III. Posterior and Prior Distributions – Basic Setup –

A. Let **θ** and **Y** be two events, then in classical probability theory:

$$\mathbf{P(\theta|Y)} = \frac{P(\theta \cap Y)}{P(Y)} \quad \text{and} \quad \mathbf{P(Y|\theta)} = \frac{P(\theta \cap Y)}{P(\theta)}$$

hence

$$\mathbf{P(\theta \cap Y) = P(\theta|Y)P(Y) = P(Y|\theta)P(\theta)}$$

and

$$\mathbf{P(\theta|Y)} = \frac{P(Y\,|\,\theta)P(\theta)}{P(Y)}$$

In the Bayesian framework, **Y is the observed data**, the **θ are the parameters**, **P(Y|θ) is the joint distribution of the sample**, **P(θ)** is the *prior distribution* of the parameters, **P(Y)** is the marginal distribution of the sample, and **P(θ|Y)** is the *posterior distribution*. Because **P(Y)** is a constant, the posterior distribution is proportional to the product of the joint distribution of the sample (which is proportional to the likelihood function) and the prior distribution; that is:

$$\mathbf{P(\theta|Y) \propto P(Y|\theta)P(\theta)}$$

B. Suppose we have a ***random sample*** (a set of independent and identically distributed random variables) from a probability distribution **f(x|θ).** Then the joint distribution of the sample is:

$$f_n(x_1, x_2, x_3, \ldots, x_n|\theta) = f_n(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i\theta)$$

This is commonly referred to as a ***likelihood function***. (Technically, when $\mathbf{f_n(\underline{x}|\theta)}$ is regarded as a function of $\boldsymbol{\theta}$ for a given vector $\underline{\mathbf{x}}$, it is called a likelihood function. Note that $\boldsymbol{\theta}$ is *not* a random variable in this context.) If $\boldsymbol{\theta}$ is a random variable then $\mathbf{f_n(\underline{x}|\theta)}$ is equal to the ratio of the joint distribution of the sample and $\boldsymbol{\theta}$ to the marginal distribution of $\boldsymbol{\theta}$ using the standard formula for conditional probability:

$$f_n(x_1, x_2, x_3, \ldots, x_n \mid \theta) = \frac{h(x_1, x_2, x_3, \ldots, x_n, \theta)}{\xi(\theta)}$$

and

$$h(x_1, x_2, x_3, \ldots, x_n, \theta) = f_n(x_1, x_2, x_3, \ldots, x_n \mid \theta)\xi(\theta)$$

where $\boldsymbol{\xi(\theta)}$ is the probability distribution of $\boldsymbol{\theta}$. In the Bayesian framework $\boldsymbol{\xi(\theta)}$ is known as the ***prior distribution of θ***.

The ***posterior distribution*****,** the probability distribution of $\boldsymbol{\theta}$ given the vector $\underline{\mathbf{x}}$, $\boldsymbol{\xi(\theta|\underline{x})}$, is the ratio of the joint distribution of $\underline{\mathbf{x}}$ and $\boldsymbol{\theta}$, $\mathbf{h(\underline{x}, \theta)}$, and the marginal distribution of the sample. The marginal distribution of the sample is:

$$g_n(x_1, x_2, x_3, \ldots, x_n) = \int_\theta h(x_1, x_2, x_3, \ldots, x_n, \theta)d\theta = \int_\theta f_n(x_1, x_2, x_3, \ldots, x_n|\theta)\xi(\theta)d\theta$$

Applying the standard formula for conditional probability:

$$\xi(\theta| x_1, x_2, x_3, \ldots, x_n) = \frac{f_n(x_1, x_2, x_3, \ldots, x_n \mid \theta)\xi(\theta)}{g_n(x_1, x_2, x_3, \ldots, x_n)}$$

which is in the same form as Bayes Theorem. Stated more compactly:

$$\xi(\theta|\underline{x}) = [f_n(\underline{x}|\theta)\xi(\theta)]/g_n(\underline{x})$$

C. Most of the time it is not practical to compute $g_n(\underline{x})$. However, this is not a problem because *with respect to the posterior distribution it is a constant*. That is:

$$\xi(\theta|\underline{x}) \propto f_n(\underline{x}|\theta)\xi(\theta)$$

B. The Confusion Over Likelihood Functions!

Consider the standard development of MLE in most textbooks. The Likelihood setup for the Bernoulli distribution is:

$$L(p \mid \underline{x}) = \prod_{i=1}^{n} f(x_1 \mid p)f(x_2 \mid p)\ldots f(x_n \mid p) = p^{\sum_{i=1}^{n} X_i}(1 - p)^{n-\sum_{i=1}^{n} X_i}$$

this is usually written as $L(p \mid \underline{x}) = p^y(1 - p)^{n-y}$ where $y = \sum_{i=1}^{n} X_i$. Standard Calculus can now be used to solve for the value of p that maximizes $\mathbf{L(p|\underline{x})}$; namely,

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n .$$

1. Note that $\mathbf{L(p|\underline{x})}$ ***is not a probability distribution because p is NOT a random variable!*** It is simply a ***function***. In this example, *even if* **p** is treated as a random variable $\mathbf{L(p|\underline{x})}$ is not a proper probability distribution. Although $\mathbf{L(p|\underline{x})} \geq 0$ it does not integrate to 1:

$$\int_0^1 L(p \mid \underline{x})dp = \int_0^1 p^y(1-p)^{n-y}dp = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

2. **The joint distribution of the sample** is:

$$f_n(\underline{x} \mid p) = \prod_{i=1}^{n} f(x_1 \mid p) f(x_2 \mid p) \ldots f(x_n \mid p) = p^{\sum_{i=1}^{n} X_i} (1 - p)^{n - \sum_{i=1}^{n} X_i}$$

Technically:

$$f_n(\underline{x} \mid p) = \begin{cases} p^{\sum_{i=1}^{n} X_i} (1 - p)^{n - \sum_{i=1}^{n} X_i} & \begin{matrix} x_1 = 0,1 \\ x_2 = 0,1 \\ \ldots \\ x_n = 0,1 \end{matrix} \\ 0 \text{ otherwise} & \end{cases}$$

This is a probability distribution because $\mathbf{f_n(\underline{x}|p)} \geq 0$ and

$$\sum_{X_1=0}^{1} \sum_{X_2=0}^{1} \sum_{X_3=0}^{1} \ldots \sum_{X_n=0}^{1} p^{\sum_{i=1}^{n} X_i} (1 - p)^{n - \sum_{i=1}^{n} X_i} = 1$$

3. Confusion reigns because $\mathbf{f_n(\underline{x}|p)}$ and $\mathbf{L(p|\underline{x})}$ ***are the same equation for a random sample from a known parametric distribution***. What shifts is what variables are taken as ***random*** and the Likelihood function is ***not*** a probability distribution it is a ***function***.

4. Gary King resolves this problem by simply stating as an axiom what is true in practice that:

$$L(\theta \mid \underline{x}) \propto f_n(\underline{x} \mid \theta)$$

so that

$$\boldsymbol{\xi(\theta|\underline{x}) \propto f_n(\underline{x}|\theta)\xi(\theta) \propto L(p|\underline{x})\xi(\theta)}$$