5 Bayesian estimation

Bayesians probably do it.

5.1 Prior and posterior distributions

Bayesian statistics, (named after the Rev. Thomas Bayes, an amateur 18th century mathematician), represents a different approach to statistical inference. Data are still assumed to come from a distribution belonging to a known parametric family. However, whereas classical statistics considers the parameters to be *fixed but unknown*, the Bayesian approach treats them as random variables in their own right. Prior beliefs about θ are represented by the **prior distribution**, with a prior probability density (or mass) function, $p(\theta)$. The **posterior distribution** has posterior density (or mass) function, $p(\theta \mid x_1, \dots, x_n)$, and captures our beliefs about θ after they have been *modified* in the light of the observed data.

By Bayes' celebrated formula,

$$p(\theta \mid x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n \mid \theta)p(\theta)}{\int f(x_1, \dots, x_n \mid \phi)p(\phi)\, d\phi}.$$

The denominator of the above equation does not involve θ and so in practice is usually not calculated. Bayes' rule is often just written,

$$p(\theta \mid x_1, \dots, x_n) \propto p(\theta) f(x_1, \dots, x_n \mid \theta).$$

Example 5.1 Consider the Smarties example addressed in Example 2.1 (a) and suppose our prior belief is that the number of colours is either 5, 6, 7 or 8, with prior probabilities 1/10, 3/10, 3/10 and 3/10 respectively. On seeing the data $x =$'red, green, red' we have $f(x \mid k) = (k-1)/k^2$. Similarly, if the fourth Smartie is orange, $f(x \mid k) = (k-1)(k-2)/k^3$. Then

$x =$ **'red, green, red'**

k	$p(k)$	$f(x \mid k)$	$p(k)f(x \mid k)$	$p(k \mid x)$
5	.1	.160	.016	.13
6	.3	.139	.042	.33
7	.3	.122	.037	.29
8	.3	.109	.033	.26

$x =$ **'red, green, red, orange'**

k	$p(k)$	$f(x \mid k)$	$p(k)f(x \mid k)$	$p(k \mid x)$
5	.1	.096	.010	.11
6	.3	.093	.028	.31
7	.3	.087	.026	.30
8	.3	.082	.025	.28

There is very little modification of the prior. This analysis reveals, in a way that the maximum likelihood approach did not, that the data obtained from looking at just 4 Smarties is not very informative. However, as we sample more Smarties the posterior distribution will come to concentrate on the true value of k.

5.2 Conditional pdfs

The discrete case

Thus Bayesians statistics relies on calculation of conditional distributions. For two events A and B (measurable subsets of the sample space) with $\mathbb{P}(B) \neq 0$, we define

$$\mathbb{P}(A \mid B) := \mathbb{P}(A \cap B)/\mathbb{P}(B).$$

We can harmlessly agree to define $\mathbb{P}(A \mid B) := 0$ if $\mathbb{P}(B) = 0$.

If X and Y are RVs with values in $\mathbb{Z}$, and if $f_{X,Y}$ is their joint pmf:

$$\mathbb{P}(X = x; Y = y) = f_{X,Y}(x, y),$$

then we define

$$f_{X|Y}(x \mid y) := \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x; Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) \neq 0$. We can safely define $f_{X|Y}(x \mid y) := 0$ if $f_Y(y) = 0$. Of course,

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x; Y = y) = \sum_x f_{X,Y}(x, y).$$

Example 5.2 *Suppose that X and R are independent RVs, where X is Poisson with parameter λ and R is Poisson with parameter μ. Let $Y = X + R$.*

Then

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{y-x} e^{-\mu}}{(y-x)!} \Bigg/ \sum_{x,r:x+r=y} \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^r e^{-\mu}}{r!} \\ &= \frac{y!}{x!(y-x)!} \lambda^x \mu^{(y-x)} \Bigg/ \sum_{x,r:x+r=y} \frac{y!}{x!r!} \lambda^x \mu^r \\ &= \binom{y}{x} \left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{y-x}. \end{aligned}$$

Hence $(X \mid Y = y) \sim B(y, p)$, where $p = \lambda/(\lambda + \mu)$.

The continuous case

Let $Z = (X, Y)$ be a RV with values in $\mathbb{R}^{m+n}$, X having values in $\mathbb{R}^m$ and Y values in $\mathbb{R}^n$. Assume that Z has nice pdf $f_Z(z)$ and write

$$f_Z(z) = f_{X,Y}(x, y), \qquad (z = (x, y), x \in \mathbb{R}^m, y \in \mathbb{R}^n).$$

Then the pdf of Y is given by

$$f_Y(y) = \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, dx.$$

We define $f_{X|Y}$, the conditional pdf of X given Y, by

$$f_{X|Y}(x \mid y) := \begin{cases} f_{X,Y}(x,y)/f_Y(y) & \text{if } f_Y(y) \neq 0, \\ 0 & \text{if } f_Y(y) = 0. \end{cases}$$

The intuitive idea is: $\mathbb{P}(X \in dx \mid Y \in dy) = \mathbb{P}(X \in dx; Y \in dy)/\mathbb{P}(Y \in dy)$.

Examples 5.3

(a) *A biased coin is tossed n times. Let x_i be 1 or 0 as the ith toss is or is not a head. Suppose we have no idea how biased the coin is, so we place a uniform prior distribution on θ, to give a so-called 'noninformative prior' of*

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1\,.$$

Let t be the number of heads. Then the posterior distribution of θ is

$$p(\theta \mid x_1, \ldots, x_n) = \theta^t (1-\theta)^{n-t} \times 1 \Big/ \int_0^1 \phi^t (1-\phi)^{n-t} \times 1\, d\phi\,.$$

We would usually not bother with the denominator and just write

$$p(\theta \mid x) \propto \theta^t (1-\theta)^{n-t}\,.$$

By inspection we recognise that if the appropriate constant of of proportionality is inserted on the r.h.s. then we have the density of beta$(t+1, n-t+1)$, so this is the posterior distribution of θ given x.

(b) *Suppose $X_1, \ldots, X_n \sim N(\mu, 1)$, $p(\mu) \sim N(0, \tau^{-2})$ for known τ^{-2}.* Then

$$\begin{aligned} p(\mu \mid x_1, \ldots, x_n) &\propto \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^n (x_i - \mu)^2 + \mu^2\tau^2\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(n+\tau^2)\left(\mu - \frac{\sum x_i}{n+\tau^2}\right)^2\right\} \\ \mu \mid x_1, \ldots, x_n &\sim N\left(\frac{\sum x_i}{n+\tau^2}, \frac{1}{n+\tau^2}\right) \end{aligned}$$

Note that as $\tau \to 0$ the prior distribution becomes less informative.

(c) *Suppose $X_1, \ldots, X_n \sim$ IID $\mathcal{E}(\lambda)$, and the prior for λ is given by $\lambda \sim \mathcal{E}(\mu)$, for fixed and known μ.* Then

$$p(\lambda \mid x_1, \ldots, x_n) \propto \mu e^{-\lambda\mu} \prod_i \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(\mu + \sum_{i=1}^n x_i)}\,,$$

i.e., gamma$(n+1, \mu + \sum x_i)$.

5.3 Estimation within Bayesian statistics

The Bayesian approach to the parameter estimation problem is to use a **loss function** $L(\theta, a)$ to measure the loss incurred by estimating the value of a parameter to be a when its true value is θ. Then $\hat{\theta}$ is chosen to minimize $\mathbb{E}[L(\theta, \hat{\theta})]$, where this expectation is taken over θ with respect to the posterior distribution $p(\theta \mid x)$.

Loss functions for quadratic and absolute error loss

(a) $L(\theta, a) = (a-\theta)^2$ is the **quadratic error loss** function.

$$\mathbb{E}[L(\theta, a)] = \int L(\theta, a) p(\theta \mid x_1, \ldots, x_n)\, d\theta = \int (a-\theta)^2 p(\theta \mid x_1, \ldots, x_n)\, d\theta\,.$$

Differentiating with respect to a we get

$$2\int (a-\theta) p(\theta \mid x_1, \ldots, x_n)\, d\theta = 0 \implies a = \int \theta p(\theta \mid x_1, \ldots, x_n)\, d\theta\,.$$

Therefore quadratic error loss is minimized by taking $\hat{\theta}$ to be the **posterior mean**.

(b) $L(\theta, a) = |a - \theta|$ is the **absolute error loss** function.

$$\begin{aligned} \mathbb{E}[L(\theta, a)] &= \int L(\theta, a) p(\theta \mid x_1, \ldots, x_n)\, d\theta \\ &= \int_{\theta=-\infty}^{a} (a-\theta) p(\theta \mid x_1, \ldots, x_n)\, d\theta + \int_a^\infty (\theta - a) p(\theta \mid x_1, \ldots, x_n)\, d\theta\,. \end{aligned}$$

Differentiating with respect to a we find that the minimum is where

$$\int_{-\infty}^a p(\theta \mid x_1, \ldots, x_n)\, d\theta - \int_a^\infty p(\theta \mid x_1, \ldots, x_n)\, d\theta = 0\,.$$

The minimum is achieved when both integrals are equal to $\frac{1}{2}$, i.e., by taking $\hat{\theta}$ to be the **posterior median**.

Example 5.4 *Let $X_1, \ldots, X_n \sim P(\lambda)$, $\lambda \sim \mathcal{E}(1)$ so that $p(\lambda) = e^{-\lambda}$, $\lambda \geq 0$.*

The posterior distribution is

$$p(\lambda \mid x_1, \ldots, x_n) = e^{-\lambda} \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \propto e^{-\lambda(n+1)} \lambda^{\sum x_i}\,,$$

i.e., gamma$\big(\sum x_i + 1, (n+1)\big)$. So under quadratic error loss,

$$\hat{\theta} = \text{posterior mean} = \frac{\sum_{i=1}^n x_i + 1}{n+1}\,.$$

Under absolute error loss, $\hat{\theta}$ solves

$$\int_0^{\hat{\theta}} \frac{e^{-\lambda(n+1)} \lambda^{\sum x_i} (n+1)^{\sum x_i + 1}}{(\sum x_i)!}\, d\lambda = \frac{1}{2}\,.$$