BAYESIAN STATISTICS

Chapter 4

Instructor: Seokho Lee

Hankuk University of Foreign Studies

4. Prior Distributions and Posterior Distributions

4.1. Introduction

The **parameter**, θ, is a unknown quantity which may affect the statistical decision making. It is natural to consider all possible values of θ in the decision making. The set of all possible parameter values is called **parameter space**, denoting it by Θ.

To acquire the information on θ, we observe a random variable X repeatedly in the same condition. We call them a **sample** and denote them by $(X_1, X_2, \ldots, X_n)$.

The set of all possible values of a sample $(X_1, X_2, \ldots, X_n)$ is the **sample space** or Ω.

$$\Omega = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n | (X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\}$$

Usually the distribution of a random variable X depends on the unknown parameter θ. Thus, the probability distribution P_θ is defined as

$$P_\theta(A) = \begin{cases} \sum_{x \in A} f(x|\theta) & \text{if } X \text{ is discrete} \\ \int_A f(x|\theta) dx & \text{if } X \text{ is continuous} \end{cases}$$

for any subset A of the sample space Ω.

The **likelihood function** or **likelihood** is a function of a parameter θ given the value of the sample $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$:

$$\ell(\mathbf{x}|\theta) = \ell(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \quad \mathbf{x} \in \Omega.$$

The traditional statistics uses the likelihood function in the inference for θ and the decision making. The value of θ which achieves the maximum of the likelihood function is called a **maximum likelihood estimator** (**MLE**) and used in the estimation.

The **prior distribution** represents the prior information or belief on θ.

Bayesian statistics uses the **posterior distribution** for the estimation on θ. The posterior distribution is obtained by combining the prior information and data information.

4.2. Subjective prior distribution

The **subjective probability** of a certain event represents the subjective belief on how probable that event occurs.

Consider an event A. If you believe that the event A is likely to occur twice more than the its complement A^c, you may define

$$\Pr(A) = 2/3, \quad \Pr(A^c) = 1/3.$$

Contrast to the subjective probability, the **objective probability** relies on the relative frequency from the repeated experiments under the same condition. However, the objective probability is not available when the experiment cannot be repeated under the same condition. In such case, the subjective probability from the personal insight is useful.

The problem is that the subjective probability is too subjective.

4.3. Noninformative prior

Sometimes we don't have any information on the parameter *apriori*. The **noninformative prior** is used in Bayesian inference when there is no preference on the parameter.

The noninformative prior is called the **vague prior** meaning that the most probable value of the parameter is not highlighted or vague in the prior distribution.

Suppose we would like to estimate the amount of potassium (θ) in 100cc serum. We are supposed to know nothing but $\theta < 100$cc. Thus, the noninformative prior is given as the uniform distribution on $\Theta = [0, 100]$:

$$\pi(\theta) = \frac{1}{100}, \quad 0 \leq \theta \leq 100.$$

When the parameter space Θ is the bounded interval $[a, b]$ in the real line, the noninformative prior for θ is the uniform distribution on $[a, b]$ and the prior density function is

$$\pi(\theta) = \frac{1}{b-a}, \quad a \leq \theta \leq b.$$

Consider we are interested in the mean θ for the normal distribution. Its parameter space becomes a whole real line: $\Theta = \mathbb{R}$. If we are still ignorant of θ at all, then we should impose the same weight on all possible values of θ. Thus, the noninformative prior of θ should be the form of

$$\pi(\theta) = c, \quad \theta \in \mathbb{R}$$

for a relevant $c > 0$.

Note that $\pi(\theta)$ cannot be a probability density function because $\int_{-\infty}^{\infty} \pi(\theta) d\theta = \infty$ for any positive c.

However, this kind of prior is useful in Bayesian statistics. And, moreover, the value c is not important. Laplace proposed to use

$$\pi(\theta) = 1, \quad \theta \in \mathbb{R}.$$

The prior is called the **improper prior** if it is not integrable.

4.4. Calculation of posterior

When the random sample $(X_1, X_2, \ldots, X_n)$ is observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the **posterior probability function** is defined by

$$h(\theta|\mathbf{x}) = \frac{h(\theta, \mathbf{x})}{m(\mathbf{x})} = \begin{cases} \dfrac{\pi(\theta)\ell(\mathbf{x}|\theta)}{\sum_{\vartheta\in\Theta}\pi(\vartheta)\ell(\mathbf{x}|\vartheta)} & \text{if } \theta \text{ is discrete} \\ \dfrac{\pi(\theta)\ell(\mathbf{x}|\theta)}{\int_\Theta \pi(\vartheta)\ell(\mathbf{x}|\vartheta)d\vartheta} & \text{if } \theta \text{ is continuous} \end{cases}$$

where $h(\theta, \mathbf{x})$ is the joint distribution of θ and $\mathbf{X}$, $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, and $\ell(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$ is the likelihood function.

$m(\mathbf{x})$ in the denominator has nothing to do with the parameter θ. Thus, the posterior is proportional to the multiplication of the likelihood function and a prior:

$$h(\theta|\mathbf{x}) \propto h(\theta, \mathbf{x}) = \ell(\mathbf{x}|\theta)\pi(\theta).$$

Example (4-1)

We are interested in the success probability p of the binomial experiment $B(5, p)$. Suppose we observe $X = 2$ and our past experience suggests the prior for p is

$$\pi(\theta) = \begin{cases} 0.1 & p = 0.3 \\ 0.9 & p = 0.6. \end{cases}$$

Compute the posterior $h(p|2)$.

(solution) The likelihood function is

$$\ell(2|p) = \binom{5}{2} p^2(1-p)^{5-2} = 10p^2(1-p)^3.$$

Since the prior is positive only when $p = 0.3$ and 0.6, the posterior is positive on the same values. Thus,

$$h(0.3|2) = \frac{\pi(0.3)\ell(2|0.3)}{\pi(0.3)\ell(2|0.3)+\pi(0.6)\ell(2|0.6)} \approx 0.1296$$

$$h(0.6|2) = \frac{\pi(0.6)\ell(2|0.6)}{\pi(0.3)\ell(2|0.3)+\pi(0.6)\ell(2|0.6)} \approx 0.8704.$$

Note that the posterior is different from the prior after considering the information from the data (likelihood function).

Example (4-2)

A company recently launches a new product and is interested in whether it will make a hit. Suppose that p is the current market share. The company selects 5 consumers, introduces the new product to them, and asks whether they will buy it in the future. Compute the posterior distribution of the market share p after x of 5 consumers answer yes.

(solution) Let X be the number of consumers that say yes. Then $X \sim \mathrm{B}(5, p)$. Thus the likelihood function is

$$\ell(x|p) = \binom{5}{x} p^x (1-p)^{5-x}, \quad x = 0, 1, 2, 3, 4, 5.$$

The market share of a new product is between 0 and 1 and, usually, higher probability on the low market share and lower probability on the high market share. Therefore, we assume that the prior distribution is

$$\pi(p) = 2(1-p), \quad 0 < p < 1.$$

With them, the posterior distribution of p becomes

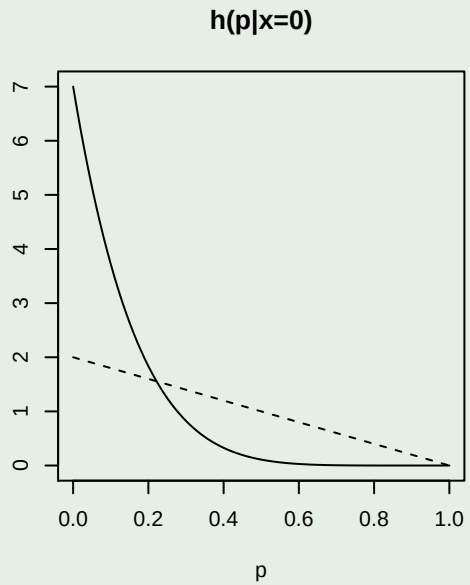
$$\begin{aligned} h(p|x) &= \frac{\pi(p)\ell(x|p)}{\int_{-\infty}^{\infty} \pi(u)\ell(x|u)du} = \frac{2(1-p)\binom{5}{x}p^x(1-p)^{5-x}}{\int_0^1 2(1-u)\binom{5}{x}u^x(1-u)^{5-x}du} \\ &= \frac{1}{B(x+1, 7-x)} p^{(x+1)-1}(1-p)^{(7-x)-1}. \end{aligned}$$

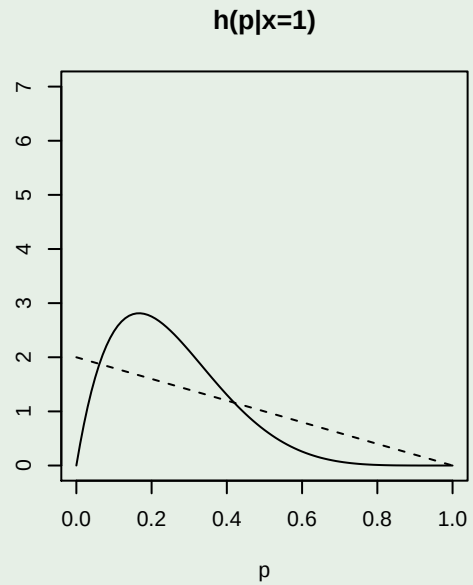
This is Beta$(x+1, 7-x)$, which is totally different from the prior distribution.

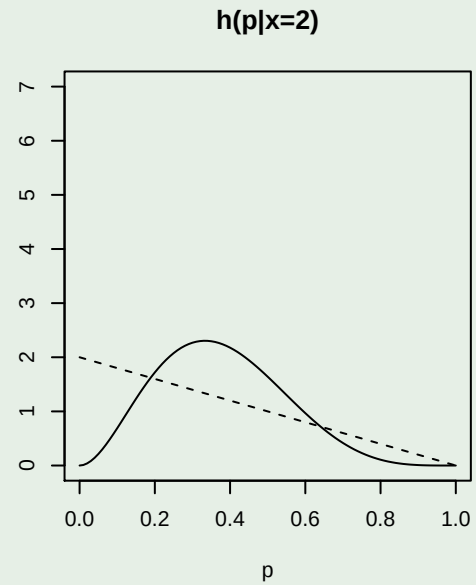
Example (4-2 continue)

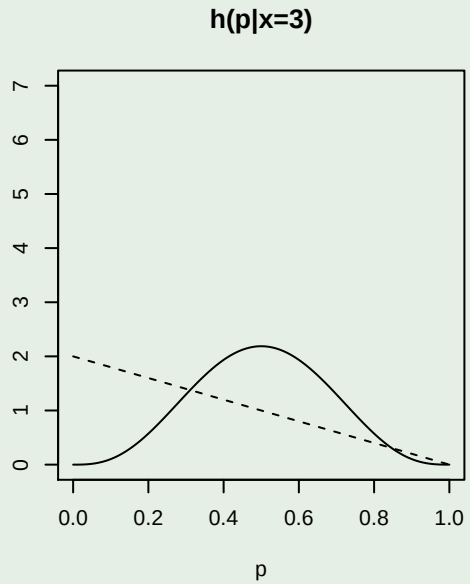
```
> p<-seq(0,1,by=0.01)
> par(mfrow=c(2,3))
> x<-0
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=0)')
> lines(p,2*(1-p),lty=2)
> x<-1
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=1)')
> lines(p,2*(1-p),lty=2)
> x<-2
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=2)')
> lines(p,2*(1-p),lty=2)
> x<-3
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=3)')
> lines(p,2*(1-p),lty=2)
> x<-4
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=4)')
> lines(p,2*(1-p),lty=2)
> x<-5
> plot(p,dbeta(p,x+1,7-x),ylim=c(0,7),type='l',xlab='p',ylab='',main='h(p|x=5)')
> lines(p,2*(1-p),lty=2)
> dev.copy2pdf(file='fig4-1.pdf')
```

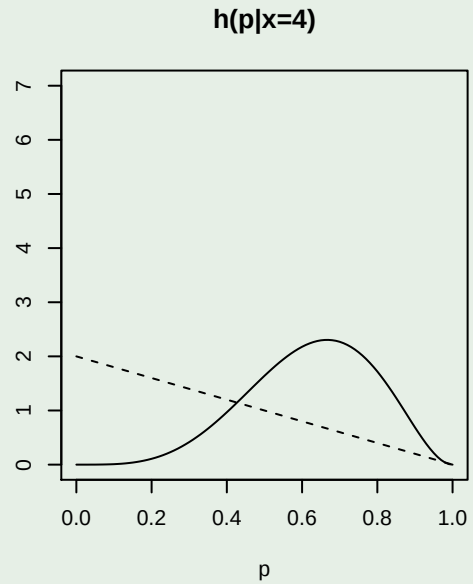
Example (4-2 continue)

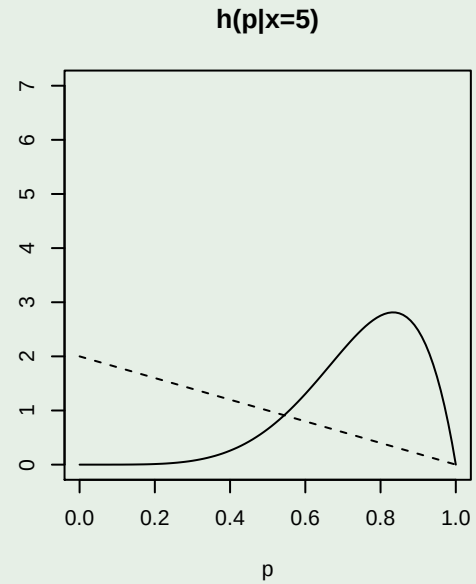












Example (4-3)

$X \sim N(\theta, 1)$ and the prior of θ is $N(0, 1)$. What is the posterior?

(solution)

$$\begin{aligned} h(\theta, x) &= \pi(\theta) f(x|\theta) = \frac{1}{2\pi} e^{-\frac{1}{2}\{\theta^2 + (x-\theta)^2\}} \\ &= \frac{1}{\sqrt{(2\pi)(1/2)}} e^{-\frac{1}{2}\frac{(\theta - x/2)^2}{1/2}} \cdot \frac{1}{\sqrt{(2\pi)(2)}} e^{-\frac{1}{2}\frac{x^2}{2}} \end{aligned}$$

and

$$m(x) = \int_{-\infty}^{\infty} h(\vartheta, x) d\vartheta = \frac{1}{\sqrt{(2\pi)(2)}} e^{-\frac{1}{2}\frac{x^2}{2}}.$$

Thus, the posterior of θ, $h(\theta|x) = h(\theta, x)/m(x)$, is $\mathrm{N}(x/2, 1/2)$.

4.5. Conjugate prior distribution

Generally, it is hard to compute the posterior distribution $h(\theta|x)$ analytically. Thus, the numerical analysis is used to its computation.

The **conjugate prior** is a prior distribution belonging to the distribution family to which the posterior distribution belongs.

The posterior distribution is easily computed if the conjugate prior is used.

The conjugate prior has the same form of the likelihood function. This fact is useful to find out a conjugate prior for the parameter.

Example (4-4)

Suppose that the random sample, $X_1, X_2, \ldots, X_n$, comes from Poisson distribution Poisson(θ). Find the conjugate prior distribution for θ.

(solution) The likelihood for the sample is

$$\ell(\theta|x) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{n\bar{x}} e^{-n\theta}}{\prod_{i=1}^{n} x_i!},$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$. When we view the likelihood as a function of θ, the likelihood has the same form of gamma distribution. Thus, we consider a gamma distribution as a prior distribution.
With a prior Gamma(α, β) with the relevant α and β,

$$h(\theta, \mathbf{x}) = \pi(\theta)\ell(\theta|\mathbf{x}) = C \cdot \theta^{n\bar{x}+\alpha-1} e^{-\theta(n+1/\beta)}, \quad \theta > 0.$$

Here $C^{-1} = \Gamma(\alpha)\beta^{\alpha} \prod_{i=1}^{n} x_i!$. The posterior is obtained by dividing $h(\theta, \mathbf{x})$ by $m(\mathbf{x})$. However, we don't necessarily compute $m(\mathbf{x})$ exactly because we can figure out the posterior is Gamma($n\bar{x} + \alpha, \frac{1}{n+1/\beta}$) without knowing $m(\mathbf{x})$.
The marginal distribution $m(\mathbf{x})$ is obtained as

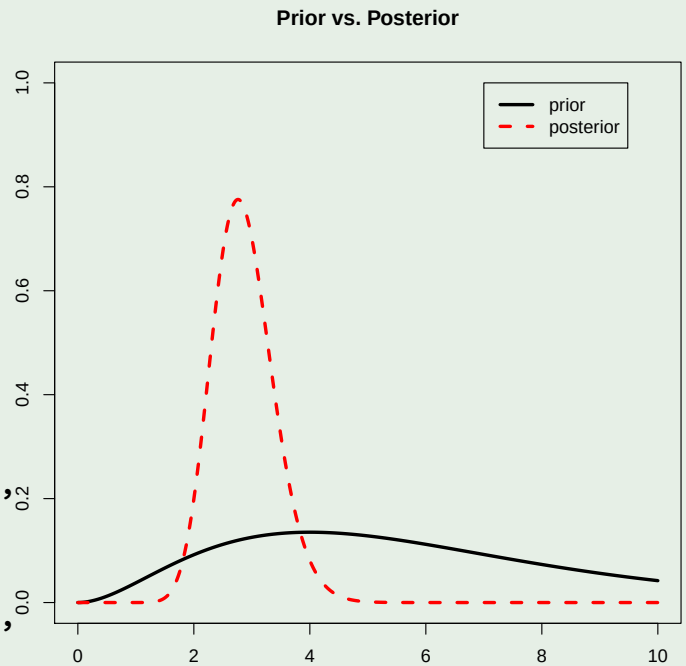
$$m(\mathbf{x}) = \frac{h(\theta,\mathbf{x})}{m(\mathbf{x})} = \frac{\Gamma(n\bar{x}+\alpha)(n+1/\beta)^{-(n\bar{x}+\alpha)}}{\Gamma(\alpha)\beta^{\alpha} \prod_{i=1}^{n} x_i!}.$$

Example (4-4 continue)

Let us schematize this:

$$\theta \sim \text{Gamma}(\alpha, \beta) \quad \xrightarrow[update]{d} \quad \theta|\mathbf{x} \sim \text{Gamma}(n\bar{x} + \alpha, \frac{1}{n+1/\beta})$$

```
> theta<-2 # true theta
> n<-10; X<-rpois(n,theta)
> X
 [1] 3 4 3 2 1 1 3 4 4 2
> alpha<-3; beta<-2 # parameters for prior
> z<-seq(0,10,by=0.01)
> plot(z,dgamma(z,shape=alpha,scale=beta),
type='l',xlab='',ylab='',ylim=c(0,1),
lwd=3) # prior
> alpha<-sum(X)+alpha # updated alpha
> beta<-1/(n+1/beta) # updated beta
> lines(z,dgamma(z,shape=alpha,scale=beta),
lty=2,lwd=3,col=2) # posterior
> title('Prior vs. Posterior')
> legend(7,1,legend=c('prior','posterior'),
lty=c(1,2),lwd=3,col=c(1,2))
> dev.copy2pdf(file='fig4-2.pdf')
```



sample distribution	conjugate prior	posterior
$X \sim \mathrm{B}(n, p)$	$p \sim \mathrm{Beta}(\alpha, \beta)$	$p\|X \sim \mathrm{Beta}(\alpha + X, n - X + \beta)$
$X \sim \mathrm{Poisson}(\theta)$	$\theta \sim \mathrm{Gamma}(\alpha, \beta)$	$\theta\|X \sim \mathrm{Gamma}(\alpha + X, (1 + 1/\beta)^{-1})$
$X \sim \mathrm{Exp}(\theta)$	$1/\theta = \lambda \sim \mathrm{Gamma}(\alpha, \beta)$	$\lambda\|X \sim \mathrm{Gamma}(\alpha + 1, (X + 1/\beta)^{-1})$
$X \sim \mathrm{N}(\mu, \sigma^2)$ (σ is known)	$\mu \sim \mathrm{N}(a, b^2)$	$\mu\|X \sim \mathrm{N}(\frac{Xb^2 + a\sigma^2}{b^2 + \sigma^2}, \frac{b^2\sigma^2}{b^2 + \sigma^2})$
$X \sim \mathrm{N}(\mu, \sigma^2)$ (μ is known)	$1/\sigma^2 = \tau \sim \mathrm{Gamma}(\alpha, \beta)$	$\tau\|X \sim \mathrm{Gamma}\left(\frac{1}{2} + \alpha, \left\{\frac{(X-\mu)^2}{2} + \frac{1}{\beta}\right\}^{-1}\right)$

Figure out why.

4.6. Improper prior and its posterior

Consider $X \sim \mathrm{N}(\theta, 1)$ and the improper prior $\pi(\theta) = 1$ $(-\infty < \theta < \infty)$. Figure out the posterior.

$$\begin{aligned} h(\theta, x) &= \pi(\theta) f(x|\theta) = f(x|\theta) \\ m(x) &= \int_{-\infty}^{\infty} h(\vartheta, x) d\vartheta = \int_{-\infty}^{\infty} f(x|\vartheta) d\vartheta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\vartheta)^2}{2}} d\vartheta = 1 \end{aligned}$$

Thus, the posterior $h(\theta|x)$ is

$$h(\theta|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}, \quad -\infty < \theta < \theta$$

which is $\mathrm{N}(x, 1)$. Note that the posterior is proper even though the prior is improper.

Suppose the prior $\pi(\theta) = c$ $(c > 0)$. Then,

$$h(\theta|x) = \frac{c \cdot f(x|\theta)}{\int_{\Theta} c \cdot f(x|\vartheta) d\vartheta} = \frac{f(x|\theta)}{\int_{\Theta} f(x|\vartheta)} d\vartheta.$$

So the constant c does not affect the posterior.