Bayesian Statistics

In contrast to “frequentist” or “classical” statistics, the Bayesian approach to statistics regards parameters as unknowns about which we will in general have partial information.

When we collect data, we learn more about the parameters and as a consequence our beliefs about their possible values change.

It is natural to summarize our uncertainty about unknown quantities via probability distributions. (There is an axiom-based approach which shows that the only rational way to summarize uncertainty, where rationality here means obeying certain minimalist assumptions, is via probability distributions.)

In this sense assignments of probability are just subjective beliefs.

In Bayesian statistics, the probability distribution, $\pi(\theta)$, which summarises our uncertainty about the parameter θ *before* we have observed the data is called the *prior distribution* of the parameter.

The probability distribution, $\pi(\theta|x)$, which summarises our uncertainty *after* we have observed the data x is called the *posterior distribution*.

The posterior distribution can be calculated from the prior distribution by Bayes theorem:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{P(x)} \tag{1}$$

The three components on the rhs are:

1. the prior distribution π;
2. the likelihood $f(x|\theta)$ (written here just as the probability of the data given the parameter);
3. the (unconditional) probability of the data, $P(x) = \int_\theta f(x|\theta)\pi(\theta)d\theta$.

Remarks

1. In effect, in a Bayesian approach anything which is unknown, (whether data before it is observed, or a parameter) is treated as a random variable. Probability distributions just summarize our uncertainty about these unknowns. When the available information changes through the collection of data, so too will our beliefs about the parameters.

To a classical statistician, a parameter is an unknown but fixed quantity. It doesn't make sense in the classical/frequentist framework to talk for example about the probability that the parameter is greater than zero.

2. The “answer” in a Bayesian statistical framework is the posterior distribution. This tells us what we believe about the parameters in the light of the data.

Sometimes it is helpful to summarize the posterior. For example, the mean of the posterior distribution is a natural Bayesian point estimator of the parameter.

A range A of possible parameter values with the property that the posterior probability that the parameter lies in A is α is called a 100α% credible region for the parameter. (Credible regions are the Bayesian equivalent of frequentist confidence intervals, but note that they actually mean what many people would like to think confidence intervals mean.)

3. A common criticism of Bayesian statistics is that it is based on subjective assumptions, and hence is inappropriate for doing science, since the scientific method is objective.

Aside from general (and interesting!) arguments that even science is socially constructed, this critique is naive. There are subjective assumptions, such as the choice of statistical model, in all statistical approaches.

One useful approach is to do the Bayesian analysis with a range of different prior distributions. If the answers to questions of interest do not depend greatly on the choice of prior, that is reassuring. On the other hand, if the conclusions do depend on prior assumptions, that is saying that there is limited information in the data, and it is often preferable to realize this explicitly!

Example

Suppose we have data X which is binomially distributed with parameters n and p, and that we adopt a Beta(a, b) prior distribution for p:

$$\pi(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1},$$

for $a, b > 0$.

Then the posterior distribution is

$$\begin{aligned}
\pi(p|x) &= \frac{\pi(p) f(x|p)}{P(x)} \\
&= C p^{a-1}(1-p)^{b-1} \binom{n}{x} p^x (1-p)^{n-x} \\
&= C' p^{a+x-1}(1-p)^{b+n-x-1}
\end{aligned}$$

from which it follows that $\pi(p|x)$ is a Beta$(a+x, b+n-x)$ distribution.

Recall that the prior mean for p is $a/(a+b)$, and the prior variance is $ab/[(a+b)^2(a+b+1)]$.

The mean of the posterior is

$$\frac{a+x}{a+b+n}.$$

Compare this with the frequentist m.l.e. $\widehat{p} = x/n$. For large n (lots of data) they will be similar. For small n note the influence of the prior.

The variance of the posterior decays like n^{-1} for large n.

A certain Hospital has $x =$zero deaths from 45 emergency cardiac operations. What is its cardiac failure probability p? If we have no knowledge of these events we might take $a = 1, b = 1$ for our prior. We estimate posterior mean $\widehat{p} = 1/47$. The posterior probability $p < b$ is 0.95 at $b = 0.063$.

To adopt a Bayesian statistical approach, one needs to be able to calculate the posterior distribution.

Fundamentally, this requires one to be able to compute expectations over the posterior: $E(g(\theta)) = \int_{\Theta} g(\theta)\pi(\theta|x)d\theta$.

In some cases, for particular combinations of prior distributions and statistical models (hence likelihoods) this is possible analytically.

In the preceding example the likelihood was binomial, and for a Beta prior distribution, the posterior distribution was also Beta, with the parameters changed in a simple way. In this case, where the posterior has the same functional form as the prior, the prior and the likelihood are said to be *conjugate*.

Example adapted from Bustamante *et al.* (2003)

`doi:10.1016/S0040-5809(02)00050-3`

The site-frequency sprectrum $Y = (4, 3, 1, 1, 1)$ for mutations at $n = 5$ segregating sites from a single genomic region is approximately Poisson with mean Λ.

$$(Y_i|\Lambda) \sim \text{Poisson}(\Lambda) \qquad i = 1, 2, ..., 5.$$

We would like to estimate Λ.

Experience with this sort of data shows that the Poisson rate Λ is itself drawn from a Gamma distribution with known parameters α, β. This is our prior.

$$\Lambda \sim \text{Gamma}(\alpha, \beta).$$

The posterior probability density at $\Lambda = \lambda$ is

$$\begin{aligned} \pi(\lambda|y) &\propto \left[\prod_{i=1}^{n} f(y_i|\lambda)\right] \pi(\lambda) \\ &\propto \left[\prod_i e^{-\lambda}\lambda^{y_i}\right] \lambda^{\alpha-1}\exp(-\beta\lambda) \\ &= \lambda^{\sum_i y_i+\alpha-1}\exp(-n\lambda-\beta\lambda) \end{aligned}$$

and so $(\Lambda|Y=y) \sim \text{Gamma}(n\bar{y}+\alpha, n+\beta)$.

For this region $\alpha = 2, \beta = 2$. The prior mean is one, and since

$$(\Lambda|Y=y) \sim \text{Gamma}(12, 7),$$

the posterior mean is 12/7. A posterior confidence interval $[a, b]$ satisfies $\Pr(a \leq \Lambda \leq b|y) = 0.95$. The Gamma cdf gives us these limits, at 95%,

$$0.89 \leq \Lambda \leq 2.8.$$

In general, and in particular in many applications in genetics, one simply can't evaluate expectations $E(g(\theta)) = \int_\Theta g(\theta)\pi(\theta|x)d\theta$ analytically. Part of the problem is the normalizing constant for $\pi(\theta|x)$ in (1) $P(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$. This is already a hard calculation, and it is just part of the expectation.

One work-around is through numerical integration methods. More recently, there has been a massive increase in the applicability of Bayesian methods through the use of so-called *Markov chain Monte Carlo* (MCMC) methods.

The remainder of this week is practicals. The first 3 days of next week build up MCMC.