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.

In most practical Bayesian analyses, the posterior distribution cannot be derived in closed form; it is often known only up to a normalization constant. This intractability posed a major challenge for Bayesian inference until roughly the past 30 years, during which Monte Carlo methods for sampling from the posterior became widespread. We will consider two fundamental Monte Carlo techniques:

- **Rejection sampling**, for drawing exact samples from the posterior distribution; and
- **Importance sampling**, for performing approximate posterior inference using samples drawn from a proposal distribution.

Markov Chain Monte Carlo (MCMC) methods are extremely flexible and powerful. They generate dependent samples by constructing a Markov chain whose stationary distribution is the target posterior; the details are beyond the scope of this course.

Posterior Inference

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.

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 $\theta_1, \ldots, \theta_B \overset{iid}{\sim} g(\theta)$. What if we know only the kernel of $h(\theta)$?

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

How should we choose a prior distribution? Several schools of thought answer this question differently:

- Subjective Bayesianism: The prior should reflect in as much detail as possible the researcher's prior knowledge of and uncertainties about the problem. These should be determined through *prior elicitiation*.

- Objective Bayesianism: The prior should incorporate as little subjective information as possible. Priors with this property are known as *non-informative*.

- Robust Bayesianism: Reasonable people may hold different priors, and it is difficult to precisely express even one person's prior; we should therefore consider the *sensitivity* of our inferences to changes in the prior.

A Bayesian analysis will often incorporate more than one of these ideas.

The simplest kind of non-informative prior places a uniform distribution on θ. When the range of θ is bounded, this prior gives a valid PDF, since it integrates to 1.

It is also possible to assign a uniform prior when the range of θ is not bounded. For example, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. We could take $f(\theta) \propto 1$. This prior is called "improper," since $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$.

However, we can still apply the Bayesian machinery to get

$$\begin{aligned} f(\theta|x^n) &\propto f(x^n|\theta)f(\theta) \\ &\propto \exp\left\{-\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n]\right\} \end{aligned}$$

which is the kernel of a $N(\bar{X}_n, 1/n)$ distribution for θ. Therefore we still have a "proper posterior."

In a Bayesian analysis, hypotheses, like parameters, can be described using probability distributions.

The simplest case is when the hypotheses describe regions into which θ can fall, and these all have positive prior probability. If $H_0 : \theta \in \Theta_0$, then

$$\begin{aligned}\text{Prior probability: } P(H_0) &= \int_{\Theta_0} f(\theta) d\theta \\ \text{Posterior probability: } P(H_0|x^n) &= \int_{\Theta_0} f(\theta|x^n) d\theta\end{aligned}$$

Suppose $H_0, \ldots, H_{K-1}$ are K hypotheses under consideration. (Typically $K = 2$, but in theory we can have more.) Suppose that under H_k, $\theta \sim f(\theta|H_k)$. θ may mean different things under the various hypotheses.

Note that

$$P(H_k|x^n) = \frac{f(x^n|H_k)P(H_k)}{\sum_{k=1}^{K} f(x^n|H_k)P(H_k)}$$

Therefore, the posterior odds of H_i relative to H_j equals

$$\frac{P(H_i|x^n)}{P(H_j|x^n)} = \frac{f(x^n|H_i)}{f(x^n|H_j)} \times \frac{P(H_i)}{P(H_j)}$$

The term $f(x^n|H_i)/f(x^n|H_j)$ is called the Bayes Factor for comparing H_i to H_j. I'll denote it BF_{ij}.

Computing the Bayes Factor

When H_i and H_j represent regions of the parameter space, it's easier to calculate the prior and posterior odds, and from this compute the Bayes Factor.

If $H_i : \theta = \theta_i$ and $H_j : \theta = \theta_j$, then the Bayes Factor is just the ratio of likelihoods under the two values.

More generally,

$$f(x^n|H_i) = \int_\Theta f(x^n|\theta, H_i) f(\theta|H_i) d\theta,$$

which is called the marginal likelihood. If $f(\theta|H_i)$ is conjugate, it can be calculated in closed form. Otherwise, we use sampling to approximate it. For example, we could use MC integration, sampling from $f(\theta|H_i)$.

Example: Albert Pujols (St. Louis Cardinals) was voted the "most feared hitter in baseball" in 2010. However, Ichiro Suzuki (Seattle Mariners) has a very similar batting average. Here are their career statistics from 2001 to 2010, when they both played major league baseball.

Pujols: 5146 at bats; 1717 hits
Suzuki: 6099 at bats; 2030 hits

If we consider that each player has a "true" batting average p, around which their actual batting average fluctuates, we might be interested in looking at evidence for/against the hypothesis $p_{Pujols} = p_{Suzuki}$.

Suppose $X|p_1 \sim Bin(n, p_1)$ and $Y|p_2 \sim (m, p_2)$. Under $H_0 : p_1 = p_2$, we assign prior distribution $p_1 \sim Unif(0, 1)$ (and $p_2 = p_1$) and under $H_1 : p_1 \neq p_2$, we assign independent priors $p_1 \sim Unif(0, 1)$ and $p_2 \sim Unif(0, 1)$.

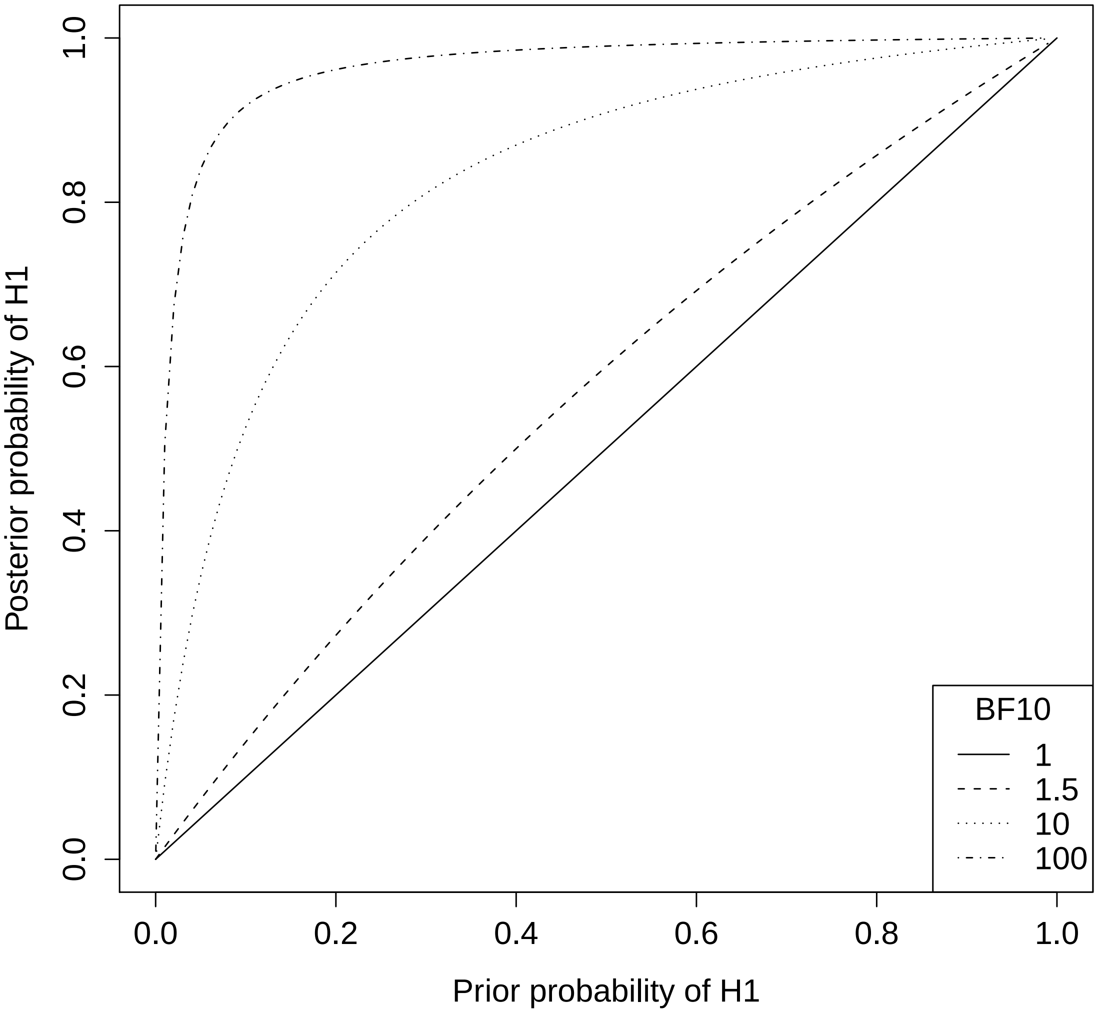
Calculate $f(x, y|H_1)$. (The rest of the problem is in the homework.)

Let BF_{10} be the Bayes Factor for comparing H_1 to H_0. We might classify BF_{10} as a measure of evidence against H_0 and in favor of H_1 as follows

$log_{10}(BF_{10})$	BF_{10}	Evidence
$0 - 1/2$	$1 - 3/2$	Weak
$1/2 - 1$	$3.2 - 10$	Moderate
$1 - 2$	$10 - 100$	Strong
> 2	> 100	Decisive

The key to this interpretation is to note that if $p = P(H_1)$ and $p^* = P(H_1|Data)$, then

$$p^* = \frac{\frac{p}{1-p} BF_{10}}{1 + \frac{p}{1-p} BF_{10}}$$


Posterior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
Prior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
BF10
1
1.5
10
100

More on noninformative priors

One property we might want a noninformative prior to possess is that it be **transformation invariant**. For example, if θ represents a distance, our inference shouldn't depend on whether θ is expressed in miles or kilometers.

That is, if instead of expressing the likelihood given θ, we express it given $\phi = g(\theta)$, we want a rule for choosing the priors f_θ and f_ϕ such that

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|$$

The Jeffreys prior does just this. For 1-dimensional θ, we have $f(\theta) \propto I(\theta)^{1/2}$.