The formalism of the Bayesian methodology is not too hard to explain, but the philosophical points of view are very contentious.

The discussion starts with Bayes’ theorem, a familiar probability result. The essential manipulation is this:

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)} \]

This is a non-controversial formula, but you should be aware of its potential for time reversal. Suppose that event A comes earlier in time. Then \( P(B \mid A) \) asks about the probability of a current event, given a past event. Time is reversed in \( P(A \mid B) \), which asks about a past event, given a current event.

You often see Bayes’ formula explained in terms of a partition. Let \( A_1, A_2, \ldots, A_n \) be a partition. “Partition” means that

\[ P(A_i \cap A_j) = 0 \text{ for } i \neq j \]

and also that

\[ P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 \]

Bayes’ formula might now be written

\[ P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)} \]

Now imagine that we have a standard parametric problem dealing with a likelihood \( f(x \mid \theta) \). Here \( x \) is a stand-in for the data that we’ll see, while \( \theta \) represents the parameter which will forever remain unknown. The conventional approach to this problem (called the likelihood approach) will use the likelihood as the primary object of interest. Estimates of \( \theta \) will be based on the likelihood, and the method of maximum likelihood has paramount importance. The parameter \( \theta \) is considered nonrandom.

The Bayesian, however, believes that previous knowledge about \( \theta \) (or even prejudices about \( \theta \)) can be incorporated into a prior distribution. Imagine that \( \pi(\theta) \) denotes this prior distribution. Now think of \( \theta \) as random with probability law \( \pi(\theta) \).

One can now engage in a petty argument which distinguishes (i) situations in which there is a genuine random process (amenable to modeling) which creates \( \theta \),
from (ii) situations in which \( \theta \) is created exactly once (so that modeling is impossible and irrelevant). The Bayesian has no need to make this distinction. It is the state of his mental opinion about \( \theta \) which is subjected to modeling.

It follows that the joint density of the random \((\theta, X)\) is given by \( \pi(\theta) f(x \mid \theta) \). Integration of \( \theta \) will give the marginal law of \( X \). Let’s denote this as \( m(x) \).

\[
m(x) = \int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta
\]

We can now create the conditional law of \( \theta \), given the data \( x \). Specifically, this is

\[
f(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{\int_{\Theta} \pi(\theta') f(x \mid \theta') d\theta'} = \frac{\pi(\theta) f(x \mid \theta)}{m(x)}
\]

In the denominator, \( \theta' \) rather than \( \theta \) has been used as the dummy of integration, just to avoid confusion. More simply, we could note that

\[
f(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{(\text{factor without } \theta)}
\]

This means that the denominator is just a fudge factor involving \( x \) and some numbers (but not \( \theta \)), so that we can supply the denominator in whatever way is needed to make \( f(\theta \mid x) \) a legitimate density. (A later example will make this point clear.)

The Bayesian calls \( f(\theta \mid x) \) as the posterior density. If another experiment is to be done, then the posterior becomes the prior for this next experiment.

The Bayesian regards \( f(\theta \mid x) \) as the best summary of the experiment.

If you wanted an estimate of \( \theta \), the Bayesian would supply you with a measure of location from \( f(\theta \mid x) \), possibly the mean or median. If you have gone through the formality of creating a loss function, the Bayesian would minimize the expected posterior loss.

If you wanted a 95% confidence interval for \( \theta \), the Bayesian would give you an interval \((a, b)\) with the property that

\[
\int_{a}^{b} f(\theta \mid x) d\theta = 0.95
\]

presumably choosing this interval so that \( b - a \) is as short as possible.
Consider now this simple example. Suppose that $\theta$ is a binomial parameter. Suppose that you want to estimate $\theta$ based on a random variable $Y$, which is binomial $(n, \theta)$.

The maximum likelihood person uses $\hat{\theta} = \frac{Y}{n}$ with no further mental anguish.

The Bayesian will invoke a prior distribution $\pi(\theta)$ for $\theta$. For this problem, this will likely be an instance of the beta distribution. Specifically, he might choose

$$
\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} 1(0 < \theta < 1)
$$

This is the beta distribution with parameters $(\alpha, \beta)$.

For this problem $f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$. Then the joint density of $(\theta, Y)$ is

$$
\pi(\theta)f(y|\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \left( \binom{n}{y} \theta^y (1-\theta)^{n-y} \right)
$$

and the posterior is

$$
f(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{m(y)} I(0 < \theta < 1)
$$

Observe that this has the form

$$
f(\theta|y) = \text{(expression without } \theta) \theta^{y+\alpha-1} (1-\theta)^{(n-y)+\beta-1} I(0 < \theta < 1)
$$

This reveals that $f(\theta|y)$ must be the beta distribution with parameters $(\alpha+y, \beta+n-y)$.

Suppose that you wanted to estimate $\theta$ with squared-error loss. That is, $L[\theta, w(y)] = (\theta - w(y))^2$. The Bayesian minimizes this expected loss with regard to the posterior distribution. The task is to select $w(y)$ to minimize

$$
\int_{\Theta} (\theta - w(y))^2 f(\theta|y) d\theta
$$
In our example, Θ corresponds to the interval (0, 1).

This minimizing can be accomplished by selecting the mean of the posterior distribution.

ASIDE: For any density \( f(z) \), the integral \( \int_{-\infty}^{\infty} (z-t)^2 f(z) \, dz \) is minimized by selecting \( t = \int_{-\infty}^{\infty} z f(z) \, dz \), which is the mean of the distribution. You can show this by an easy differentiation.

For the Bayesian, the mean of the posterior is his estimate (when using squared-error loss), and here this is \( \hat{\theta} = \frac{y+\alpha}{n+\alpha+\beta} \). This estimate has many interesting properties.

The prior distribution had mean \( \frac{\alpha}{\alpha+\beta} \) and variance \( \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \). If you think of \( \omega = \frac{\alpha}{\alpha+\beta} \) and \( N = \alpha+\beta \), the variance is about \( \frac{\omega(1-\omega)}{N} \).

The posterior has mean \( \frac{y+\alpha}{n+\alpha+\beta} \). The prior carries an interpretation of an experiment of \( \alpha+\beta \) trials with \( \alpha \) successes. (Of course, \( \alpha \) and \( \beta \) need not be integers. Also, not every prior distribution admits such a facile interpretation.)

The Bayesian estimate is a weighted average of the prior mean \( \frac{\alpha}{\alpha+\beta} \) and the sample average \( \frac{y}{n} \). In particular,

\[
\frac{y+\alpha}{n+\alpha+\beta} = \frac{\alpha+\beta}{\alpha+\beta+N} \times \frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n} \times \frac{y}{n}
\]

If \( n \) is large compared to \( \alpha+\beta \), then the prior distribution matters very little. The Bayesian estimate will not differ materially from the traditional \( \frac{y}{n} \).

Here’s an example that makes the Bayesian look very clever. Suppose that you take a quarter dollar and flip it ten times. Suppose that you get four heads. The likelihood-based estimate of the probability of heads is \( \frac{4}{10} = 0.4 \). The Bayesian regards this as
nonsense. In fact, the Bayesian might have a beta prior distribution with $\alpha = \beta = 200$. The Bayesian estimate (using squared-error loss and hence the mean of the posterior distribution) is 

$$\frac{4 + 200}{10 + 200 + 200} = \frac{204}{410} \approx 0.4976.$$

Another potent argument in favor of the Bayesian methodology is the ability to make organized updates of new information. The Bayesian will use today’s posterior as tomorrow’s prior. This gives him a rational way to combine evidence.

The Bayesian has additional ammunition in that one can show that a likelihood-based method is within $\varepsilon$ (in some technical sense) of a Bayesian method with some particular prior. That is, the Bayesian claims that a likelihood-based method is very close to using a particular prior by default. The practitioner is better off selecting a prior that really reflects his feelings.

The Bayesian also avoids a strange ambivalence of the likelihood method. Traditionally, we regard the unseen and unobservable $\theta$ as a fixed quantity and regard the available numeric data $x$ as random. This is clearly silly. The Bayesian posterior $f(\theta|x)$ clearly thinks of $x$ as given and $\theta$ (the unknown) as random.
A Summary of the Bayesian Method and Bayesian Point of View

In spite of these arguments, the writer of this document, Gary Simon, is a NON-Bayesian. The major non-Bayesian arguments are these:

Prior distributions will differ from one person to another. As a result, the output of an analysis depends on who is doing it. This vagueness cannot be tolerated in the scientific world.

Prior distributions are difficult to formulate. The Bayesians have made considerable progress at eliciting priors in one-parameter or simple multi-parameter cases. Most real work involves complicated situations (such as a regression with 15 parameters or a survival analysis with 80 parameters) in which it is hard to elicit a meaningful prior. Worse yet, the priors that are naturally suggested tend to be overly simplistic and are deficient in interaction structure.

To cover the case of complete ignorance about a parameter, the Bayesian will sometimes suggest a “diffuse” prior such as the uniform distribution on \((0, \infty)\). Such improper priors represent a philosophical cop-out.

I have not lately been asked for a serious analysis of ten flips of a coin.

The world wants modular analyses. This is just the way that all human activity seems to be conducted. The Bayesian is trying to build a coherent structure for all decision making, and the world just doesn’t want it.

A useful reference on this controversy is Bradley Efron’s “Why Isn’t Everyone a Bayesian?” This appeared in *The American Statistician*, 1986.

There is a series of articles on the teaching of Bayesian methods in elementary courses in *The American Statistician*, August 1997. Perhaps the most important of these articles is the piece by David Moore in which he surveys wide areas of statistical applications and finds Bayesian methods nearly completely absent!

There may be a future for Bayesian methods, since they furnish workable solutions for a number of otherwise intractable problems. The notion of a prior distribution provides a useful averaging strategy in some cases in which the likelihood method is ambiguous.

A long-standing criticism of Bayesian methods centered around their obsession with simple tractable forms for \(\pi(\theta)\) and \(f(x | \theta)\). In fact, there was a need to have convenient “conjugate” priors \(\pi(\theta)\) to make the problem work. This difficulty has largely been removed by the MCMC (Monte Carlo Markov Chain) idea, a very clever computational advancement. Helping the computation however does not help the philosophy!