Bayesian Statistics

James M. Flegal

Agenda

- Bayesian inference
- Priors
- Point estimates
- Bayesian hypothesis testing
- Bayes factors

Bayesian Inference

- Everything we have done up to now is frequentist statistics, Bayesian statistics is very different
  - Bayesians don't do confidence intervals and hypothesis tests
  - Bayesians don't use sampling distributions of estimators
  - Modern Bayesians aren't even interested in point estimators
- So what do they do? Bayesians treat parameters as random variables

To a Bayesian probability is the only way to describe uncertainty. Things not known for certain – like values of parameters – must be described by a probability distribution

Bayesian Inference

- Suppose you are uncertain about something, which is described by a probability distribution called your prior distribution
- Suppose you obtain some data relevant to that thing
- The data changes your uncertainty, which is then described by a new probability distribution called your posterior distribution
- Posterior distribution reflects the information both in the prior distribution and the data
- Most of Bayesian inference is about how to go from prior to posterior

Bayesian Inference

- Bayesians go from prior to posterior is to use the laws of conditional probability, sometimes called in this context Bayes rule or Bayes theorem
- Suppose we have a PDF g for the prior distribution of the parameter θ, and suppose we obtain data x whose conditional PDF given θ is f
- Then the joint distribution of data and parameters is conditional times marginal

$$f(x|\theta)g(\theta)$$

- May look strange because most of your training on considers the frequentist paradigm
- Here both x and θ are random variables

Bayesian Inference

- The correct posterior distribution, according to the Bayesian paradigm, is the conditional distribution of θ given x, which is joint divided by marginal

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta}$$

- Often we do not need to do the integral if we recognize that

$$\theta \mapsto f(x|\theta)g(\theta)$$

  is, except for constants, the PDF of a brand name distribution, then that distribution must be the posterior

Binomial Data, Beta Prior

Suppose the prior distribution for p is Beta(α_1, α_2) and the conditional distribution of x given p is Bin(n, p). Then

$$f(x|p) = \binom{n}{p} p^x (1-p)^{n-x}$$

and

$$g(p) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}.$$

Then

$$f(x|p)g(p) = \binom{n}{p} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{x+\alpha_1 - 1}(1-p)^{n-x+\alpha_2 - 1}$$

and this, considered as a function of p for fixed x is, except for constants, the PDF of a Beta$(x + \alpha_1, n - x + \alpha_2)$ distribution. So that is the posterior.

Binomial Data, Beta Prior

Why?

$$
\begin{aligned}
h(p|x) &= \frac{f(x|p)g(p)}{\int f(x|p)g(p)dp} \\
&\propto f(x|p)g(p) \\
&= \binom{n}{p} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1} \\
&\propto p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1}
\end{aligned}
$$

And there is only one PDF with support $[0, 1]$ of that form, i.e. a Beta$(x + \alpha_1, n - x + \alpha_2)$ distribution. So that is the posterior.

Bayesian Inference

- In Bayes rule, **constants**, meaning anything that doesn't depend on the parameter, are irrelevant
- We can drop multiplicative constants that do not depend on the parameter from $f(x|\theta)$ obtaining the likelihood $L(\theta)$
- We can also drop multiplicative constants that do not depend on the parameter from $g(\theta)$ obtaining the unnormalized prior
- Multiplying them together gives the unnormalized posterior

$$\text{likelihood} \times \text{unnormalized prior} = \text{unnormalized posterior}$$

Bayesian Inference

In our example we could have multiplied likelihood

$$p^x(1-p)^{n-x}$$

times unnormalized prior

$$p^{\alpha_1-1}(1-p)^{\alpha_2-1}$$

to get unnormalized posterior

$$p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1}$$

which, as before, can be recognized as an unnormalized beta PDF.

Bayesian Inference

- It is convenient to have a name for the parameters of the prior and posterior. If we call them parameters, then we get confused because they play a different role from the parameters of the distribution of the data.
- The parameters of the distribution of the data, p in our example, the Bayesian treats as random variables. They are the random variables whose distributions are the prior and posterior.
- The parameters of the prior, α_1 and α_2 in our example, the Bayesian treats as known constants. They determine the particular prior distribution used for a particular problem. To avoid confusion we call them **hyperparameters**.

Bayesian Inference

- Parameters, meaning the parameters of the distribution of the data and the variables of the prior and posterior, are unknown constants. The Bayesian treats them as random variables because probability theory is the correct description of uncertainty.
- Hyperparameters, meaning the parameters of the prior and posterior, are known constants. The Bayesian treats them as non- random variables because there is no uncertainty about their values.
- In our example, the hyperparameters of the prior are α_1 and α_2, and the hyperparameters of the posterior are $x + \alpha_1$ and $n - x + \alpha_2$.

Example: Normal

Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ where σ^2 is known. Suppose further we have a prior $\theta \sim N(\mu, \tau^2)$. Then the posterior can be obtained as follows,

$$
\begin{aligned}
f(\theta|x) &\propto f(\theta) \prod_{i=1}^{n} f(x_i|\theta) \\
&\propto \exp\left\{ -\frac{1}{2} \left(\frac{(\theta-\mu)^2}{\tau^2} + \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma^2} \right) \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \frac{\left(\theta - \frac{\mu/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2} \right)^2}{\frac{1}{1/\tau^2 + n/\sigma^2}} \right\}.
\end{aligned}
$$

Example: Normal

Or $f(\theta|x) \sim N(\mu_n, \tau_n^2)$ where

$$\mu_n = \left(\frac{\mu}{\tau^2} + \frac{n\bar{x}}{\sigma^2} \right) \tau_n^2 \quad \text{and} \quad \tau_n^2 = \frac{1}{1/\tau^2 + n/\sigma^2}.$$

We will call this a **conjugate** Bayes model. Also note a 95% credible region for θ is given by (this is also the HPD, highest posterior density)

$$(\mu_n - 1.96\tau_n, \mu_n + 1.96\tau_n).$$

For large n, the data will overwhelm the prior.

Example: Normal

- If $f(\theta) \propto 1$, an improper prior, then a 95% credible region for θ is the same as a 95% confidence interval since $f(\theta|x) \sim N(\bar{x}, \sigma^2/n)$ (try to show this at home).
- Usually, we specify a prior and likelihood that result in an posterior that is intractable. That is, we can't work with it analytically or even calculate the appropriate normalizing constant c.
- However, it is often easy to simulate a Markov chain with $f(\theta|x)$ as its stationary distribution.

Conjugate Priors

- Given a data distribution $f(\theta|x)$, a family of distributions is said to be conjugate to the given distribution if whenever the prior is in the conjugate family, so is the posterior, regardless of the observed value of the data
- Our first example showed that, if the data distribution is binomial, then the conjugate family of distributions is beta
- Our second example showed that, if the data distribution is normal with known variance, then the conjugate family of distributions is normal

Improper Priors

- A subjective Bayesian is a person who really buys the Bayesian philosophy. Probability is the only correct measure of uncertainty, and this means that people have probability distributions in their heads that describe any quantities they are uncertain about. In any situation one must make one's best effort to get the correct prior distribution out of the head of the relevant user and into Bayes rule.
- Many people, however, are happy to use the Bayesian paradigm while being much less fussy about priors. When the sample size is large, the likelihood outweighs the prior in determining the posterior. So, when the sample size is large, the prior is not crucial.

Improper Priors

- Such people are willing to use priors chosen for mathematical convenience rather than their accurate representation of uncertainty.
- They often use priors that are very spread out to represent extreme uncertainty. Such priors are called "vague" or "diffuse" even though these terms have no precise mathematical definition.
- In the limit as the priors are spread out more and more one gets so-called improper priors.

Improper Priors

- There is no guarantee that

$$\text{likelihood} \times \text{improper prior} = \text{unnormalized posterior}$$

  results in anything that can be normalized. If the right-hand side integrates, then we get a proper posterior after normalization. If the right-hand does not integrate, then we get complete nonsense.
- You have to be careful when using improper priors that the answer makes sense. Probability theory doesn't guarantee that, because improper priors are not probability distributions.

Improper Priors

- Improper priors are questionable
  - Subjective Bayesians think they are nonsense. They do not correctly describe the uncertainty of anyone.
  - Everyone has to be careful using them, because they don't always yield proper posteriors. Everyone agrees improper posteriors are nonsense.
  - Because the joint distribution of data and parameters is also improper, paradoxes arise. These can be puzzling.
- However they are widely used and need to be understood.

Objective Bayesian Inference

- The subjective, personalistic aspect of Bayesian inference bothers many people. Hence many attempts have been made to formulate **objective** priors, which are supposed to be priors that many people can agree on, at least in certain situations.
- However, none of the proposed **objective** priors achieve wide agreement.

Flat Priors

- One obvious **default** prior is flat (constant), which seems to give no preference to any parameter value over any other
- If the parameter space is unbounded, then the flat prior is improper
- One problem with flat priors is that they are only flat for one parameterization
- Another alternative is **Jeffreys priors**

Bayesian Point Estimates

- Bayesians have little interest in point estimates of parameters. To them a parameter is a random variable, and what is important is its distribution. A point estimate is a meager bit of information as compared, for example, to a plot of the posterior density.
- However, Bayesian point estimates are widely reported and something we will be estimating using MCMC
- Bayesian point estimates most commonly used are the posterior mean, the posterior median, the posterior mode, and the endpoints of Bayesian credible regions
- Frequentists too have little interest in point estimates except as tools for constructing tests and confidence intervals

Bayesian Credible Intervals

- Not surprisingly, when a Bayesian makes an interval estimate, it is based on the posterior.
- Many Bayesians do not like to call such things **confidence intervals** because that names a frequentist notion. Hence the name **credible intervals** which is clearly something else.
- One way to make credible intervals is to find the marginal posterior distribution for the parameter of interest and find its $\alpha/2$ and $1 - \alpha/2$ quantiles. The interval between them is a $100(1 - \alpha)\%$ Bayesian credible interval for the parameter of interest called the equal tailed interval.

Bayesian Point Estimates

- Suppose the data x is Bin(n, p) and we use the conjugate prior Beta(α_1, α_2), so the posterior is Beta($x + \alpha_1, n - x + \alpha_2$)
- Since we know the mean of a beta distribution, we can see the posterior mean is

$$E(p|x) = \frac{x + \alpha_1}{x + \alpha_1 + \alpha_2}$$

- The posterior median has no simple expression, but we can calculate it using the R

```
qbeta(0.5, x + alpha1, n - x + alpha2)
```

- The endpoints of Bayesian credible regions can also be found using R, say for an 80% credible region

```
qbeta(0.1, x + alpha1, n - x + alpha2)
qbeta(0.9, x + alpha1, n - x + alpha2)
```

Bayesian Point Estimates

- Suppose $\alpha_1 = \alpha_2 = 1/2$, $x = 2$, and $n = 10$.

```
alpha1 <- alpha2 <- 1 / 2
x <- 2
n <- 10
(x + alpha1) / (n + alpha1 + alpha2)
```

```
## [1] 0.2272727
```

```
qbeta(0.5, x + alpha1, n - x + alpha1)
```

```
## [1] 0.2103736
```

```
cbind(qbeta(0.1, x + alpha1, n - x + alpha2), qbeta(0.9, x + alpha1, n - x + alpha2))
```

```
##             [,1]      [,2]
## [1,] 0.08361516 0.3948296
```

Bayesian Hypothesis Tests

- Not surprisingly, when a Bayesian does a hypothesis test, it is based on the posterior.
- To a Bayesian, a hypothesis is an event, a subset of the sample space. Remember that after the data are seen, the Bayesian considers only the parameter random. So the parameter space and the sample space are the same thing to the Bayesian.
- The Bayesian compares hypotheses by comparing their posterior probabilities.
- All but the simplest such tests must be done by computer.

Bayesian Hypothesis Tests

- Suppose the data x is Bin(n, p) and we use the conjugate prior Beta(α_1, α_2), so the posterior is Beta($x + \alpha_1, n - x + \alpha_2$)
- Suppose the hypotheses in question are

$$H_0 : p \geq 1/2$$
$$H_1 : p < 1/2$$

- We can calculate the probabilities of these two hypotheses by the the R expressions

```
pbeta(0.5, x + alpha1, n - x + alpha2)
```

```
## [1] 0.9739634
```

```
pbeta(0.5, x + alpha1, n - x + alpha2, lower.tail = FALSE)
```

```
## [1] 0.02603661
```

Bayesian Hypothesis Tests

- Suppose $\alpha_1 = \alpha_2 = 1/2$, $x = 2$, and $n = 10$.

```
alpha1 <- alpha2 <- 1 / 2
x <- 2
n <- 10
pbeta(0.5, x + alpha1, n - x + alpha2)
```

```
## [1] 0.9739634
```

```
pbeta(0.5, x + alpha1, n - x + alpha2, lower.tail = FALSE)
```

```
## [1] 0.02603661
```

Bayesian Hypothesis Tests

- Bayes tests get weirder when the hypotheses have different dimensions
- In principle, there is no reason why a prior distribution has to be continuous
  - It can have degenerate parts that put probability on sets a continuous distribution would give probability zero
  - But many users find this weird
- Bayes Factors tend to be more widely used

Bayes Factors

- Let M be a finite or countable set of models. For each model $m \in M$ we have the prior probability of the model $h(m)$. It does not matter if this prior on models is unnormalized.
- Each model m has a parameter space Θ_m and a prior $g(\theta|m)$, $\theta \in \Theta_m$
- The spaces Θ_m can and usually do have different dimensions. That's the point. These within model priors must be normalized proper priors. The calculations to follow make no sense if these priors are unnormalized or improper.
- Each model m has a data distribution

$$f(x|\theta, m)$$

which may be a PDF or PMF.

Bayes Factors

The unnormalized posterior for everything, models and parameters within models, is

$$f(x|\theta, m)g(\theta|m)h(m)$$

To obtain the conditional distribution of x given m, we must integrate out the nuisance parameters θ

$$q(x|m) = \int_{\Theta_m} f(x|\theta, m)g(\theta|m)h(m)d\theta$$

$$h(m) \int_{\Theta_m} f(x|\theta, m)g(\theta|m)d\theta$$

These are the unnormalized posterior probabilities of the models.

The normalized probabilities are

$$p(m|x) = \frac{q(x|m)}{\sum q(x|m)}$$

Bayes Factors

It is useful to define

$$b(x|m) = \int_{\Theta_m} f(x|\theta, m) g(\theta|m) d\theta$$

so

$$q(x|m) = b(x|m)h(m)$$

Then the ratio of posterior probabilities of models m_1 and m_2 is

$$\frac{p(m_1|x)}{p(m_2|x)} = \frac{q(x|m_1)}{q(x|m_2)} = \frac{b(x|m_1)h(m_1)}{b(x|m_2)h(m_2)}$$

This ratio is called the **posterior odds** of the models (a ratio of probabilities is called an odds) of these models.

Bayes Factors

The **prior odds** is

$$\frac{h(m_1)}{h(m_2)}$$

The term we have not yet named in

$$\frac{p(m_1|x)}{p(m_2|x)} = \frac{b(x|m_1)h(m_1)}{b(x|m_2)h(m_2)}$$

is called the **Bayes factor**

$$\frac{b(x|m_1)}{b(x|m_2)}$$

the ratio of posterior odds to prior odds.

The prior odds tells how the prior compares the probability of the models. The Bayes factor tells us how the data shifts that comparison going from prior to posterior via Bayes rule.

Bayes Factors

- Suppose the data x is Bin(n, p) and the models (hypotheses) in question are

$$m_1 : p = 1/2$$
$$m_2 : p \neq 1/2$$

- The model m_1 is concentrated at one point $p = 1/2$, hence has no nuisance parameter. Hence $g(\theta|m_1) = 1$. Suppose we use the within model prior Beta(α_1, α_2) for model m_2.
- Then

$$b(x|m_1) = f(x|1/2) = \binom{n}{x}(1/2)^x(1-1/2)^{n-x} = \binom{n}{x}(1/2)^n$$

Bayes Factors

Then

$$
\begin{aligned}
b(x|m_2) &= \int_0^1 f(x|p)g(p|m_2)dp \\
&= \int_0^1 \binom{n}{x} \frac{1}{B(\alpha_1, \alpha_2)} p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1} dp \\
&= \binom{n}{x} \frac{B(x+\alpha_1, n-x+\alpha_2)}{B(\alpha_1, \alpha_2)}
\end{aligned}
$$

where

$$
B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}
$$

by properties of the Beta distribution

Bayes Factors

```
alpha1 <- alpha2 <- 1 / 2
x <- 2
n <- 10
p0 <- 1 / 2
b1 <- dbinom(x, n, p0)
b2 <- choose(n, x) * beta(x + alpha1, n - x + alpha2) / beta(alpha1, alpha2)
BayesFactor <- b1 / b2
BayesFactor
```

```
## [1] 0.5967366
```

```
pvalue <- 2 * pbinom(x, n, p0)
pvalue
```

```
## [1] 0.109375
```

Bayes Factors

- For comparison, we calculated not only the Bayes factor 0.597 but also the frequentist p-value 0.109
  - Bayes factors and p-values are sort of comparable, but are **not** identical
- In fact, it is a theorem that in situations like this the Bayes factor is always larger than the p-value, at least asymptotically
- This makes Bayesian tests more conservative, less likely to reject the null hypothesis, than frequentists
- Either the frequentists are too optimistic or the Bayesians are too conservative, or perhaps both

Summary

- Bayesian inference, priors, Bayesian point estimates, Bayesian hypothesis testing, and Bayes factors
- Many applications including pattern recognition, span detection, search for lost objects, ...
- Calculations are trivial in our examples so far, not usually the case