7.4 Bayes Estimators

The prior and posterior distributions tell us how a Bayesian statisticians interprets parameters in light of data. But one of the fundamental goals of statistics is **estimating** the values of parameters.

Before we get to the problem of estimating parameters using *data*, it is worth discussing how one might estimate the value of a random variable in general.

Suppose Y is random with distribution $f(y)$; in particular, suppose we already know the formula for f, and based on this, want to make a prediction a for Y that is optimal in some sense. To determine what is optimal, we need to consider two factors:

1. The loss incurred if we guess a and the true value is y. We denote this loss as $L(y, a)$.
2. A decision rule determining what aspect of loss we wish to minimize.

There are several common choices of loss functions used throughout statistics, economics, and information theory. Each has its own relative advantages and disadvantages, which we won't dwell on too much here:

- Squared Loss: $L(y, a) = (y - a)^2$
- Absolute Loss: $L(y, a) = |y - a|$
- 0-1 loss: $L(y, a) = 1 - I_{y=a}$ *For discrete variables*
- 0-1 loss with ϵ tolerance: $L(y, a) = 1 - I_{|y-a|<\epsilon}$ *For continuous variables*
- Entropy loss: $L(y, a) = \frac{a}{y} - \log\left(\frac{a}{y}\right) - 1$

After determining the loss function $L(y, a)$, we still also need to determine a decision rule. Since Y is random variable, then $L(Y, a)$ is also a random variable, and so we need to decide what statistic for this random variable to minimize/maximize.

Some common examples:

- Minimize **expected** loss
- Minimize **maximal** loss (minimax decision rule)
- **Invariance** under change of coordinates

Once we have both a loss function and a decision rule, we can calculate the optimal estimate for Y.

Thm: Using squared loss $L(y, a) = (y - a)^2$ and minimizing expected loss, the optimal estimator of Y is the mean $a = E[Y]$.

Proof. Using estimate a, expected loss is

$$E[L(Y, a)] = E[(Y - a)^2] = E[Y^2] - 2aE[Y] + a^2.$$

Differentiating w.r.r. a:

$$\frac{\partial}{\partial} E[L(Y, a)] = -2E[Y] + 2a$$

which is 0 when $a = E[Y]$. Since the second derivative is positive, then $a = E[Y]$ corresponds to a local minimum. □

Thm: Using absolute loss $L(y, a) = |Y - a|$ and minimizing expected loss, the optimal estimator of Y is the median $a = \text{Med}(Y)$.

Proof. Note that $E[L(Y,a)] = E|Y-a|$; since the absolute value function isn't differentiable, we cannot solve this optimization problem by differentiating. Instead, we proceed by a different technique. Suppose that Y is a continuous variable, so that $f(y)$ is a density with $\int f(y)\,dy = 1$. Let $m = E[Y]$ and suppose that $a < m$. We'll show that $E[L(Y,a)] - E[L(Y,m)] \geq 0$.

$$\begin{aligned}
E[L(Y,a)] - E[L(Y,m)] &= \int_{-\infty}^{\infty} |y-a| - |y-m|\,dy \\
&= \int_{-\infty}^{a} (a-m)f(y)\,dy + \int_{a}^{m} [(y-a)-(m-y)]f(y)\,dy + \int_{m}^{\infty} (m-a)f(y)\,dy \\
&= \int_{-\infty}^{a} (a-m)f(y)\,dy + \int_{a}^{m} [2y-(a+m)]f(y)\,dy + \int_{m}^{\infty} (m-a)f(y)\,dy \\
&= (a-m)P(Y \leq a) + (m-a)P(Y \geq m) + \int_{a}^{m} [2y-(a+m)]f(y)\,dy \\
&\geq (a-m)P(Y \leq a) + (m-a)P(Y \geq m) + \int_{a}^{m} [2a-(a+m)]f(y)\,dy \\
&= (a-m)P(Y \leq a) + (m-a)P(Y \geq m) + (a-m)P(a < Y < m) \\
&= (m-a)[P(Y \geq m - P(Y < m)] \geq 0
\end{aligned}$$

□

Now, let's relate this discussion to the Bayesian framework for parameters.

Suppose θ is a parameter of a model, treated as a random variable with prior distribution ξ. Before observing data, we can estimate the value of θ using the loss-decision framework.

The estimate of θ that minimized expected loss with squared loss function is the mean of the prior distribution of θ.

Of course, rarely will we be estimating θ just based on the prior distribution. Instead, we would like to estimate θ in light of data, which means our formula for our estimate of θ should be a function of the data.

Def: Let $\mathbf{X}$ be observations whose joint distribution is indexed by $\theta \in \mathbb{R}$. An **estimator** of θ is a real-valued function $\delta(\mathbf{X})$. If $\mathbf{X} = \mathbf{x}$, then $\delta(\mathbf{x})$ is called the **estimate** of θ.

After observing data, we update the prior distribution $\xi(\theta)$ to produce the posterior distribution $\xi(\theta|\mathbf{x})$. But by definition, this is the distribution of the random variable $\theta|\mathbf{x}$. We have tools for finding the best estimates of random variables: loss functions and decisions rules.

That is, the expected loss that arises when estimation $\theta|\mathbf{x}$ using a is

$$E[L(\theta,a)|\mathbf{x}] = \int_{\Omega} L(\theta,a)\xi(\theta|\mathbf{x})\,d\theta$$

Since we should change our estimate a based on data, we can treat a as a function of $\mathbf{x}$. Define $\delta^*(\mathbf{x}) = a$ to be the value of a which minimizes the expected loss given the data $\mathbf{x}$.

Def: Let $L(\theta,a)$ be a loss function. The **Bayes Estimator** is the function $\delta^*(\mathbf{x})$ given by

$$\delta^*(\mathbf{x}) = \text{argmin}_a E[L(\theta,a)|\mathbf{x}]$$

Thm: Suppose we use a squared loss function $L(\theta,a) = (\theta-a)^2$. Then the Bayes estimator is $\delta^*(\mathbf{X}) = E[\theta|\mathbf{X}]$.

That is, using squared loss and minimizing expected loss, the best estimate for $\theta|\mathbf{x}$ is the mean of the conditional distribution $\xi(\theta|\mathbf{x})$.

Ex 1: When a DNA sequence is transcribed, errors are introduced at a rate of θ per 10^6 base pairs. Assume that θ is unknown, but the prior on θ is Gamma with $E[\theta] = \frac{2}{5}$ and $\alpha = 2, \beta = 5$. Suppose 2 defects are found in a sequence with $2 \cdot 10^7$ base pairs. What is the Bayes estimate of the average number of defects?

Solution. By assumption, the prior distribution of θ is

$$\xi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta} \propto \theta e^{-5\theta}$$

We assume that, given θ, the number of defects per 10^6 base pairs is approximately $Pois(\theta)$. For $1 \leq i \leq 20$, let X_i be the number of defects in the ith set of 10^6 base pairs. Note that $x_1 + \cdots + x_{20} = 2$. The likelihood function is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{20} f(x_i|\theta) = \prod_{i=1}^{20} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \propto e^{-20\theta}\theta^{x_1+\ldots x_{20}} = e^{-20\theta}\theta^2$$

Thus, the posterior distribution is

$$\xi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\xi(\theta) \propto e^{-25\theta}\theta^{4-1}$$

and so $\theta|\mathbf{x} \sim \text{Gamma}(4, 25)$.

The Bayes' estimate is $E[\theta|\mathbf{X}] = \frac{4}{25}$. ■

Def: A sequence of estimators δ_n^* of θ is **consistent** provided the sequence converges in probability to θ.

Thm: (Slutsky's Theorem) Let (X_n, Y_n) be a random vector. If X_n converges in distribution to the random variable X and Y_n converges in probability to the constant c, then

1. $X_n + Y_n$ converges in distribution to $X + c$
2. X_nY_n converges in distribution to cX
3. X_n/Y_n converges in distribution to X/c (provided $c \neq 0$)

The theorem remains true if all instances of convergence in distribution are replaced with convergence in probability.

Ex 2: The Bayes Estimator for θ is consistent.