ay something about subjective probability being the key tenet of the Bayesian approach. Then:

Following the subjective interpretation, our probability assessments about X are made conditional on our background knowledge $\mathcal{H}$. Usually, $\mathcal{H}$ is large, very complex, of high dimension and may be mostly irrelevant to X. What we need is some way of abridging $\mathcal{H}$ so that it is more manageable. This introduces the idea of a parameter and a parametric model. We assume that there is another random quantity θ that summarises the information in $\mathcal{H}$ about X, and hence makes X and $\mathcal{H}$ independent. By the partition law:

$$P(X \mid \mathcal{H}) \;=\; \sum_{\theta} P(X \mid \theta, \mathcal{H})\, P(\theta \mid \mathcal{H}) \;=\; \sum_{\theta} P(X \mid \theta)\, P(\theta \mid \mathcal{H}). \tag{1}$$

Now our probability distribution for X is a function of two probability distributions: the first, $P(X \mid \theta)$, is the *probability model* for X with *parameter* θ. The second, $P(\theta \mid \mathcal{H})$, is called the *prior distribution* of θ.

For a set of random quantities X_1, X_2, $\ldots, X_n$ one can still use the model and prior approach by writing:

$$P(X_1, X_2, \ldots, X_n \mid \mathcal{H}) \;=\; \sum_{\theta} P(X_1, X_2, \ldots, X_n \mid \theta)\, P(\theta \mid \mathcal{H}). \tag{2}$$

In many situations, where the X_i are a random sample of a quantity, it makes sense to assume that each X_i is independent of the others conditional on θ, thus:

$$P(X_1, X_2, \ldots, X_n \mid \mathcal{H}) \;=\; \sum_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta)\, P(\theta \mid \mathcal{H}). \tag{3}$$

Equation 3 is fundamental to statistical learning about unknown quantities from data. Under the assumptions of this equation, there are two quantities that we can learn about:

1. Our beliefs about likely values of the parameter θ given the data;

2. Our beliefs about the X's given observation of X_1, $\ldots$, X_n. In particular, we might want to assess the probable values of the next observation X_{n+1} in light of the data.

For Bayesian learning, in the spirit of the belief that probability is the only way to describe uncertainty, Bayesian inference strives to produce a probability distribution for the unknown quantities of interest. For inference on the parameter θ, the *posterior* distribution given the data, $P(\theta \mid X_1, \ldots, X_n, \mathcal{H})$, is the natural expression to look at; by Bayes' formula it can be written in terms of the model and prior:

$$P(\theta \mid X_1, \ldots, X_n, \mathcal{H}) \;=\; \frac{\prod_{i=1}^{n} P(X_i \mid \theta) P(\theta \mid \mathcal{H})}{\sum_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta) P(\theta \mid \mathcal{H})}. \tag{4}$$

Since the X_i are known, $\prod_i P(X_i \mid \theta)$ is known as the likelihood. As in Equation **??**, we can then write

$$P(\theta \mid X_1, \ldots, X_n, \mathcal{H}) \;\propto\; \prod_{i=1}^{n} P(X_i \mid \theta) P(\theta \mid \mathcal{H}) \tag{5}$$

or, in words,

$$\text{posterior} \;\propto\; \text{prior} \;\times\; \text{likelihood}. \tag{6}$$

For our belief about the next observation, we calculate the distribution of X_{n+1} conditional on the observations and $\mathcal{H}$; this is given by Equation 1, but with the posterior distribution of θ replacing the prior:

$$P(X_{n+1} \,|\, X_1, ..., X_n, \mathcal{H}) \;=\; \sum_{\theta} P(X_{n+1} \,|\, \theta)\, P(\theta \,|\, X_1, \ldots, X_n, \mathcal{H}). \tag{7}$$

This is called the posterior *predictive* distribution of X.

This is what happens when you try to cite something....[1]

References

[1] M. A. Tanner. *Tools for statistical inference.* Springer, New York, Third edition, 1997.

[2] S. M. Ross. *Simulation.* Academic Press, London, Third edition, 2002.