Bayesian Statistics

Bayesian statistics is built upon a subjective interpretation of probability. What exactly is meant by "subjective" is a source of controversy. It can mean simply that probability statements are judgements made by the statistical practitioner. However others accuse Bayesian statistics of allowing the practitioner to impose his/her own biases.

Another way of saying this is that a Bayesian statistician uses the language of probability to reflect two different kinds of uncertainty about a problem:

- aleatory uncertainty: due to inherent randomness in a system or observations of the system; used in frequentist statistics too

- epistemic uncertainty: due to our own incomplete understanding of the system; the point of scientific inquiry is to reduce this

Bayesian statistics is built upon Bayes Theorem. If $x^n = (x_1, \ldots, x_n)$ represents the observed data, we have

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{f(x^n)}$$

To use Bayes Theorem for inference, we attach interpretations.

- $f(\theta)$: the prior density – reflects knowledge of θ before seeing the data
- $f(x^n|\theta)$: the likelihood – joint density of the data for particular θ
- $f(\theta|x^n)$: the posterior density – reflects knowledge of θ after seeing the data
- $f(x^n)$: the normalizing constant – the marginal distribution for the data; can be hard to calculate

First consider a special class of problems in which the calculations on the previous page can be done in closed form.

A conjugate prior distribution for θ is one for which $f(\theta)$ and $f(\theta|x^n)$ belong to the same parametric family. In these cases, the key to calculation is to identify the kernel of the density and see how it is changed by the likelihood.

The kernel of a density for θ is the part that depends on θ, ignoring any constant multiplicative terms.

Example: Suppose $\theta \sim N(m, v)$ where v is known. What is the kernel?

Examples using conjugate priors

1. Suppose $X_1, \ldots, X_n|\lambda \overset{iid}{\sim} Poisson(\lambda)$, and the prior is $\lambda \sim Gamma(a, b)$. Find the posterior distribution for λ.

2. Suppose $X_1, \ldots, X_n|\theta \overset{iid}{\sim} N(\theta, \sigma^2)$, where σ^2 is known. Let the prior be $\theta \sim N(a, b^2)$. Find the posterior distribution for θ.

3. Suppose $X_1, \ldots, X_n|\sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-\alpha-1}\exp\{-\beta/\sigma^2\}$$

   Find the posterior distribution for σ^2.

In Bayesian statistics, all inference is based on the posterior distribution. We can use the posterior to calculate quantities similar to those under frequentist statistics (point estimates and intervals), or we can examine the posterior probability of *any* event of interest.

The posterior mean is a commonly used point estimator:

$$E[\theta|X_1,\ldots,X_n] = \int \theta f(\theta|X_1,\ldots,X_n)d\theta$$

It can often be written as a weighted average of the prior mean and the MLE. For example, in the second example on the previous page,

$$\begin{aligned} E[\theta|X_1,\ldots,X_n] &= \frac{b^2\sum_{i=1}^n X_i + a\sigma^2}{nb^2+\sigma^2} \\ &= \frac{nb^2}{nb^2+\sigma^2}\bar{X}_n + \frac{\sigma^2}{nb^2+\sigma^2}a \end{aligned}$$

A $1-\alpha$ credible interval for θ (also called a posterior interval) is an interval C_n satisfying

$$P(\theta \in C_n | X_1, \ldots, X_n) = 1 - \alpha$$

Note a few differences compared to a confidence interval:

- The probability statement is about θ, not C_n. C_n is a function of $X_1, \ldots, X_n$, which we are conditioning on in the probability statement.

- The statement is an equality. This is different from a frequentist interval, which puts a lower bound on the probability of coverage. Here we're not making a guarantee; we're just providing one summary of the posterior distribution.

- The intervals constructed this way may or may not have good frequentist coverage rates.

Note that C_n is not uniquely defined. There are several popular methods for finding such intervals.

A $1-\alpha$ equal-tail credible interval is an interval (a, b) such that

$$\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$$

A $1-\alpha$ highest posterior density (HPD) region R_n is defined such that

1. $P(\theta \in R_n|x^n) = 1-\alpha$

2. $R_n = \{\theta : f(\theta|x^n) > k\}$ for some k.

When $f(\theta|x^n)$ is unimodal, R_n is an interval.

Often it is more informative to plot $f(\theta|x^n)$ than it is to report an interval.

It is typically the case that the posterior distribution can't be calculated in closed form. This difficulty was a major roadblock for Bayesian statistics until the last 30 years or so, during which Monte Carlo sampling from the posterior became widespread.

We'll consider two basic methods for sampling from the posterior:

- Rejection sampling: produces an exact, iid sample, but may be difficult to tailor to a given problem

- Importance sampling: more generally applicable, but only approximates the distribution

Markov Chain Monte Carlo (MCMC) methods construct a Markov chain that has the posterior distribution as its stationary distribution. These are very flexible but are beyond the scope of this course.

Suppose we have $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta|x^n)$. The basic Monte Carlo approximation to the posterior mean of any function $q(\theta)$ is

$$\begin{aligned} E[q(\theta)|x^n] &= \int q(\theta) f(\theta|x^n) d\theta \\ &\approx \frac{1}{B} \sum_{i=1}^{B} q(\theta_i) \end{aligned}$$

This is broader than it might seem at first glance. For example, q could be an indicator function, giving us a way of approximating the posterior probability of any event.

We can also construct the ECDF to estimate the posterior CDF, or use a histogram or kernel density estimate (later in the course) to estimate the posterior PDF.

General algorithm for rejection sampling:

Suppose we can easily sample from some density $g(\theta)$, but what we want is a sample from $h(\theta)$, and we know $h(\theta)$ up to some proportionality constant. That is, suppose we know $k(\theta)$, where $h(\theta) = k(\theta)/\int k(\theta)d\theta$.

Moreover, suppose that we can find $M > 0$ such that

$$k(\theta) \leq Mg(\theta)\ \forall\theta \quad \text{(envelope condition)}$$

Then the following algorithm produces B iid draws from $h(\theta)$.

1. Draw $\theta^{cand} \sim g(\theta)$.
2. Generate $u \sim Unif(0, 1)$.
3. If $u \leq k(\theta^{cand})/Mg(\theta^{cand})$, accept θ^{cand}; otherwise reject it.

Repeat 1-3 until B values of θ^{cand} have been accepted.

Now consider rejection sampling where we first sample from the prior, with $g(\theta) = f(\theta)$, and the target is the posterior distribution, with $h(\theta) \propto k(\theta) = f(x^n|\theta)f(\theta)$. Note that by definition,

$$\frac{k(\theta)}{g(\theta)} = \frac{f(x^n|\theta)f(\theta)}{f(\theta)} = f(x^n|\theta) \leq f(x^n|\hat{\theta}_n) \equiv M$$

where $\hat{\theta}_n$ is the MLE. So the rejection sampling algorithm becomes

1. Draw $\theta^{cand} \sim f(\theta)$.
2. Generate $u \sim Unif(0, 1)$.
3. If $u \leq f(x^n|\theta^{cand})/f(x^n|\hat{\theta}_n)$, accept θ^{cand}; otherwise reject it.

Repeat 1-3 until B values of θ^{cand} have been accepted.

Importance sampling is an adaptation to the usual Monte Carlo integration that allows us to sample from an "importance function" g rather than the target density h. Note that

$$\begin{aligned} E_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)} \end{aligned}$$

where $\theta_1, \ldots, \theta_B \overset{iid}{\sim} g(\theta)$.

We can use this principle to obtain an approximation to $E[q(\theta)|x^n]$.

Sample from the prior: $\theta_1, \ldots, \theta_B \overset{iid}{\sim} f(\theta)$, then for each $i = 1, \ldots, B$, calculate

$$w_i = \frac{\mathcal{L}_n(\theta_i)}{\sum_{i=1}^{B} \mathcal{L}_n(\theta_i)}$$

Then $E[q(\theta)|x^n] \approx \sum_{i=1}^{B} q(\theta_i) w_i$.