Bayesian Inference

Bayesian Inference

The goal is to learn the parameter θ given the observed data

we want posterior distribution of θ : $P(\theta \mid \text{data})$

We typically have:

$P(\text{data} \mid \theta)$ likelihood, our model of the data

$P(\theta)$ prior

$P(\text{data})$ marginal probability of data

In Bayesian Inference, we want to learn the distribution of the parameters θ of the model given the observed data. θ can be a scalar or a vector with many parameters values. We accomplish this by calculating the posterior distribution of θ, P(θ|data). After we decide on how to model the data, we write down the likelihood, P(data|θ). The prior P(θ) quantifies our knowledge about θ before observing any data. Using the Bayes rule, we can calculate the posterior distribution of the parameter θ using the likelihood, the prior and the marginal probability of the data P(data).

Exercise: "Invert" the Probability Statement

$P(\text{data} \mid \theta)$ (likelihood)

$P(\theta \mid \text{data})$ (posterior)

$$P(\theta \mid \text{data}) \quad \left[\begin{array}{l} P(\text{data} \mid \theta) \\ P(\theta) \\ P(\text{data}) \end{array} \right.$$

hint Bayes' Rule: $P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$

Now, you are going to "invert" the probability statement, in other words, we want to get P(θ|data) using P(data|θ), P(θ), and P(data).

Example: I get a test for a disease, say HIV and the result is positive. The data here is that my test is positive. The parameter is my actual disease status, which could be I have the disease or I am healthy. What would I be more interested in? P(positive| diseased) or P(diseased| positive)? In medicine, P(positive|diseased) is also called sensitivity of the test and P(diseased|positive) is the positive predictive value. In our context, P(positive|diseased) is the likelihood and P(diseased|

Bayesian Inference

$$P(\theta \mid \text{data}) \propto P(\text{data} \mid \theta) \cdot P(\theta)$$

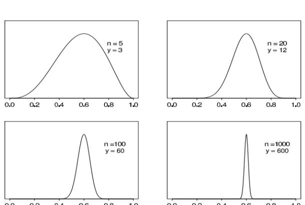


Figure 2.1 *Unnormalized posterior density for binomial parameter θ, based on uniform prior distribution and y successes out of n trials. Curves displayed for several values of n and y.*

Bayesian Data Analysis 3rd Ed - Gelman, Carlin, Stern, Dunson, Vehtari, Rubin

The expression at the top is the key relationship used for Bayesian inference. To get the posterior distribution, we use the product of the likelihood and the prior. The α symbol means that the LHS is proportional to the RHS.

Why do you think we drop the proportionality constant 1/P(data)?

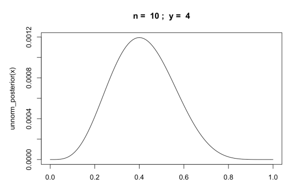
The marginal probability of the data P(data) can be computationally expensive to calculate. Since P(data) does not depend on θ there are several tricks to get the posterior without explicitly calculating it. Sometimes, when the RHS has a known distribution function, then the proportionality constant are already known or tabulated. In more complex cases, MCMC methods are used to obtain a sample of the posterior distribution.

When θ is a single parameter, plotting the RHS can give us a good sense of the posterior distribution. The figure shows the posterior density for the probability of success in n = 3, 20, 100, and 1000 trials, with 3, 12, 60, and 600 successes. As the number of trials goes up, more information about the parameter θ is collected, leading to the probability mass becoming more concentrated nearby 0.6, the true parameter.

Unnormalized Posterior with Simulated Binomial r.v.

A simulation of n=10 tosses with θ = 0.6 yielded 4 successes. Assuming uniform prior on θ, P(θ)=1, the unnormalized posterior is

$$P(\theta \,|\, \text{data}) \propto P(\text{data} \,|\, \theta) \cdot P(\theta) \propto \theta^{10-4}(1-\theta)^4$$



https://hakyimlab.github.io/hgen471/L1-binomial-parameter-posterior.html

Bayesian Data Analysis 3rd Ed - Gelman, Carlin, Stern, Dunson, Vehtari, Rubin

Here we simulate a binomial random variable as the number of success in n = 10 trials. In R this is accomplished with the command y = rbinom(1, 10, theta=0.6). In this particular simulation, we got 4 successes. The figure shows the unnormalized posterior (that means that all factors that do not depend on θ have been dropped) density of θ assuming a flat prior on θ.

What would be the mle estimate of theta? Does this posterior look consistent with the true parameter, which we know here because we simulated the data?

Try out other simulations to get a sense of the variability of the mle using the code here https://github.com/hakyimlab/hgen471/blob/master/analysis/L1-binomial-parameter-posterior.Rmd

---

Example: Proportion of Female Birth

241,945 girls and 251,527 boys were born in Paris from 1745 to 1770.

Calculate the posterior distribution of the proportion of female birth given the observed data $P(\theta \,|\, \text{data})$

we have 241,945 successes from 493,472 trials

write down the likelihood $P(\text{data} \,|\, \theta)$

choose the prior $P(\theta)$

try not to have to deal with the marginal prob of data $P(\text{data})$

Bayesian Data Analysis 3rd Ed - Gelman, Carlin, Stern, Dunson, Vehtari, Rubin

Laplace rediscovered Bayes rule to answer the question whether there were more boys than girls born during the second half of the 18th century. Let's calculate the posterior P(θ|data). Our data is that there were 241,945 girls out of 493,472 births. We can model the number of girls as the sum of 493,472 independent trials with 241,945 successes.

Which of the known distributions should we use?

To indicate that we don't have any prior knowledge of the probability of girl, let's use a uniform prior for θ.

Exercise: Proportion of Female Birth

we have 241,945 successes from 493,472 trials

write down the likelihood $P(\text{data} \mid \theta)$

choose the uniform prior $P(\theta)$

hint: number of girls born can be modeled as a binomial r.v.   hint: binomial pmf = $\binom{n}{y}\theta^y(1-\theta)^{n-y}$

calculate the posterior $P(\theta \mid \text{data})$

see posterior plotted in https://hakyimlab.github.io/hgen471/L1-female-birth-rate.html