Chapter 15 Decision Theory and Bayesian Inference

Bayesian Estimation and Inference:

In this lecture we will discuss **Bayesian estimation**. A good reference for Bayesian methods in an introductory math-stat textbook is DeGroot's and Schervish's *Probability and Statistics*.

Consider our usual inference problem in which observations are drawn from a pdf $f(x;\theta)$ for some θ in a parameter space Ω.

In some cases, before any data are collected, it is possible for the statistician to use expert knowledge or previous information to construct a distribution on where in Ω the parameter θ is likely to be.

In that regard, we can think of Θ as a random variable with space Ω, and we call the pdf of Θ, $\pi(\theta)$, the pdf of the **prior distribution**. This terminology stems from the notion that we have specified the relative likelihood of the parameter falling in various regions of the parameter space prior to collecting any data.

Once a random sample from $f(x;\theta)$ has been collected, we can incorporate information from the likelihood function as well as the prior distribution to find the conditional distribution of Θ given the observations. This conditional distribution is known as the **posterior distribution**.

The posterior distribution has density function

$$k(\theta|x_1, ..., x_n) = \frac{\pi(\theta)L(\theta; x_1, ..., x_n)}{g(x_1, ..., x_n)}$$

where

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

and

$$g(x_1, ..., x_n) = \int_\Omega L(\theta; x_1, ..., x_n)\pi(\theta)d\theta.$$

If we consider the data as fixed, note that $k(\theta|x_1, ..., x_n) \propto \pi(\theta)L(\theta; x_1, ..., x_n)$ and $g(x_1, ..., x_n)$ can be viewed as the normalizing constant so that $k(\theta|x_1, ..., x_n)$ integrates to 1.

Example 1: As an example of this structure, consider an example taken from a clinical trial. Prien et al. (1984) studied three treatments for depression: imipramine, lithium carbonate, and a combination. An additional treatment arm received an inactive substance referred to as a **placebo**. A total of 150 patients were randomized to the four groups after an episode of depression, and after being treated they were followed to see how many had a relapse.

A Bayesian approach to analyzing such a trial would be to begin by incorporating the opinions of expert psychiatrists to specify prior distributions for the probability of success for each treatment.

For example, consider imipramine. Let Θ be the probability of no relapse after a period of treatment with imipramine. Θ can take possible values in $\Omega = (0, 1)$.
A popular model for Θ would be a beta-distribution with parameters α and β. In other words,

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

for $0 < \theta < 1$, $\alpha > 0$, $\beta > 0$.

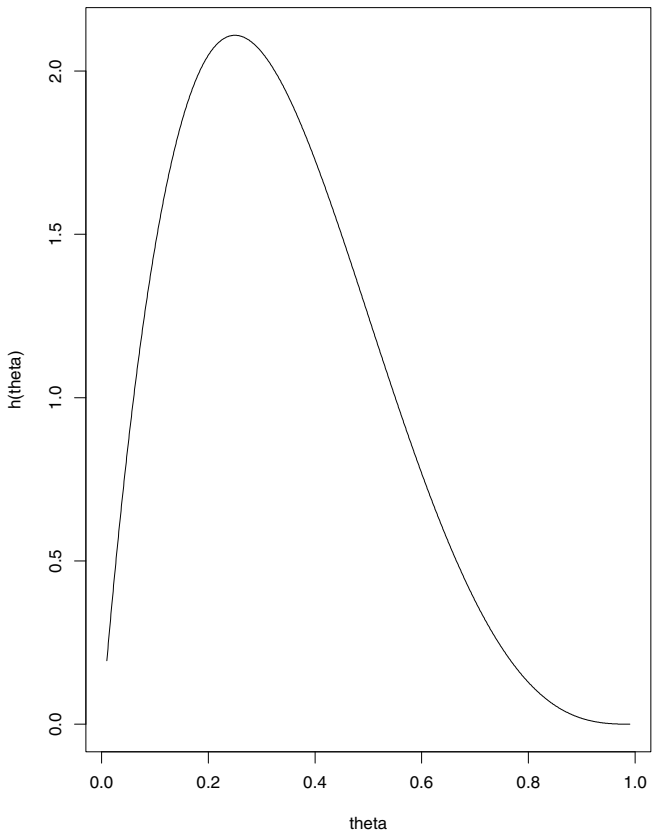
Recall that mean of a beta-distributed random variable is $\alpha/(\alpha+\beta)$ and the variance is $(\alpha\beta)/[(\alpha+\beta)^2(\alpha+\beta+1)]$.

Suppose the doctors agree that the likely success rate of imipramine will be about 1/3. Then, one possible choice for the prior distribution would be a beta-distribution with $\alpha = 2$ and $\beta = 4$.

```
theta<-seq(0.01,0.99,length=100)
h<-dbeta(theta,2,4)

plot(theta,h,xlab="theta",ylab="h(theta)",type="l")
```

Figure 1: Prior distribution for success of imipramine, beta(2,4).



Now let's look at the results of the study

Results

Group	Relapse	No Relapse	Success
Imipramine	18	22	55%
Lithium	13	25	66%
Combination	22	16	42%
Placebo	24	10	29%

Conditional on $\Theta = \theta$, the correct model for each of the 40 observations in the imipramine group is Bernoulli with success probability θ.

Thus, the posterior probability density function is

$$k(\theta|x_1, ..., x_{40}) \propto \theta^{22}(1-\theta)^{18}\pi(\theta)$$

where

$$\theta^{22}(1-\theta)^{18}\pi(\theta) \propto \theta^{22}(1-\theta)^{18}\theta^{1}(1-\theta)^{3}$$

Thus we see that

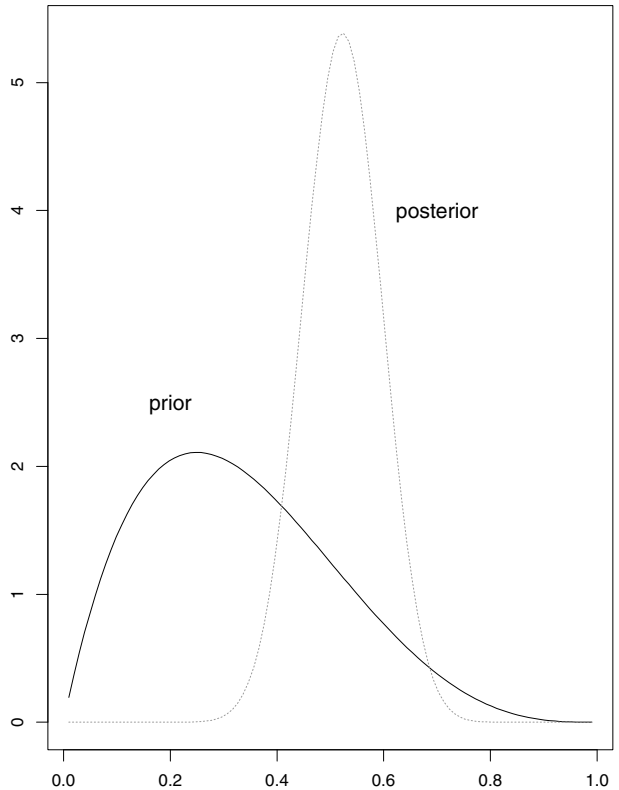
$$k(\theta|x_1, ..., x_{40}) \propto \theta^{23}(1-\theta)^{21}$$

which implies that the posterior distribution is a beta-distribution with $\alpha = 24$ and $\beta = 22$. Note that the mean of the posterior distribution is 24/(24+22)=.521, which is greater than the mean of the prior distribution, but slightly less than the observed success rate.

Let's look at plots of the prior and posterior densities.

```
theta<-seq(.01,.99,length=100)
prior<-dbeta(theta,2,4)
posterior<-dbeta(theta,24,22)
matplot(theta,cbind(prior,posterior),type="ll")
text(.2,2.5,"prior",cex=1.25)
text(.7,4,"posterior",cex=1.25)
```

Figure 2: Prior and posterior distributions for success of imipramine.



Now let's imagine that early in the study an interim analysis was done, and it was found that 3 out of 5 patients on imipramine had no relapse, slightly better than the final success rate.

Let's see what the posterior distribution would look like at this point in the study.

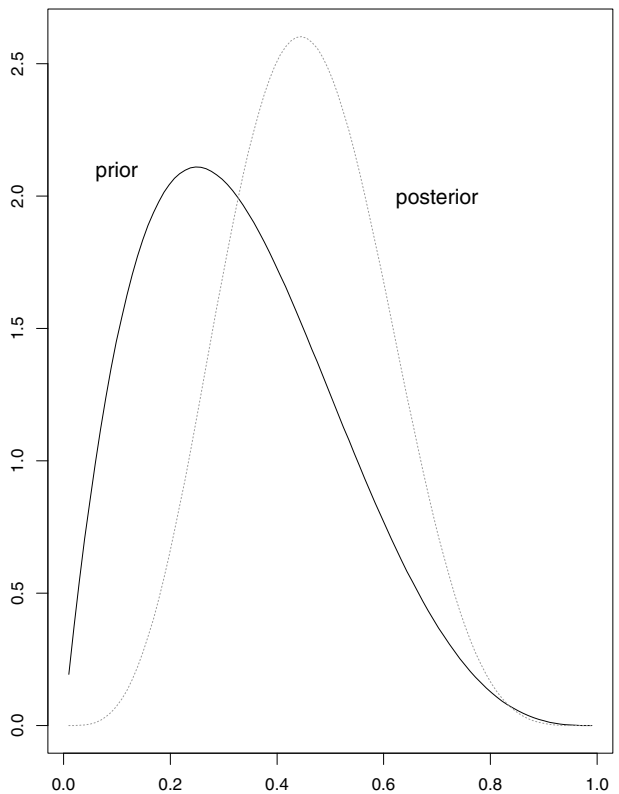
$$\begin{aligned} k(\theta|x_1, ..., x_2) &\propto \theta^3(1-\theta)^2\pi(\theta) \\ &\propto \theta^4(1-\theta)^5. \end{aligned}$$

Thus, the posterior distribution of Θ is a beta-distribution with $\alpha = 5$ and $\beta = 6$, which implies the posterior mean is $5/11$. We can see that before much data have been collected, the prior has a substantial influence on the posterior.

This diminshes with sample size, and the likelihood function begins to dominate.

```
theta<-seq(.01,.99,length=100)
prior<-dbeta(theta,2,4)
posterior<-dbeta(theta,5,6)
matplot(theta,cbind(prior,posterior),type="ll")
text(.1,2.1,"prior",cex=1.25)
text(.7,2,"posterior",cex=1.25)
```

Figure 3: Prior and posterior distributions for success of imipramine after a fictional analysis of 5 observations.



As you might imagine, use of prior distributions can be controversial. Some statisticians believe that a prior distribution is a subjective probability distribution that represents an experimenter's information and beliefs about the true value of θ, and that science can benefit from using this information in most statistical analyses. Such statisticians adhere to the Bayesian philosophy of statistics.

Other statisticians do not believe it is appropriate to think of θ as having a probability distribution and being a random variable. They feel it is better to simply view θ as a fixed but unknown point in the space of possible parameters. One objection they have about Bayesian methods is that two scientists with the same data but with different prior distributions may reach different conclusions.

In our example of a Bernoulli variable in a clinical trial, we saw that selecting a beta-distribution for the prior distribution of the success probability resulted in a posterior that was also a beta-distribution. Finding such situations is mathematically convenient, when possible.

If the prior distribution is chosen from a family of distributions such that the posterior distribution will also be in that family, the family of distributions is called a **conjugate family**.

Let's consider some examples of this.

Sampling from a Bernoulli Distribution

Suppose we are sampling from the distribution

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}$$

for $x = 0, 1$ and $0 < \theta < 1$. Also, suppose that a beta-distribution with $\alpha > 0$ a $\beta > 0$ is used as the prior distribution of Θ.

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

for $0 < \theta < 1$, $\alpha > 0$, $\beta > 0$. Then the posterior distribution is a beta-distribution with parameters $\alpha + \Sigma\, x_i$ and $\beta + n - \Sigma\, x_i$.

Proof: The probability density function of the posterior distribution is

$$k(\theta|x_1, ..., x_n) \propto L(\theta; x_1, ..., x_n)\pi(\theta)$$

$$\propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{\alpha+\sum x_i-1}(1-\theta)^{\beta+n-\sum x_i-1}.$$

Thus, the posterior must be a beta-distribution with the stated parameters.

Sampling from a Poisson Distribution

Suppose we are sampling from a Poisson distribution

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

for $x = 0, 1, 2, ...$ and $0 < \theta < \infty$. If we select the prior distribution to have a gamma distribution with parameters α and β, the posterior distribution will be gamma with parameters $\alpha + \Sigma\, x_i$ and $1/[n + 1/\beta]$.

$$\pi(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{-\theta/\beta}$$

for $\alpha > 0$, $\beta > 0$, and $0 < \theta < \infty$.

Recall that

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

Proof:

$$k(\theta|x_1, ..., x_n) \propto L(\theta; x_1, ..., x_n)\pi(\theta)$$

$$\propto \left[\prod_{i=1}^{n} \theta^{x_i} e^{-\theta}\right] \theta^{\alpha-1} e^{-\theta/\beta}$$

$$= \theta^{\alpha+\sum x_i-1} e^{-(n+\frac{1}{\beta})\theta}$$

Sampling from a normal distribution

Suppose that we are sampling from the pdf

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

for $-\infty < x < \infty$, σ^2 is known and $-\infty < \theta < \infty$. We use a normal prior distribution for θ with mean μ and variance λ^2.

$$\pi(\theta) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\lambda^2}}$$

Then after taking a random sample $X_1, ..., X_n$, the posterior distribution will be normal with mean

$$\eta = \frac{\sigma^2\mu + n\lambda^2\bar{x}}{\sigma^2 + n\lambda^2}$$

and variance

$$\omega^2 = \frac{\sigma^2 \lambda^2}{\sigma^2 + n\lambda^2}$$