Bayesian Inference

Informally, to make inference about θ is to learn about the unknown θ from data X, i.e., based on the data, explore which values of θ are probable, what might be plausible numbers as estimates of different components of θ and the extent of uncertainty associated with such estimates. In addition to having a model $f(x|\theta)$ yielding a likelihood function, the Bayesian needs a distribution for θ. The distribution is called a prior distribution or simply a prior because it quantifies her uncertainty about θ prior to seeing data. The prior may represent a blending of her subjective belief and knowledge, in which case it would be a subjective prior. Alternatively, it could be a conventional prior supposed to represent small or no information. Such a prior is called an objective prior.

Given all the above ingredients, the Bayesian calculates the conditional probability density of θ given $X = x$ by Bayes formula. First, the joint density of X and θ is $h(x, \theta) = f(x|\theta)\pi(\theta)$ on $\mathcal{X} \times \Theta$. The marginal (or predictive) density of X is

$$m(x) = \int_\Theta f(x|\theta)\pi(\theta)\,d\theta, \text{ or } \sum_\theta f(x|\theta)\pi(\theta).$$

Then,

$$\begin{aligned}\pi(\theta|x) &= \frac{h(x,\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{m(x)} \\ &\propto f(x|\theta)\pi(\theta) \text{ for observed data, } x,\end{aligned}$$

is called the post-experimental or posterior distribution (density) of θ given x.

This summarizes all the post-data information about θ, and is a quantification of our uncertainty about θ in the light of data. The transition from $\pi(\theta)$ to $\pi(\theta|x)$ is what we have learnt from the data. All inferences about θ must be based on this posterior distribution. Nothing beyond $l(\theta|x)$ from the experiment is needed. Two different experiments with the same likelihood lead to identical inference. $\pi(\theta|x) \propto l(\theta|x) \propto f(x|\theta)$ if $\pi(\theta) \equiv 1$. In this case $\pi(\theta|x)$ does not use any information other than what is in $l(\theta|x)$.

Suppose $T = T(X)$ is sufficient for θ (or $P_\theta, \theta \in \Theta$) or for $f(x|\theta)$, $\theta \in \Theta$.

Theorem. Posterior distribution of θ given $X = x$ depends on x only through $T(x)$.

Proof. We will assume the factorization theorem: $f(x|\theta) = g(T(x), \theta)h(x)$. If $T(x) = t$, then

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|u)\pi(u)\,du} \\ &= \frac{g(T(x),\theta)h(x)\pi(\theta)}{\int g(T(x),u)h(x)\pi(u)\,du} \\ &= \frac{g(t,\theta)\pi(\theta)}{\int g(t,u)\pi(u)\,du}. \end{aligned}$$

Example. Consider an urn with Np red and $N(1-p)$ black balls, p is unknown but N is a known large number. Balls are drawn at random one by one and with replacement, selection is stopped after n draws. For $i = 1, 2, \ldots, n$, let

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th ball drawn is red;} \\ 0 & \text{otherwise.} \end{cases}$$

Then Y_i's are i.i.d Binomial$(1, p)$, i.e., *Bernoulli* with probability of success p. Therefore the likelihood function for p given the data is proportional to

$$f(y_1, \ldots, y_n) = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} = p^x(1-p)^{n-x},$$

where $x = \sum_{i=1}^n y_i$. Since $X = \sum_{i=1}^n Y_i$ (the number of red balls drawn) is sufficient for p, we get the same likelihood function (the part involving p) if we assume that we have observed $X = x$ from a Binomial(n, p). Let p have a prior distribution $\pi(p)$. We will consider a family of priors for p that simplifies the calculation of posterior distribution and then consider some commonly used priors from this family. Let

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \le p \le 1; \alpha > 0, \beta > 0.$$

This is the density of the Beta distribution. Equivalently, under the prior distribution, the unknown parameter, $p \sim$ Beta(α, β). (Note that for convenience we take p to assume all values between 0 and 1, rather than only $0, 1/N, 2/N$, etc.) The prior mean and variance are $\alpha/(\alpha+\beta)$ and $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$, respectively, which may be obtainable from past data.

To derive the posterior density, note that

$$\begin{aligned} h(x,p) &= \binom{n}{x}p^x(1-p)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, x = 0, \ldots, n; 0 < p < 1 \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, x = 0, \ldots, n; 0 < p < 1, \end{aligned}$$

so that

$$\begin{aligned} m(x) &= \int_0^1 h(x,p)\,dp = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\,dp \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}, x=0,\ldots,n. \end{aligned}$$

Therefore, we obtain,

$$\begin{aligned} \pi(p|x) &= \frac{h(x,p)}{m(x)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, 0<p<1. \end{aligned}$$

i.e., $p|X=x \sim \mathrm{Beta}(x+\alpha, n-x+\beta)$. Note, however, that the computation of $m(x)$ is not needed here to derive the posterior density; it can be deduced from simply noting the functional form of the density in $h(x,p)$, namely, $p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$, which is just the (unnormalized) density of the Beta distribution. This is due to the choice of the prior, and will be explored further later. Before that note the following. As a straightforward and immediate estimate of p, one could look at the most 'probable' value of p (under $\pi(p|x)$). The *highest posterior density* or HPD estimate of p is the value, denoted $\hat{p}_{\text{hpd}}$, which maximizes $\pi(p|x)$. In the example above, $\hat{p}_{\text{hpd}} = (x+\alpha-1)/(n+\alpha+\beta-2)$.
(i) If we take $\alpha = 1 = \beta$, i.e., $\pi(p) \equiv 1$, we get

$$\pi(p|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}p^x(1-p)^{n-x}, 0<p<1.$$

As a function of p, $\pi(p|x)$ and $l(p|x)$ are the same. Therefore,
MLE of $p = x/n =$ HPD estimate of p. i.e., $\hat{p}_{\text{mle}} = \hat{p}_{\text{hpd}}$, but their interpretations are different.
Given x, $\hat{\theta}_{\text{hpd}}$ is the most probable value of θ, or $P_{\hat{\theta}_{\text{hpd}}}$ is most 'probably' the correct model (for X), whereas $\hat{p}_{\text{mle}}$ is that value of θ, or the parameter of that model which most 'likely' produced x.

Now, coming back to the Beta prior for Binomial, note that what simplied the computation of the posterior density is that the prior and likelihood have the same functional form.