehat{\theta}^*; 0.46|\underline{X}\right)\right] = \frac{n+1}{\widehat{\beta}^{-1} + n\overline{X}} = \frac{50+1}{2.17+101.5} = 0.492.$$

which is fairly close to the MLE of the parameter θ, the standard Bayes and empirical Bootstrap estimate.

The analytical form of the $100(1-\alpha)\%$ credible interval for the true parameter θ, under the jackknife empirical Bayes estimate, is given by:

$$\int_0^a \frac{(103.67)^{51}}{\Gamma(51)} \theta^{(50)} \, e^{-\theta(103.67)} d\theta = \frac{\alpha}{2},$$

and

$$\int_b^0 \frac{(103.67)^{51}}{\Gamma(51)} \theta^{(50)} \, e^{-\theta(103.67)} d\theta = \frac{\alpha}{2},$$

where $(a, \; b)$ is the credible interval and α is the level of significance.

The following R-code gives the 90% and 95% credible intervals.

```
> jackconf← rgamma(50, shape = 51, scale = .0096)
> library(EnvStats)
```

```
> eqgamma ( jackconf, p = 0.5, method = "mle", ci = TRUE, ci.type = "two-sided",
+ conf.level = 0.95, normal.approx.transform = "Kulkarni.powar", digits = 0)
```

Thus, the 90% credible interval under jackknife Bayes estimate $\widehat{\theta}$ is (0.47, 0.5). That is,

$$P[0.47 \leq \theta \leq 0.5] \geq 90\%.$$

The confidence range is $0.5 - 0.47 = 0.03$. Similarly, the 95% credible interval is (0.457, 0.495), that is,

$$P[0.459 \leq \theta \leq 0.495] \geq 95\%,$$

with a confidence range of 0.038.

Note that all the estimates of the true parameter θ, MLE, standard Bayes, empirical Bayes, bootstrapping, and jackknife, are all very close.

In the literature, there are many different empirical Bayes models and various applications are available. The purpose of this section is mainly to introduce the concept of empirical Bayes.

Exercises 10.6

10.6.1. You are given the following observation, $n = 60$, in Table 10.12, that characterizes the behavior of a certain phenomenon, A, about which we are interested in analyzing and learning as much as possible. Assume that the given data were randomly obtained.

- **(a)** Through goodness-of-fit testing identify, if possible, the pdf that characterizes probabilistically the behavior of phenomenon A, say, $f(x;\ \theta)$.
- **(b)** From (a) obtain the MLE of the parameter, or the parameter that drives the pdf, $f(x;\ \theta)$.
- **(c)** Determine the 90% and 95% confidence limits of the two parameters in $f(x;\ \theta_1, \theta_2)$.
- **(d)** Interpret the meanings and usefulness of (a), (b), and (c), with respect to phenomenon A.

10.6.2. For the data given in Exercise 10.6.1, assume that the parameter θ, in the pdf $f(x;\ \theta)$ that you have identified, behaves as a random variable with prior pdf, $\pi(\theta)$, which follows the exponential pdf. Assume a mean square error loss function, and proceed to answer the following questions:

- **(a)** What is the standard Bayes estimate of the true parameter θ?
- **(b)** What are the 90% and 95% credible intervals of the true parameter θ?
- **(c)** Interpret the meaning of your results in comparison with those found in Exercise 10.6.1.

10.6.3.

- **(a)** Obtain an empirical Bayes estimate of the true θ, using the data in Exercise 10.6.1, by estimating the prior $\pi(\theta)$ of the parameter θ, in $f(x|\theta)$, using the bootstrapping resampling method with an $n = 50$ sample.
- **(b)** Obtain 90% and 95% credible intervals for the true parameter θ.
- **(c)** Interpret your results.
- **(d)** Compare your findings with the MLE of θ, the standard Bayesian estimate of θ, and the empirical Bayes by bootstrapping of θ.

10.6.4. Repeat Exercise 10.6.3, but instead of using bootstrapping, use jackknife resampling and compare the four estimates of θ, that is, parametric, standard Bayes, empirical Bayes using bootstrapping, and jackknife estimates of the prior.

TABLE 10.12 The Data.

0.69	**0.54**	**0.08**	**2.33**	**0.47**	**0.88**	**0.07**	**1.31**	**0.19**	**0.15**
0.39	0.25	1.52	0.84	0.85	0.29	0.05	0.32	0.06	0.37
1.03	1.32	0.41	0.14	0.70	0.29	0.09	0.07	0.42	0.50
0.16	0.63	0.70	0.76	1.49	1.19	0.53	1.67	0.86	0.27
0.71	0.65	1.01	0.75	0.11	0.33	0.41	0.50	1.91	0.83
0.02	1.15	0.85	0.72	0.03	1.04	2.78	0.94	2.32	0.86

TABLE 10.13 Laboratory Mice Data.

0.56	2.06	1.12	1.34	0.50	1.49	0.67	1.09	1.10	0.81
2.54	0.50	0.65	2.60	1.36	0.19	1.91	1.28	1.92	0.35
0.33	1.24	0.18	0.36	3.53	0.87	0.87	0.80	3.68	1.34
1.90	0.11	2.90	0.77	0.87	1.04	6.37	1.54	1.60	1.09
2.05	0.41	2.86	0.34	0.75	0.66	3.47	0.13	1.73	1.21
0.93	1.36	0.10	0.18	4.88	0.95	0.26	1.84	0.85	2.15

10.6.5. We were told by a laboratory scientist that she conducted an experiment and measured the behavior of a certain characteristic of 60 mice, and the data she collected are given in Table 10.13.
The laboratory scientist also told us that the data follow a two-parameter gamma pdf, $f(x;\ \alpha,\ \beta)$.
(a) Through goodness-of-fit testing confirm the fact that the data follow a gamma pdf. Through the process you have identified the MLE of the shape parameter α and location parameter β.
(b) If in (a) the data follow the gamma pdf, $f(x;\ \alpha,\ \beta)$, find the 95% confidence limit on the true parameter α.
(c) Plot the pdf and its cumulative probability distribution of (a).

10.6.6. **(a)** If the given data follow the gamma pdf, assume the shape parameter behaves as a random variable with exponential pdf. Using mean square error obtain the standard Bayesian estimate of α.
(b) Obtain a 95% credible interval for the true parameter α and its confidence range. Interpret the meaning of your results.
(c) Compare and discuss the results of the parametric analysis in Exercise 10.6.5 with the standard Bayesian results.

10.6.7. **(a)** We want to estimate the prior pdf, $\pi(\alpha)$, rather than assume or guess it as in Exercise 10.6.6 using bootstrap resampling from the given data, that is, estimating the prior, $\pi(\widehat{\alpha})$, and proceed to obtain an empirical Bayes estimate of the true α.
(b) Obtain a 95% credible interval of the true parameter α and its confidence range.
(c) Compare the results of the parametric analysis, Exercise 10.6.5; standard Bayesian, Exercise 10.6.6; and empirical Bayes estimates using bootstrapping.

10.6.8. Repeat Exercise 10.6.7 (a), (b), and (c) using jackknife resampling to obtain the empirical Bayesian estimate of the shape parameter $\alpha\alpha$. Compare and discuss your current results with the results of Exercises 10.6.5, 10.6.6, and 10.6.7.

10.7 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. Empirical Bayes method calculations are illustrated through an example.

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
- Empirical Bayes

10.8 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 10.2 can be performed using Minitab. Refer to the book *Bayesian Computation Using Minitab,* by Jim Albert (Wadsworth, 1996). For R help, we suggest the book *Bayesian Computation with R* (second edition), by Jim Albert (Springer, 2009). The methods explained in this book can also be used in Chapter 13.

10.8.1 Examples with R

To do the R-codes in this section, download the R package *LearnBayes*.

EXAMPLE 10.8.1: Using the data of Example 10.2.1, write an R-code to obtain the posterior.

Solution

We use p = θ.

```
p=seq(0.8, 0.9, by = 0.02)
prior=c(0.13, 0.15, 0.22, 0.25, 0.15, 0.10)
prior=prior/sum(prior)
plot(p, prior, type="h", ylab="Prior Probability")
data=c(13, 2)
post=pdisc(p, prior, data)
post=pdisc(p, prior, data)
round(cbind(p, prior, post), 2)
```

Output:

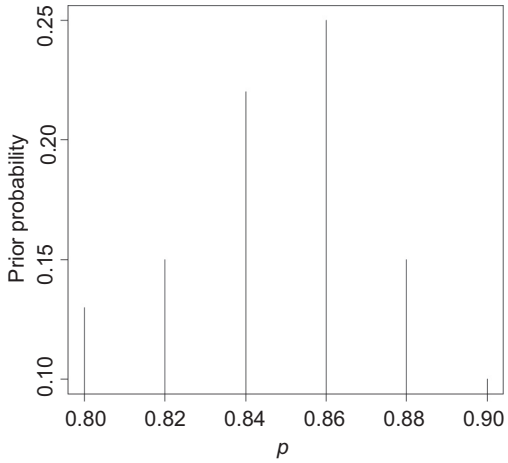


Figure: Discrete prior distribution for a proportion p.

```
p prior post
[1,] 0.80 0.13 0.11
[2,] 0.82 0.15 0.14
[3,] 0.84 0.22 0.23
[4,] 0.86 0.25 0.27
[5,] 0.88 0.15 0.16
[6,] 0.90 0.10 0.10
```

EXAMPLE 10.8.2 (Posterior calculation) Consider Example 10.2.4 with $\mu_p = 100$, $\sigma_p = 15$, and $x = 115$. Write an R-code to find the posterior.

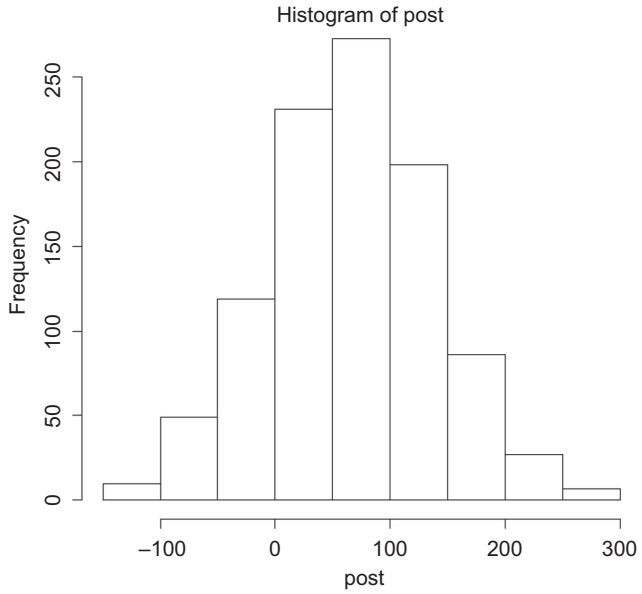
Solution

R-code

```
library (LearnBayes)
mup=100
sigmp=15
sigma=10
x=115
post=rnorm(1000,((sigmâ2*mup/(sigmp̂2+sigmâ2))+(sigmâ2*x/(sigmp̂2+sigmâ2))),(sigmâ2*sigmp̂2/(sigmp̂2+sigmâ2)))
post
hist(post)
```

Output

Along with many posterior sample values, we will get the following histogram for the posterior.



EXAMPLE 10.8.3 (Credible interval) Obtain a 95% credible interval for the posterior obtained in Example 10.8.2.

Solution

Once we have the posterior stored in post, the following will give us the credible interval.

R-code

```
quantile(post, c(0.025,0.5,0.975))
```

Output

2.5%	50%	97.5%
−76.84277	66.83870	207.86700

EXAMPLE 10.8.4 (Bayesian hypothesis testing)

The following are 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

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

Solution

R-code

```
y=c(.92, 7.42, 1.05, 1.76, 5.53, .01, 3.64, 2.69, -4.47, 1.54,
+ -2.60, 3.97, .71, 1.34, -3.66, -1.63, 1.38, -1.24, 3.87, -4.78)
pop.s=3
norpar=c(0,4) # vector of mean and standard deviation of the normal prior distribution
m0=0 # value of the normal mean to be tested
mnormt.onesided(m0,normpar,data)
```

Output

$BF *(Bayes factor in support of the null hypothesis)*

[1] 0

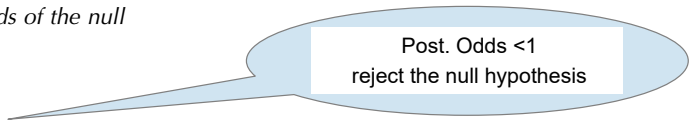
Post. Odds <1
reject the null hypothesis

$prior.odds *(prior odds of the null hypothesis)*

[1] 0.7621303

$post.odds *(posterior odds of the null hypothesis)*

[1] 0



$postH (posterior probability of null hypothesis)

[1] 0

Project for Chapter 10

10A 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 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:

$$f\Big(y|x) = \int f(y,\theta|x\Big)d\theta$$

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

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

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], and (2) $N(1, 1/16)$. Assume $f(x|\theta) \sim N(\theta, 1)$. Find the predictive distributions given the sample $X_1, X_2, \ldots, X_n$.