Simple Examples of Bayesian Statistics

The following examples are intended to give very simple illustrations of the use of Bayes Theorem to compute posterior distributions from prior distributions. They also introduce some important terminology:

- The *posterior mean* for θ is $E(\theta|D)$. It is often used as a point estimate for θ. Another natural point estimate is the *posterior median*. Also used is the *posterior mode*, sometimes referred to as the MAP (maximum a posteriori) estimate. This is less natural in the continuous case (e.g. it is not parameterization invariant). We will see later how these different estimates can be motivated as attempting to minimize certain loss functions.

- A $(1-\alpha)$ *Credible Interval* (CI) for the parameter θ is any interval $[a, b]$ such that $p(\theta \in [a, b]) = 1 - \alpha$, where here the probability is a statement about θ (i.e. θ is considered a random variable; a and b are not). A more general concept is a "credible set": a $1-\alpha$ credible set is a set A such that $p(\theta \in A) = 1 - \alpha$. Sometimes the term "credible region" is also used.

  There are often many ways to create a $1-\alpha$ CI for a given posterior distribution (indeed, for continuous parameters, usually an infinite number of ways). Two ways are particularly common. The first is to use the *symmetric CI*: that is, choose a and b so that $p(\theta < a) = p(\theta > b) = \alpha/2$. The second is to use a Highest Posterior Density (HPD) interval, which is any set of the form $\{c : p_{f|D}(c) > T\}$ for some threshold T.

- The *predictive distribution* for x_{n+1} (given data $x_1, \ldots, x_n$) means the conditional distribution $p(x_{n+1}|x_1, \ldots, x_n)$. When the x_is are iid given θ it is given by

$$p(x_{n+1}|x_1, \ldots, x_n) = \int p(x_{n+1}|\theta)p(\theta|x_1, \ldots, x_n)d\theta \tag{1}$$

- A particular parametric family $\mathcal{F}$ is said to be *conjugate* for a particular inference problem if, when the prior distribution is $\in \mathcal{F}$ then the posterior distribution is also $\in \mathcal{F}$. (See examples below.)

1 Estimate of a Frequency from Binomial Data

Suppose we sample n alleles from a population and observe n_A of type A. What is the frequency, f, of the A allele? Can you give a point estimate for f? (e.g. posterior mean).

An interval estimate? (e.g. 95% CI). What is the probability that the next allele sampled will be an A? (i.e. what is the predictive distribution of the next observation given the previous observations).

Assume a prior distribution for f that is Beta(α, β). Then by Bayes Theorem

$$\begin{aligned} p(f|D) &\propto p(D|f)p(f) && (2)\\ &\propto f^{n_A}(1-f)^{n-n_A} f^{\alpha-1}(1-f)^{\beta-1} && (3)\\ &\propto f^{n_A+\alpha-1}(1-f)^{n-n_A+\beta-1} && (4)\\ &\sim \text{Beta}(n_A+\alpha, n-n_A+\beta). && (5) \end{aligned}$$

Note: from this we see that the Beta distribution is the *conjugate prior* distribution for the frequency f in a binomial likelihood.

A natural point estimate is the *Posterior expectation*, $E(f|D) = (n_A+\alpha)/(n+\alpha+\beta)$.

A natural interval estimate is the *symmetric* $1-\alpha$ *CI*, which is $[A, B]$ where A, B are chosen such that $p(f < A) = \alpha/2$ and $p(f > B) = \alpha/2$, so that $p(f \in [A, B]) = 1-\alpha)$. Note that these probability statements are about f, not about A and B! Here a credible interval differs from a confidence interval in its interpretation.

A better (shorter), but harder-to-compute, CI is the Highest Posterior Density $1-\alpha$ credible interval (set), $\text{HPD}(T) = \{c : p_{f|D}(c) > T\}$, where T is chosen so that $p(f \in \text{HPD}(T)) = 1-\alpha$.

Now to compute the predictive distribution of the next sample. Given f, the probability that the next allele is an A is f. So

$$\begin{aligned} p(\text{Next is } A|D) &= \int p(\text{Next is A}|f, D)p(f|D)\, df && (6)\\ &= \int f p(f|D), && (7) \end{aligned}$$

which from the above is $(n_A+\alpha)/(n+\alpha+\beta)$.

2 Inference of normal mean

See also Robert pp161-162.

Suppose we have a single observation $X = x$. Model X as a $N(\theta, \phi)$ given the values of the parameters θ and ϕ. Suppose ϕ is known and take a prior for θ which is $N(\theta_0, \phi_0)$, for fixed θ_0 and ϕ_0.

The posterior for θ is

$$p(\theta \mid x) \propto \frac{1}{\sqrt{2\pi\phi_0}} e^{-\frac{1}{2\phi_0}(\theta-\theta_0)^2} \frac{1}{\sqrt{2\pi\phi}} e^{-\frac{1}{2\phi}(x-\theta)^2} \tag{8}$$

$$\propto \exp\left(-\frac{1}{2}\theta^2(\frac{1}{\phi_0}+\frac{1}{\phi}) + \theta\left(\frac{\theta_0}{\phi_0}+\frac{x}{\phi}\right)\right) \tag{9}$$

Now, write

$$\phi_1 = \frac{1}{\frac{1}{\phi_0}+\frac{1}{\phi}}, \tag{10}$$

and

$$\theta_1 = \phi_1\left(\frac{\theta_0}{\phi_0}+\frac{x}{\phi}\right), \tag{11}$$

so that

$$p(\theta|x) \propto \exp\left(-\frac{1}{2}\frac{\theta^2}{\phi_1} + \frac{\theta\theta_1}{\phi_1}\right) \tag{12}$$

$$\propto \exp\left(-\frac{1}{2\phi_1}(\theta-\theta_1)^2\right). \tag{13}$$

It follows that the posterior for θ is $N(\theta_1, \phi_1)$. Note: from this we see that the normal distribution is the conjugate prior distribution for the mean parameter in a normal likelihood with known variance.

Remarks:

1. The posterior mean for θ is:

$$\theta_1 = \phi_1\left(\frac{\theta_0}{\phi_0}+\frac{x}{\phi}\right), \tag{14}$$

$$= \theta_0\frac{\phi_0^{-1}}{\phi_0^{-1}+\phi^{-1}} + x\frac{\phi^{-1}}{\phi_0^{-1}+\phi^{-1}}, \tag{15}$$

   a weighted average of the prior mean and the observed value with the weights depending on the precision (inverse of the variance) of the prior and the model. For example, if the prior variance is large relative to the model variance the posterior mean will be close to the observed data x.

2. The posterior variance is half of the harmonic mean of the prior and the model variances. In particular, since

$$\frac{1}{\phi_0}+\frac{1}{\phi} > \frac{1}{\phi}, \frac{1}{\phi_0}, \ \phi_1 = \frac{1}{\frac{1}{\phi_0}+\frac{1}{\phi}} < \phi, \phi_0,$$

   so the posterior variance is smaller than the prior and model variance.

3. For credible intervals, given x,

$$P\left(-z_{\alpha/2} < \frac{\theta - \theta_1}{\sqrt{\phi_1}} < z_{\alpha/2}\right) = 1 - \alpha,$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)\times 100$th percentile of the N(0,1) distribution. In particular as the prior variance $\phi_0 \to \infty$, $\phi_1 \to \phi$, and $\theta_1 \to x$. Thus, as $\phi_0 \to \infty$,

$$P\left(x - z_{\alpha/2}\sqrt{\phi} < \theta < x + z_{\alpha/2}\sqrt{\phi} \mid x\right) \to 1 - \alpha.$$

Thus, $\left(x - z_{\alpha/2}\sqrt{\phi}, x + z_{\alpha/2}\sqrt{\phi}\right)$ is a $\times(1-\alpha)$ posterior credible interval for θ. This is the usual $(1-\alpha)$ confidence interval for θ in frequentist statistics. Note the different interpretations of

$$P\left(x - z_{\alpha/2}\sqrt{\phi} < \theta < x + z_{\alpha/2}\sqrt{\phi}\right).$$

In the frequency approach, the statement applies before x is observed, and the randomness relates to the distribution of x. In the Bayesian approach (with a limiting flat prior for θ) the analysis is conditional on the observed value of x and the randomness relates to the uncertainty over the value of θ.

3 Classifying a sample, continued: Propogating Uncertainty in estimated frequencies

We now extend the example from the first handout to make it a little more realistic, replacing the unrealistic assumption that we "know" the frequencies of A in both savannah and forest subspecies with a more realistic assumption. Specifically, we now assume that we have available random samples of n individuals from each subspecies, and observed n_s and n_f copies of A in the savannah and forest samples respectively. Based on these data the maximum likelihood estimates for f_S and f_F are $\hat{f_S} = n_s/n$ and $\hat{f_F} = n_f/n$. A naive approach would be to replace the frequencies in the previous example, wherever they occur, with these estimates. (This is sometimes called the "plug-in" approach to classification.) However, this solution is not always satisfactory because it ignores the fact that these are only *estimates* of the frequencies. If n is sufficiently large then this approach will work fine because the estimates will be very accurate. However, if f_F and f_S are near 0 (which can happen easily) then the size of n required for the plug-in approach to work may be larger than what is actually available.

To show how big a problem this can be, imagine that $n = 100$, and that we see no copies of A in the savannah sample, and just one in the forest sample ($n_s = 0, n_f = 1$). Now

the maximum likelihood estimates of f_S and f_F are 0 and 0.01 respectively, and plugging these into (2) from the first sheet gives $p(H_s|D) = 0$. That is, we would conclude that the tusk was *certainly*, with no room for doubt, from a forest elephant, and not a savannah one. Furthermore we would draw this conclusion whatever the prior probabilities π_S and π_F (as long as $\pi_F \neq 0$), so even if we were *a priori* very confident that the tusk arose from a savanna elephant we would reverse this confidence based on the genetic data. Such a conclusion goes against the common sense intuition that, in this case, the amount of information obtained about the tusk origin from the genetic data is rather limited.

Now let us see how we can take a Bayesian approach to this problem. Now the unknowns include not only S, but also f_S and f_F, and the data D include not only that the tusk has allele A (an event we now denote T_A), but also the observation of n_s and n_f copies of A in n savanna and forest elephants. Thus we write $D = (T_A, n_s, n_f)$.

Since f_S and f_F are unknown, we must specify a prior distribution for them. A convenient choice here is the Beta distribution (because of conjugacy). To follow this example all you need to know is that the Beta distribution with parameters α, β, written Beta(α, β), is a distribution on $[0, 1]$ with density $p(f) \propto f^\alpha - 1(1-f)^{\beta-1}$, and expectation $\alpha/(\alpha+\beta)$. It may help to note that $\alpha = \beta = 1$ is the uniform distribution on $[0, 1]$ (see Wikipedia for more on the beta distribution). We assume that S is *a priori* independent of f_S and f_F, and that f_S and f_F are, a priori, independent $\sim$ Beta(α, β).

Recall, from the previous notes on this question, we have

$$\frac{p(H_S|D)}{p(H_F|D)} = \frac{p(D|H_S)}{p(D|H_F)} \frac{p(H_S)}{p(H_F)}. \tag{16}$$

Or, in words,

$$\text{Posterior odds} = \text{Bayes Factor} \times \text{prior odds}. \tag{17}$$

To compute the posterior odds, for any given prior odds, we have to compute the Bayes Factor, and so focus on computing $p(D|H_S)$ and $p(D|H_F)$. We have

$$p(D|H_S) = p(T_A, n_s, n_f|H_S) \tag{18}$$

$$= p(T_A|n_s, n_f, H_S)p(n_s, n_f|H_S) \tag{19}$$

and similarly

$$p(D|H_F) = p(T_A|n_s, n_f, H_F)p(n_s, n_f|H_F). \tag{20}$$

Since S is assumed independent of f_S and f_F, $p(n_s, n_f|H_S) = p(n_s, n_f|H_F)$ and these terms will cancel in the Bayes Factor $p(D|H_S)/p(D|H_F)$. Therefore we are left to compute

$p(T_A|n_s, n_f, H_S)$ and $p(T_A|n_s, n_f, H_F)$. For the first of these we have

$$p(T_A|n_s, n_f, H_S) = \int p(T_A, f_S|n_s, n_f, H_S)df_S \tag{21}$$

$$= \int p(T_A|f_S, n_s, n_f, H_S)p(f_S|n_s, n_f, H_S)df_S \tag{22}$$

$$= \int f_S p(f_S|n_s, n_f, H_S). \tag{23}$$

Now the conditional distribution $p(f_S|n_s, n_f, H_S)$ comes from Bayes theorem:

$$p(f_S|n_s, n_f, H_S) \propto p(n_s, n_f, H_S|f_S)p(f_S) \tag{24}$$

$$\propto f_S^{n_s}(1 - f_S)^{n-n_s} f_S^{\alpha-1}(1 - f_S)^{\beta-1} \tag{25}$$

$$\propto f_S^{n_s+\alpha-1}(1 - f_S)^{n-n_s+\beta-1} \tag{26}$$

from which we conclude that $f_S|n_s, n_f, H_S \sim \text{Beta}(n_s + \alpha, n - n_s + \beta)$. Now, recognizing (23) as the expectation of this distribution, we get $p(T_A|n_s, n_f, H_S) = \frac{n_s+\alpha}{n+\alpha+\beta}$.

Similarly, we obtain $p(T_A|n_s, n_f, H_F) = \frac{n_f+\alpha}{n+\alpha+\beta}$.

Return now to our example, with $n = 100$, $n_s = 0$, and $n_f = 1$. If we use $\alpha = \beta = 0.5$ (Jeffrey's prior for this problem) then the BF is $0.5/1.5 = 1/3$, which accords with the intuition that this reflects only "modest" evidence for H_F.

Of course, the BF in this case is quite sensitive to the prior - that is, the choice for α and β. However, if α, β are modest (i.e. not too small) the BF will be modest (i.e. not very strongly in favor of H_F).