Math 541: Statistical Theory II

Bayesian Approach to Parameter Estimation

Lecturer: Songfeng Zheng

1 Prior Probability and Posterior Probability

Consider now a problem of statistical inference in which observations are to be taken from a distribution for which the pdf or the mass probability function is $f(x|\theta)$, where θ is a parameter having an unknown value. It is assumed that the unknown value of θ must lie in a specified parameter space Θ.

In many problems, before any observations from $f(x|\theta)$ are available, the experimenter or statistician will be able to summarize his previous information and knowledge about where in Θ the value of θ is likely to lie by constructing a probability distribution for θ on the set Θ. In other words, before any experimental data have been collected or observed, the experimenter's past experience and knowledge will lead him to believe that θ is more likely to lie in certain regions of Θ than in others. We would assume that the relative likelihoods of the different regions can be expressed in terms of a probability distribution on Θ. This distribution is called the *prior distribution* of θ, because it represents the relative likelihood that the true value of θ lies in each of various regions of Θ prior to observing any values from $f(x|\theta)$.

The concept of a prior distribution is very controversial in statistics. This controversy is closely related to the controversy regarding the meaning of probability. Some statisticians believe that a prior distribution can be chosen for the parameter θ in every statistics problem. They believe that this distribution is a subjective probability distribution in the sense that it represents an individual experimenter's information and subjective beliefs about where the true value of θ is likely to lie. They also believe, however, that a prior distribution is no different from any other probability distribution used in the field of statistics and that all the rules of probability theory apply to a prior distribution. This school of statisticians are said to adhere to the philosophy of Bayesian statistics.

Other statisticians believe that in many problems it is not appropriate to speak of a probability distribution of θ because the true value of θ is not a random variable at all, but is rather a certain fixed number whose value happens to be unknown to the experimenter. These statisticians believe that a prior distribution can be assigned to a parameter θ only when there is extensive previous information about the relevant frequencies with which θ has taken

each of its possible values in the past, so that two different scientists would agree on the correct prior distribution to be used. This group of statisticians are known as frequentism.

Both groups of statisticians agree that whenever a meaningful prior distribution can be chosen, the theory and methods to be described in this section are applicable and useful. The method of moment and maximum likelihood are not based on the the prior distribution, while the method to be introduced in this lecture will be based on the prior distribution.

Suppose now that the n random variables $X_1, \cdots, X_n$ form a random sample from a distribution for which the pdf or the point mass function is $f(x|\theta)$. Suppose also that the value of the parameter θ is unknown, and that the prior pdf or prior point mass function is $p(\theta)$. For simplicity, we shall assume that the parameter space Θ is either an interval of the real line or the entire real line, and that $p(\theta)$ is a prior pdf on Θ, rather than a prior point mass function. However, the discussion that will be given here can be easily adapted for a problem in which $f(x|\theta)$ or $p(\theta)$ is a point mass function.

Since the random variables $X_1, \cdots, X_n$ form a random sample from a distribution for which the pdf is $f(x|\theta)$, it follows that their joint pdf $f_n(x_1, \cdots, x_n|\theta)$ is given by

$$f_n(x_1, \cdots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$$

If we use the vector notation $\mathbf{x} = (x_1, \cdots, x_n)$, the above joint pdf could be written simply as $f_n(\mathbf{x}|\theta)$.

Since the parameter θ itself is now regarded as having a distribution for which the pdf is $p(\theta)$, the joint pdf $f_n(\mathbf{x}|\theta)$ could be regarded as the conditional joint pdf of $X_1, \cdots, X_n$ for a given value of θ. If we multiply this conditional joint pdf by the prior pdf $p(\theta)$, we obtain the $(n+1)$-dimensional joint pdf of $X_1, \cdots, X_n$ and θ in the form $f_n(\mathbf{x}|\theta)p(\theta)$. The marginal joint pdf of $X_1, \cdots, X_n$ can now be obtained by integrating this joint pdf over all values of θ. Therefore, the n-dimensional marginal joint pdf $g_n(\mathbf{x})$ of $X_1, \cdots, X_n$ can be written in the form

$$g_n(\mathbf{x}) = \int_\Theta f_n(\mathbf{x}|\theta)p(\theta)d\theta.$$

Furthermore, the conditional pdf of θ given that $X_1 = x_1, \cdots, X_n = x_n$, which will be denoted by $p(\theta|\mathbf{x})$, must be the joint pdf of $X_1, \cdots, X_n$ and θ divided by the marginal joint pdf of $X_1, \cdots, X_n$. Thus, we have

$$p(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)p(\theta)}{g_n(\mathbf{x})} \quad \text{for } \theta \in \Theta \tag{1}$$

The probability distribution over Θ represented by the conditional pdf in Eqn. 1 is called the *posterior distribution* of θ, because it is the distribution of θ after the values of $X_1, \cdots, X_n$ have been observed. We may say that the prior pdf $p(\theta)$ represents the relative likelihood, before the values of $X_1, \cdots, X_n$ have been observed, that the true value of θ lies in each of various regions of Θ. The prior distribution reflects our prior belief about the parameter θ.

On the other hand, the posterior pdf $p(\theta|\mathbf{x})$ represents the relative likelihood after the values $X_1 = x_1, \cdots, X_n = x_n$ have been observed. The posterior distribution reflects our belief about the parameter θ after we observed the data. Usually, the prior distribution is different from the posterior distribution, this means that the data can change our belief about the parameter.

The denominator on the right hand side of Eqn. 1 is simply the integral of the numerator over all the possible values of θ. Although the value of this integral depends on the observed values $x_1, \cdots, x_n$, it does not depend on θ in which we are interested, therefore it can be treated as a constant when the right hand side of Eqn. 1 is regarded as a pdf of θ. We may therefore replace Eqn. 1 with the following relation:

$$p(\theta|\mathbf{x}) \propto f_n(\mathbf{x}|\theta)p(\theta). \tag{2}$$

The proportionality symbol $\propto$ is used to indicate that the left side is equal to the right side except possibly for a constant factor, the value of which may depend on the observed values $x_1, \cdots, x_n$ but does not depend on θ. The appropriate constant factor which will establish the equality of the two sides in the relation (Eqn. 2) can be determined at any time by using the fact that $\int_\Theta p(\theta|\mathbf{x})d\theta = 1$, because $p(\theta|\mathbf{x})$ is a pdf of θ.

Recall that when the joint pdf or the joint point mass function $f_n(\mathbf{x}|\theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \cdots, x_n$, it is called the likelihood function. In this terminology, the relation (2) states that the posterior pdf of θ is proportional to the product of the likelihood function and the prior pdf of θ, i.e.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

By using the proportional relation (2), it is often possible to determine the posterior pdf of θ without explicitly performing the integration. If we can recognize the right hand side of the relation (2) as being equal to one of the standard pdf's except possibly for a constant factor, then we can easily determine the appropriate factor which will convert the right side of (2) into a proper pdf of θ. We will see some examples in the following.

2 Conjugate Prior Distributions

Certain prior distributions are particularly convenient for use with samples from certain other distributions. For example, suppose that a random sample is taken from a Bernoulli distribution for which the value of the parameter θ is unknown. If the prior distribution is a beta distribution, the for any possible set of observed sample values, the posterior distribution of θ will again be a beta distribution.

Theorem 1: *Suppose that $X_1, \cdots, X_n$ form a random sample from a Bernoulli distribution for which the value of the parameter θ is unknown ($0 < \theta < 1$). Suppose also that the prior*

distribution of θ is a beta distribution with given parameter α and β ($\alpha > 0$ and $\beta > 0$). Then the posterior distribution of θ given that $X_i = x_i$ ($i = 1, \cdots, n$) is a beta distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n - \sum_{i=1}^{n} x_i$.

Proof: Let $y = \sum_{i=1}^{n} x_i$, then the likelihood function, i.e., the joint pdf $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ is given by

$$f_n(\mathbf{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \cdots \theta^{x_n}(1-\theta)^{1-x_n} = \theta^y(1-\theta)^{n-y}$$

The prior distribution is

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

Since the posterior pdf $p(\theta|\mathbf{x})$ is proportional to the product $f_n(\mathbf{x}|\theta)p(\theta)$, it follows that

$$p(\theta|\mathbf{x}) \propto \theta^{y+\alpha-1}(1-\theta)^{\beta+n-y-1} \quad \text{for} \quad 0 < \theta < 1.$$

The right side of this relation can be recognized as being, except for a constant factor, equal to the pdf of a beta distribution with parameter $\alpha+y$ and $\beta+n-y$. Therefore the posterior distribution of θ is as specified in the theorem, and we can easily determine the constant as $\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}$.

An implication of Theorem 1 is the following: In Bernoulli distribution, suppose θ, the proportion of 1, is unknown, and we have n random variables sampled from this Bernoulli distribution, i.e. a series of 0 or 1. Suppose also that the prior distribution of θ is a beta distribution with parameters α and β. If the first random variable is a 1, the posterior distribution of θ will be a beta distribution with parameters $\alpha+1$ and β; if the first random variable is a 0, the posterior distribution of θ will be a beta distribution with parameters α and $\beta+1$. The process can be continued in this way. Each time, a random variable is obtained, the current posterior distribution of θ is changed to a new beta distribution in which the value of either the parameter α or the parameter β is increased by one unit: the value of α is increased by one unit if a 1 is observed, and the value of β is increased by one unit if a 0 is obtained.

The family of beta distribution is called a *conjugate family of prior distributions* for samples from a Bernoulli distribution. If the prior distribution of θ is a beta distribution, then the posterior distribution at each stage of sampling will also be a beta distribution, regardless of the observed values in the sample. It is also said that the family of beta distribution is *closed under sampling* from a Bernoulli distribution.

When samples are taken from a Poisson distribution, the family of gamma distributions is a conjugate family of prior distributions. This relationship is shown in the next theorem.

Theorem 2: *Suppose that $X_1, \cdots, X_n$ form a random sample from a Poisson distribution for which the value of the parameter θ is unknown ($\theta > 0$). Suppose also that the prior*

distribution of θ is a gamma distribution with given parameter α and β ($\alpha > 0$ and $\beta > 0$). Then the posterior distribution of θ given that $X_i = x_i$ ($i = 1, \cdots, n$) is a gamma distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$.

Proof: Let $y = \sum_{i=1}^{n} x_i$, then the likelihood function, i.e., the joint pdf $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ is given by

$$f_n(\mathbf{x}|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \frac{\theta^{x_1} e^{-\theta}}{x_1!} \cdots \frac{\theta^{x_n} e^{-\theta}}{x_n!} = \frac{\theta^y e^{-n\theta}}{x_1! \cdots x_n!}$$

The prior pdf of θ is

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for} \quad \theta > 0.$$

Since the posterior pdf $p(\theta|\mathbf{x})$ is proportional to the product $f_n(\mathbf{x}|\theta)p(\theta)$, it follows that

$$p(\theta|\mathbf{x}) \propto \theta^{y+\alpha-1} e^{-(\beta+n)\theta} \quad \text{for} \quad \theta > 0.$$

The right side of this relation can be recognized as being, except for a constant factor, equal to the pdf of a gamma distribution with parameter $\alpha + y$ and $\beta + n$. Therefore the posterior distribution of θ is as specified in the theorem, and we can easily determine the constant as $\frac{(\beta+n)^{\alpha+y}}{\Gamma(\alpha+y)}$.

When samples are taken from a normal distribution $N(\mu, \sigma^2)$, we will discuss two cases: μ is unknown but σ^2 is known, and σ^2 is unknown but μ is known. In these two cases, the conjugate prior distributions are different as the following two theorems show.

For the convenience of our discuss, we reparameterize the normal distribution, replacing σ^2 by $\xi = \sigma^2$; ξ is called the *precision* because the variance σ^2 measures the spread of the normal distribution. Under this representation, the normal distribution can be expressed as

$$f(x|\mu, \xi) = \left(\frac{\xi}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2}\xi(x-\mu)^2\right) \tag{3}$$

Theorem 3: *Suppose that $X_1, \cdots, X_n$ form a random sample from a normal distribution for which the value of the mean μ is unknown ($-\infty < \mu < \infty$), and the value of precision ξ ($\xi > 0$) is known to be ξ_0. Suppose also that the prior distribution of μ is a normal distribution with given parameter μ_p and ξ_p. Then the posterior distribution of μ given that $X_i = x_i$ ($i = 1, \cdots, n$) is a normal distribution with mean μ_1 and precision ξ_1 as*

$$\mu_1 = \frac{n\xi_0\bar{x} + \mu_p\xi_p}{n\xi_0 + \xi_p}$$

and

$$\xi_1 = n\xi_0 + \xi_p$$

Proof: Given the observed data $X_1 = x_1, \cdots, X_n = x_n$ and the prior distribution, we can write out the posterior distribution of μ as

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto f_n(\mathbf{x}|\mu) \times p(\mu) \\ &= \left(\frac{\xi_0}{2\pi}\right)^{n/2} \prod_{i=1}^{n} \exp\left(-\frac{\xi_0}{2}(x_i-\mu)^2\right) \times \left(\frac{\xi_p}{2\pi}\right)^{1/2} \exp\left(-\frac{\xi_p}{2}(\mu-\mu_p)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\xi_0 \sum_{i=1}^{n}(x_i-\mu)^2 + \xi_p(\mu-\mu_p)^2\right]\right) \end{aligned}$$

Here we have exhibited only the terms in the posterior density that depend on μ; the last expression above shows the shape of the posterior density as a function of μ. The posterior density itself is proportional to this expression, with a proportionality constant that is determined by the requirement that the posterior density integrates to 1.

We will now manipulate the expression for above to cast it in a form so that we can recognize that the posterior density is normal. We can write

$$(x_i-\mu)^2 = (x_i-\bar{x})^2 + (\mu-\bar{x})^2 + 2(x_i-\bar{x})(\bar{x}-\mu)$$

Summing up the two sides, it is easy to verify that

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\mu-\bar{x})^2$$

Therefore, we can absorb more terms that do not depend on μ into the constant of proportionality, and find

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto \exp\left(-\frac{1}{2}\left[\xi_0 \sum_{i=1}^{n}(x_i-\bar{x})^2 + n\xi_0(\mu-\bar{x})^2 + \xi_p(\mu-\mu_p)^2\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[n\xi_0(\mu-\bar{x})^2 + \xi_p(\mu-\mu_p)^2\right]\right) \end{aligned}$$

Now, observe that this is of the form $\exp(-Q(\mu)/2)$, where $Q(\mu)$ is a quadratic polynomial. We can write $Q(\mu)$ as

$$\begin{aligned} Q(\mu) &= n\xi_0(\mu-\bar{x})^2 + \xi_p(\mu-\mu_p)^2 \\ &= n\xi_0\mu^2 + n\bar{x}^2 - 2n\xi_0\bar{x}\mu + \xi_p\mu^2 + \xi_p\mu_p^2 - 2\xi_p\mu_p\mu \\ &= (n\xi_0+\xi_p)\left[\mu^2 - 2\frac{n\xi_0\bar{x}+\mu_p\xi_p}{n\xi_0+\xi_p}\mu + C_1\right] \\ &= (n\xi_0+\xi_p)\left[\mu - \frac{n\xi_0\bar{x}+\mu_p\xi_p}{n\xi_0+\xi_p}\right]^2 + C_2 \end{aligned}$$

where C_1 and C_2 are two constant expressions which do not depend on μ. Finally, we can write the posterior distribution of μ as

$$f(\mu|\mathbf{x}) \propto \exp(-Q(\mu)/2) \propto \exp\left(-\frac{n\xi_0+\xi_p}{2}\left[\mu - \frac{n\xi_0\bar{x}+\mu_p\xi_p}{n\xi_0+\xi_p}\right]^2\right)$$

Comparing the above expression to the normal distribution density function (Eqn. 3), we can recognize that the posterior distribution is a normal distribution with mean μ_1 and precision ξ_1 as

$$\mu_1 = \frac{n\xi_0\bar{x} + \mu_0\xi_p}{n\xi_0 + \xi_p}$$

and

$$\xi_1 = n\xi_0 + \xi_p$$

The mean value μ_1 of the posterior distribution of μ can be rewritten as follows:

$$\mu_1 = \frac{n\xi_0}{n\xi_0 + \xi_p}\bar{x} + \frac{\xi_p}{n\xi_0 + \xi_p}\mu_p$$

which means that μ_1 is a weighted average of the mean μ_p of the prior distribution and the sample mean $\bar{x}$. Furthermore, it can be seen that the relative weight given to $\bar{x}$ satisfies the following three properties: i) For fixed values of ξ_0 and ξ_p, the larger the sample size n, the greater will be the relative weight that is given to $\bar{x}$; ii) For fixed values of ξ_p and n, the smaller the precision ξ_0 of each observation in the sample, the smaller will be the relative weight that is given to $\bar{x}$; iii) For fixed value of ξ_0 and n, the smaller the precision ξ_p of the prior distribution, the larger will be the relative weight is given to $\bar{x}$.

Moreover, it can be seen that the precision ξ_1 of the posterior distribution of μ depends on the number n of the observations that have been taken but not depend on the magnitudes of the observed values. Suppose, therefore, that a random sample of n observations is to be taken from a normal distribution for which the value of the mean μ is unknown and the value of the variance (or precision) is known, and suppose that the prior distribution of μ is a specified normal distribution. Then, before any observations have been taken, we can calculate the value that the variance of the posterior distribution will have. However, the value of the mean μ_1 of the posterior distribution will depend on the observed values that are obtained in the sample.

Now, let us consider the case when $\xi_p \ll n\xi_0$, which would be the case if n were sufficiently large or if ξ_p were small (as for a very flat prior). In this case, the posterior mean would be $\mu_1 \approx \bar{x}$, which is the maximum likelihood estimate.

The result in Theorem 3 can also be expressed in terms of mean value and variance value, which is

$$\mu_1 = \frac{\sigma_0^2\mu_p + n\sigma_p^2\bar{x}}{\sigma_0^2 + n\sigma_p^2}$$

and

$$\sigma_1^2 = \frac{\sigma_0^2\sigma_p^2}{\sigma_0^2 + n\sigma_p^2}$$

where σ_0 and σ_p are the known variances of the sample distribution and prior distribution, respectively.

Theorem 4: *Suppose that $X_1, \cdots, X_n$ form a random sample from a normal distribution for which the value of the mean μ is known to be μ_0, and the value of precision ξ ($\xi > 0$) is unknown. Suppose also that the prior distribution of ξ is a gamma distribution with given parameter α and β. Then the posterior distribution of ξ given that $X_i = x_i$ ($i = 1, \cdots, n$) is a gamma distribution with parameters α_1 and precision β_1 as*

$$\alpha_1 = \alpha + \frac{n}{2}$$

and

$$\beta_1 = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

Proof: Given the observed data $X_1 = x_1, \cdots, X_n = x_n$ and the prior distribution, we can write out the posterior distribution of ξ as

$$\begin{aligned} f(\xi|\mathbf{x}) &\propto f_n(\mathbf{x}|\xi) \times p(\xi) \\ &= \left(\frac{\xi}{2\pi}\right)^{n/2} \prod_{i=1}^{n} \exp\left(-\frac{\xi}{2}(x_i - \mu_0)^2\right) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\xi^{\alpha-1}e^{-\beta\xi} \\ &\propto \xi^{\alpha+\frac{n}{2}-1} \exp\left(-\xi\left[\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right]\right) \end{aligned}$$

The right side of this relation can be recognized as being, except for a constant factor, equal to the pdf of a gamma distribution with parameter $\alpha + \frac{n}{2}$ and $\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2$. Therefore the posterior distribution of ξ is as specified in the theorem.

Now, by summarizing the above examples, we can give out the formal definition of conjugate priors: if the prior distribution belongs to a family G and, conditional on the parameters from G, the data have a distribution H, then G is said to be conjugate to H if the posterior is in the family G. Conjugate priors are used used for mathematical convenience (required integrations can be done in closed form) and they can assume a variety of shapes at the parameters of the prior are varied.

3 Bayes Estimator

3.1 Review of Estimator

Suppose we have an i.i.d. random sample $X_1, \cdots, X_n$ which is taken from a distribution $f(x|\theta)$ with the unknown parameter(s) θ. Suppose also that the value of θ must lie in a given interval Θ of the real line. We also assume the the value of θ must be estimated from the observed values in the sample.

An *estimator* of the parameter θ, based on the random variables $X_1, \cdots, X_n$, is a real-valued function $\delta(X_1, \cdots, X_n)$ which specifies the estimated value of θ for each possible set of values of $X_1, \cdots, X_n$. In other words, if the observed values of $X_1, \cdots, X_n$ turn out to be $x_1, \cdots, x_n$, then the estimated value of θ is $\delta(x_1, \cdots, x_n)$. Since the value of θ must belong to the interval Θ, it is reasonable to require that every possible value of an estimator $\delta(X_1, \cdots, X_n)$ must also belong to Θ.

It will often be convenient to use vector notation and to let $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{x} = (x_1, \cdots, x_n)$. In this notation, an estimator is a function $\delta(\mathbf{X})$ of the random vector $\mathbf{X}$, and an estimate is a specific value $\delta(\mathbf{x})$.

3.2 Loss Functions and Bayes Estimators

The foremost requirement of a good estimator of δ is that it yield an estimate of θ which is close to the actual value of θ. In other words, a good estimator is one for which it is highly probable that the error $\delta(\mathbf{X}) - \theta$ will be close to 0. We assume that for each possible value of $\theta \in \Theta$ and each possible estimate $a \in \Theta$, there is a number $L(\theta, a)$ which measures the loss or cost to the statistician when the true value of the parameter is θ and his estimate is a. Typically, the greater the distance between a and θ, the larger will be the value of $L(\theta, a)$.

As before, we let $p(\theta)$ denote the prior pdf of θ on the interval Θ, and consider a problem in which the statistician must estimate the value of θ without being able to observe the values in a random sample. If the statistician chooses a particular estimate a, then his expected loss will be

$$E[L(\theta, a)] = \int_\Theta L(\theta, a) p(\theta) d\theta \tag{4}$$

This is actually the average loss considering all the possible value of the true value of θ. We will assume that the statistician wishes to choose an estimate a for which the expected loss in Eqn. 4 is a minimum. In any estimation problem, a function L for which the expectation $E[L(\theta, a)]$ is to be minimized is called a loss function.

Suppose now that the statistician can observe the value $\mathbf{x}$ of the random vector $\mathbf{X}$ before estimating θ, and let $p(\theta|\mathbf{x})$ denote the posterior pdf of θ on Θ. For any estimate a the statistician might use, the expected loss will now be

$$E[L(\theta, a)|\mathbf{x}] = \int_\Theta L(\theta, a) p(\theta|\mathbf{x}) d\theta \tag{5}$$

Now, the statistician would now choose an estimate a for which the expectation in Eqn. 5 is minimum.

For each possible value $\mathbf{x}$ of the random vector $\mathbf{X}$, let $\delta^*(\mathbf{x})$ denote a value of the estimate a for which the expected loss in Eqn. 5 is a minimum. Then the function $\delta^*(\mathbf{X})$ for which the values are specified in this way will be an estimator of θ. This estimator is called a *Bayes*

estimator of θ. In other words, for each possible value $\mathbf{x}$ of $\mathbf{X}$, the value $\delta^*(\mathbf{x})$ of the Bayes estimator is chosen so that

$$E[L(\theta, \delta^*(\mathbf{x}))|\mathbf{x}] = \min_{a \in \Theta} E[L(\theta, a)|\mathbf{x}] \tag{6}$$

It should be emphasized that the form of the Bayes estimator will depend on both the loss function that is used in the problem and the prior distribution that is assigned to θ.

The Squared Error Loss Function: By far the most commonly used loss function in estimation problems is the squared error loss function. This function is defined as

$$L(\theta, a) = (\theta - a)^2.$$

When the squared error loss function is used, the Bayes estimate $\delta^*(\mathbf{x})$ for any observed value of $\mathbf{x}$ will be the value of a for which the expectation $E[(\theta - a)^2|\mathbf{x}]$ is a minimum.

It could be shown that for any given probability distribution of θ, the expectation of $(\theta - a)^2$ will be a minimum when a is chosen to be equal to the mean of the distribution of θ. Therefore, when the expectation of $(\theta - a)^2$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be equal to the mean $E(\theta|\mathbf{x})$ of the posterior distribution. This means that when the squared error loss function is used, the Bayes estimator is $\delta^*(\mathbf{X}) = E(\theta|\mathbf{X})$.

Example 1: *Estimating the Parameter of a Bernoulli Distribution.* Suppose that a random sample $X_1, \cdots, X_n$ is to be taken from a Bernoulli distribution for which the value of the parameter θ is unknown and must be estimated, and suppose that the prior distribution of θ is a beta distribution with given parameters α and β ($\alpha > 0$ and $\beta > 0$). Suppose also that the squared error loss function is used. We shall determine the Bayes estimator of θ.

For any observed values $x_1, \cdots, x_n$, let $y = \sum_{i=1}^n x_i$. Then it follows from Theorem 1 that the posterior distribution of θ will be a beta distribution with parameters $\alpha + y$ and $\beta + n - y$, therefore the mean value of this posterior distribution of θ is $(\alpha + y)/(\alpha + \beta + n)$. The Bayes estimate $\delta^*(\mathbf{x})$ will be equal to this value for any observed vector $\mathbf{x}$, i.e.

$$\delta^*(\mathbf{x}) = \frac{\alpha + \sum_{i=1}^n x_i}{\alpha + \beta + n}$$

Example 2: *Estimating the Mean of a Normal Distribution.* Suppose that a random sample $X_1, \cdots, X_n$ is to be taken from a normal distribution for which the value of the mean μ is unknown and the value of the variance σ^2 is known, and suppose that the prior distribution of μ is a normal distribution with given values of the mean μ_p and the variance σ_p^2. Suppose also that the squared error loss function is used. We shall determine the Bayes estimator of μ.

It follows from Theorem 3 that for any observed values $x_1, \cdots, x_n$, the posterior distribution of μ will be a normal distribution for which the mean μ_1 is

$$\mu_1 = \frac{\sigma^2 \mu_p + n\sigma_p^2 \bar{x}}{\sigma^2 + n\sigma_p^2}$$

Therefore the Bayes estimator $\delta^*(\mathbf{X})$ is specified as

$$\mu_1 = \frac{\sigma^2\mu_p + n\sigma_p^2\bar{X}}{\sigma^2 + n\sigma_p^2}$$

The Absolute Error Loss Function: Another commonly used loss function in estimation problems is the absolute error loss function, which is defined as

$$L(\theta, a) = |\theta - a|.$$

For any observed value of $\mathbf{x}$, the Bayes estimate $\delta(\mathbf{x})$ will now be the value of a for which the expectation $E(|\theta - a||\mathbf{x})$ is a minimum.

It could be shown that for any given probability distribution of θ, the expectation of $|\theta - a|$ will be a minimum when a is chosen to be equal to a median of the distribution of θ. Therefore, when the expectation of $|\theta - a|$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be equal to a median of the posterior distribution of θ. It follows that when the absolute error loss function is used, the Bayes estimator $\delta^*(\mathbf{X})$ is an estimator for which the value is always equal to a median of the posterior distribution of θ.

4 Exercises

Exercises 1: Let $p(\theta)$ be a pdf which is defined as follows, for constant $\alpha > 0$ and $\beta > 0$:

$$p(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}, & \text{for} \quad \theta > 0 \\ 0, & \text{for} \quad \theta \leq 0 \end{cases}$$

(a) Please verify that $p(\theta)$ is actually a pdf by verifying that $\int_0^\infty p(\theta)d\theta = 1$. (b) Consider the family of probability distributions that can be represented by a pdf $p(\theta)$ having the given form for all possible pairs of constants $\alpha > 0$ and $\beta > 0$. Show that this family is a conjugate family of prior for samples from a normal distribution with a known value of the mean μ and an unknown value of the variance θ.

Exercises 2: The Pareto distribution with parameters $x_0 > 0$ and $\alpha > 0$ is defined by the following pdf:

$$f(\theta|x_0, \alpha) = \begin{cases} \frac{\alpha x_0^\alpha}{\theta^{\alpha+1}}, & \text{for} \quad \theta > x_0 \\ 0, & \text{for} \quad \theta \leq x_0 \end{cases}$$

Show that the family of Pareto distribution is a conjugate family of prior distribution for sample $X_1, \cdots, X_n$ from a uniform distribution on the interval $(0, \theta)$, where the value of the parameter θ is unknown. Also find out the posterior distribution (determine the proportional coefficients in the distribution).

Exercises 3: Suppose that $X_1, \cdots, X_n$ form a random sample from a distribution for which the pdf $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1}, & \text{for} \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Suppose also that the value of the parameter θ is unknown ($\theta > 0$) and that the prior distribution of θ is a gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$). Determine the mean and the variance of the posterior distribution of θ.

Exercises 4: Suppose that a random sample of size n is taken from a Bernoulli distribution for which the value of the parameter θ is unknown, and that the prior distribution of θ is a beta distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\lim_{n\to\infty} \gamma_n = 1$.

Exercises 5: Suppose that a random sample of size n is taken from a Poisson distribution for which the value of the mean θ is unknown, and that the prior distribution of θ is a gamma distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\lim_{n\to\infty} \gamma_n = 1$.

Exercises 6: Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(0, \theta)$, where the value of the parameter θ is unknown ($\theta > 0$). Suppose also that the prior distribution of θ is a Pareto distribution with parameters x_0 and α ($x_0 > 0$ and $\alpha > 0$). Suppose, finally, that the value of θ is to be determined by using the squared error loss function. Determined the Bayesian estimator of θ.