Basic Elements of Bayesian Analysis

The basic elements in a "full" Bayesian analysis are:

1. The parameter of interest, say θ. Note that this is completely general, since θ may be vector valued. So θ might be a binomial parameter, or the mean and variance from a Normal distribution, or an odds ratio, or a set of regression coefficients, etc. The parameter of interest is sometimes usefully thought of as the "true state of nature".

2. The prior distribution of θ, $f(\theta)$. This prior distrubution summarizes what is known about θ before the experiment is carried out. It is "subjective", so may vary from investigator to investigator.

3. The likelihood function, $f(x|\theta)$. The likelihood function provides the distribution of the data, x, given the parameter value θ. So it may be the binomial likelihood, a normal likelihood, a likelihood from a regression equation with associated normal residual variance, etc.

4. The posterior distribution, $f(\theta|x)$. The posterior distribution summarizes the information in the data, x, together with the information in the prior distribution, $f(\theta)$. Thus, it summarizes what is known about the parameter of interest θ after the data are collected.

5. Bayes Theorem. This theorem relates the above quantities:

$$\text{posterior distribution} = \frac{\text{likelihood of the data} \times \text{prior distribution}}{\text{a normalizing constant}},$$

   or

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{\int f(x|\theta) \times f(\theta) d\theta,}$$

   or, forgetting about the normalizing constant,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta).$$

   Thus we "update" the prior distribution to a posterior distribution after seeing the data via Bayes Theorem.

6. The action, a. The action is the decision or action that is taken after the analysis is completed. For example, one may decide to treat a patient with Drug 1 or Drug 2, depending on the data collected in a clinical trial. Thus our action will either be to use Drug 1 (so that $a = 1$) or Drug 2 (so that $a = 2$).

7. The loss function, $L(\theta, a)$. Each time we choose an action, there is some loss we incur, which depends on what the true state of nature is, and what action we decide to take. For example, if the true state of nature is that Drug 1 is in fact superior to Drug 2, then choosing action $a = 1$ will incur a smaller loss than choosing $a = 2$. Now, the usual problem is that we do not know the true state of nature, we only have data that lets us make probabilistic statements about it (ie, we have a posterior distribution for θ, but do not usually know the exact value of θ). Also, we rarely make decisions before seeing the data, so that in general, $a = a(x)$ is a function of the data. Note that while we will refer to these as "losses", we could equally well use "gains".

8. Expected Bayes Loss (Bayes Risk): We do not know the true value of θ, but we do have a posterior distribution once the data are known, $f(\theta|x)$. Hence, to make a "coherent" Bayesian decision, we minimize the Expected Bayesian Loss, defined by:

$$\text{EBL} = \int L(\theta, a(x)) f(\theta|x) d\theta$$

   In other words, we choose the action $a(x)$ such that the EBL is minimized.

The first five elements in the above list comprise a non-decision theoretic Bayesian approach to statistical inference. This type of analysis (ie, non-decision theoretic) is what most of us are used to seeing in the medical literature. However, many Bayesians argue that the main reason we carry out any statistical analyses is to help in making decisions, so that elements 6, 7, and 8 are crucial. There is little doubt that we will see more such analyses in the near future, but it remains to be seen how popular the decision theoretic framework will become in medicine. The main problem is to specify the loss functions, since there are so many possible consequences (main outcomes, side-effects, costs, etc.) to medical decisions, and it is difficult to combine these into a single loss function. My guess is that much work will have to be done on developing loss functions before the decision theoretic approach becomes mainstream. This course, therefore, will focus on elements 1 through 5. Nevertheless, for completeness, today we will look at an example using loss functions in some detail.

An Example Using A Loss Function

Loss functions can take on many different forms, but there are at least two basic types:

Mathematical loss functions: Rather generic, used for selecting the best estimator in certain "classes" of problems. Used for inferential purposes, not necessarily tied to detailed substantive decision problems. Typical example is "squared error loss", where the further an estimate is likely to be from its true value, the larger the loss will be.

Substantive loss functions: Non-generic and non-generalizable, tied to a specific real decision problem that needs to be solved.

We will now see an example of the first type. The second type are difficult to realistically implement in medicine, and more work is required before the application becomes routine.

Squared Error Loss - A generic mathematical loss function

Suppose we want to estimate a normal mean with known variance (arbitrary, so say $\sigma = \sigma^2 = 1$). Suppose we have data of sample size $n = 10$

$$x_1, x_2, \ldots, x_{10} \sim N(\mu, \sigma^2 = 1)$$

Suppose we take a normal prior on μ, $N(0, \sigma^2 = 100)$. Let our estimator be denoted by a (so that our action is the estimate we will choose for μ). Also suppose that our loss function is "squared error loss", that is,

$$L(\mu, a) = (a - \mu)^2$$

So, the further our estimate is from the true (but unknown) μ, the larger our loss, and we sqaure the difference to construct a loss function L.

To summarize, let's see exactly what our 8 steps are in this scenario:

1. The parameter of interest is μ, the normal mean.

2. The prior distribution is $N(\theta = 0, \tau^2 = 100)$. Note that θ is the prior mean, and τ is the prior SD so that τ^2 is the prior variance.

3. The likelihood function is, as usual for normally distributed data when $\sigma = 1$,

$$f(x_1, \ldots, x_n|\mu, \sigma^2 = 1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$$

4. Suppose that the mean of the observed data is 2, i.e., $\overline{x} = 2$. The posterior distribution (as seen in EPIB-607 or EPIB-613, or, if not, in a few classes) is again normally distributed

$$\mu \sim N\left(A \times \theta + B \times \overline{x}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right)$$

where
$A = \frac{\sigma^2/n}{\tau^2+\sigma^2/n} \frac{1^2/10}{100+1/10} = 0.00099901$
$B = \frac{\tau^2}{\tau^2+\sigma^2/n}$=0.999001
$n = 10$
$\sigma = 1$
$\tau = \sqrt{100} = 10$
$\theta = 0$, and
$\overline{x} = 2$

Hence the posterior distribution is $f(\mu|x) = \mu \sim N(1.998002, 0.0999001)$. As expected, the posterior density concentrates around the observed mean 2, since the prior distribution was very weak. There is only small "movement" back towards the prior mean of zero.

5. Bayes Theorem. This theorem relates the above quantities, which we have already seen.

6. The action, a, will be our choice of estimator of μ, our inferential target parameter.

7. The loss function is $L(\mu, a) = (a - \mu)^2$.

8. Expected Bayes Loss (Bayes Risk) is

$$\text{EBL } = \int L(\mu, a(x))f(\mu|x)d\theta$$

In other words, we average the loss over all possible values of μ, where the average is over the posterior density of μ. This average is minimized for some choice of a (i.e., some estimate of μ, and the challenge is to find that estimate.

technical details are omitted (complex integration), but it can be shown that the estimate that minimizes the EBL over the posterior density is the posterior mean, 1.998002. Thus, this is our best estimate under squared error loss function. It should not be too surprising the the posterior mean is a good estimate, but maybe not obvious that it is the absolutely optimal estimate.