02 Introduction to Bayes

Shravan Vasishth

SMLP 2019

Introduction to Bayesian data analysis

Recall Bayes' rule:

When A and B are observable events, we can state the rule as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

Note that $P(\cdot)$ is the probability of an event.

Introduction to Bayesian data analysis

When looking at probability distributions, we will encounter the rule in the following form.

$$f(\theta \mid \text{data}) = \frac{f(\text{data} \mid \theta) f(\theta)}{f(y)} \tag{2}$$

Here, $f(\cdot)$ is a probability density, not the probability of a single event. $f(y)$ is called a "normalizing constant", which makes the left-hand side a probability distribution.

$$f(y) = \int f(x, \theta)\, d\theta = \int f(y \mid \theta) f(\theta)\, d\theta \tag{3}$$

Introduction to Bayesian data analysis

If θ is a discrete random variable taking one value from the set $\{\theta_1, \ldots, \theta_n\}$, then

$$f(y) = \sum_{i=1}^{n} f(y \mid \theta_i) P(\theta = \theta_i) \tag{4}$$

Introduction to Bayesian data analysis

Without the normalizing constant, we have the relationship:

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{5}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{6}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The likelihood function will tell us $P(\text{data} \mid \theta)$:

```r
dbinom(46, 100, 0.5)
```

```
## [1] 0.057958
```

Note that

$$P(\text{data} \mid \theta) \propto \theta^{46}(1-\theta)^{54} \tag{7}$$

So, to get the posterior, we just need to work out a prior distribution $f(\theta)$.

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{8}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

For the prior, we need a distribution that can represent our uncertainty about the probabiliy θ of success. The Beta distribution is commonly used as prior for proportions. We say that the Beta distribution is conjugate to the binomial density; i.e., the two densities have similar functional forms.

The pdf is

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

In R, we write $X \sim \text{beta}(\texttt{shape1} = \alpha, \texttt{shape2} = \beta)$. The associated R function is $\texttt{dbeta}(\texttt{x}, \texttt{shape1}, \texttt{shape2})$.

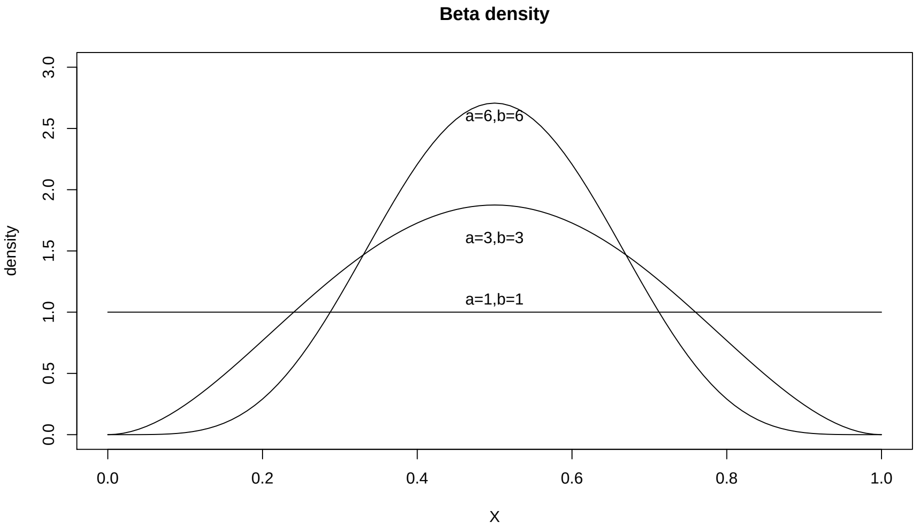
The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{9}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The Beta distribution's parameters a and b can be interpreted as (our beliefs about) prior successes and failures, and are called **hyperparameters**. Once we choose values for a and b, we can plot the Beta pdf. Here, we show the Beta pdf for three sets of values of a,b.

Example 1: Binomial Likelihood, Beta prior, Beta posterior



Example 1: Binomial Likelihood, Beta prior, Beta posterior

- If we don't have much prior information, we could use a=b=1; this gives us a uniform prior; this is called an uninformative prior or non-informative prior (although having no prior knowledge is, strictly speaking, not uninformative).
- If we have a lot of prior knowledge and/or a strong belief that θ has a particular value, we can use a larger a,b to reflect our greater certainty about the parameter.
- Notice that the larger our parameters a and b, the narrower the spread of the distribution; this makes sense because a larger sample size (a greater number of successes a, and a greater number of failures b) will lead to more precise estimates.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

Just for the sake of argument, let's take four different beta priors, each reflecting increasing certainty.

1. Beta(a=2,b=2)
2. Beta(a=3,b=3)
3. Beta(a=6,b=6)
4. Beta(a=21,b=21)

Each reflects a belief that $\theta = 0.5$, with varying degrees of (un)certainty. Now we just need to plug in the likelihood and the prior:

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{10}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The four corresponding posterior distributions would be:

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{48-1}(1-\theta)^{56-1} \quad (11)$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{49-1}(1-\theta)^{57-1} \quad (12)$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{52-1}(1-\theta)^{60-1} \quad (13)$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{67-1}(1-\theta)^{75-1} \quad (14)$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

We can now visualize each of these triplets of priors, likelihoods and posteriors. Note that I use the beta to model the likelihood because this allows me to visualize all three (prior, lik., posterior) in the same plot. The likelihood function is shown in the next slide.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

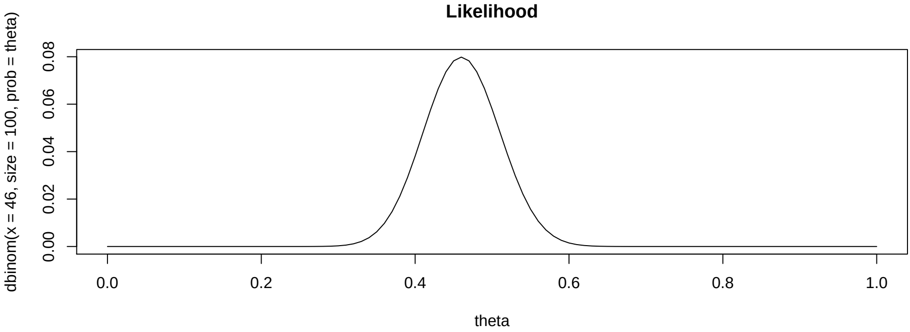


Figure 1: Binomial likelihood function.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

We can represent the likelihood in terms of the beta as well:

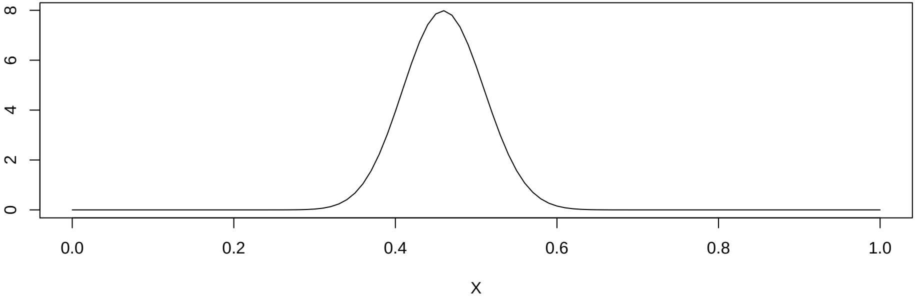
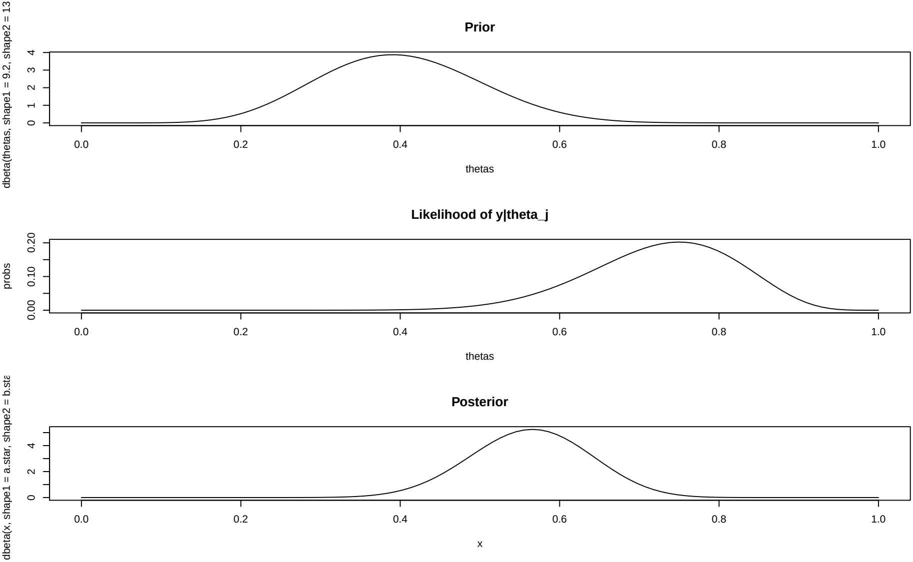


Figure 2: Using the beta distribution to represent a binomial likelihood function.

Example 1: Binomial Likelihood, Beta prior, Beta posterior



Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This is also a contrived example. Suppose we are modeling the number of times that a speaker says the word "the" per day.

The number of times x that the word is uttered in one day can be modeled by a Poisson distribution:

$$f(x \mid \theta) = \frac{\exp(-\theta)\theta^x}{x!} \tag{15}$$

where the rate θ is unknown, and the numbers of utterances of the target word on each day are independent given θ.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

We are told that the prior mean of θ is 100 and prior variance for θ is 225. This information could be based on the results of previous studies on the topic.

In order to visualize the prior, we first fit a Gamma density prior for θ based on the above information.

Note that we know that for a Gamma density with parameters a, b, the mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$. Since we are given values for the mean and variance, we can solve for a,b, which gives us the Gamma density.

If $\frac{a}{b} = 100$ and $\frac{a}{b^2} = 225$, it follows that $a = 100 \times b = 225 \times b^2$ or $100 = 225 \times b$, i.e., $b = \frac{100}{225}$.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This means that $a = \frac{100\times 100}{225} = \frac{10000}{225}$. Therefore, the Gamma distribution for the prior is as shown below (also see Fig 3):

$$\theta \sim Gamma(\frac{10000}{225}, \frac{100}{225}) \tag{16}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

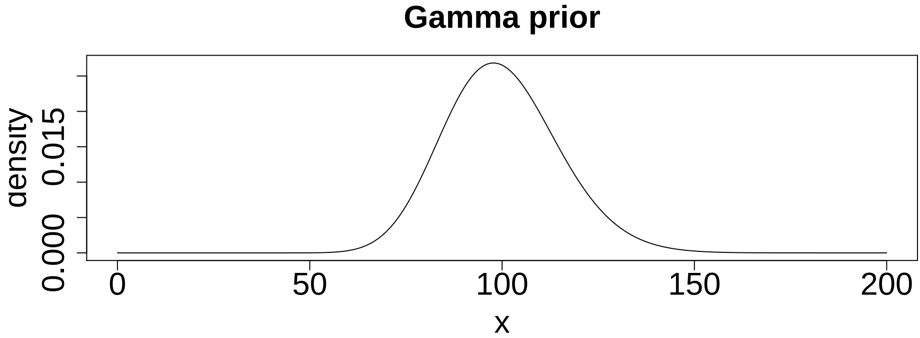


Figure 3: The Gamma prior for the parameter theta.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Given that

$$\text{Posterior} \propto \text{Prior Likelihood} \tag{17}$$

and given that the likelihood is:

$$\begin{aligned} L(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} \frac{\exp(-\theta)\theta^{x_i}}{x_i!} \\ &= \frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^{n} x_i!} \end{aligned} \tag{18}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

we can compute the posterior as follows:

$$\text{Posterior} = [\frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!}][\frac{b^a\theta^{a-1}\exp(-b\theta)}{\Gamma(a)}] \tag{19}$$

Disregarding the terms $x!, \Gamma(a), b^a$, which do not involve θ, we have

$$\begin{aligned} \text{Posterior} &\propto \exp(-n\theta)\theta^{\sum_i^n x_i}\theta^{a-1}\exp(-b\theta) \\ &= \theta^{a-1+\sum_i^n x_i}\exp(-\theta(b+n)) \end{aligned} \tag{20}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

First, note that the Gamma distribution in general is $Gamma(a, b) \propto \theta^{a-1} \exp(-\theta b)$. So it's enough to state the above as a Gamma distribution with some parameters a, b.

If we equate $a^* - 1 = a - 1 + \sum_i^n x_i$ and $b^* = b + n$, we can rewrite the above as:

$$\theta^{a^*-1} \exp(-\theta b^*) \tag{21}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This means that $a^* = a + \sum_i^n x_i$ and $b^* = b + n$. We can find a constant k such that the above is a proper probability density function, i.e.:

$$\int_{-\infty}^{\infty} k\theta^{a^*-1}\exp(-\theta b^*) = 1 \tag{22}$$

Thus, the posterior has the form of a Gamma distribution with parameters $a^* = a + \sum_i^n x_i, b^* = b + n$. Hence the Gamma distribution is a conjugate prior for the Poisson.

Concrete example given data

Suppose the number of "the" utterances is: $115, 97, 79, 131$.

Suppose that the prior is Gamma(a=10000/225,b=100/225). The data are as given; this means that $\sum_i^n x_i = 422$ and sample size $n = 4$. It follows that the posterior is

$$
\begin{aligned}
Gamma(a^* = a + \sum_i^n x_i, b^* = b + n) =& Gamma(\frac{10000}{225} + 422, 4 + \frac{100}{225}) \\
=& Gamma(466.44, 4.44)
\end{aligned}
\tag{23}
$$

The mean and variance of this distribution can be computed using the fact that the mean is $\frac{a*}{b*} = 466.44/4.44 = 104.95$ and the variance is $\frac{a*}{b*^2} = 466.44/4.44^2 = 23.66$.

Concrete example given data

```
### load data:
data<-c(115,97,79,131)

a.star<-function(a,data){
  return(a+sum(data))
}

b.star<-function(b,n){
  return(b+n)
}

new.a<-a.star(10000/225,data)
new.b<-b.star(100/225,length(data))
```

Concrete example given data

```
### post. mean
post.mean<-new.a/new.b
### post. var:
post.var<-new.a/(new.b^2)

new.data<-c(200)

new.a.2<-a.star(new.a,new.data)
new.b.2<-b.star(new.b,length(new.data))

### new mean
new.post.mean<-new.a.2/new.b.2
### new var:
new.post.var<-new.a.2/(new.b.2^2)
```

The posterior is a weighted mean of prior and likelihood

We can express the posterior mean as a weighted sum of the prior mean and the maximum likelihood estimate of θ.

The posterior mean is:

$$\frac{a*}{b*} = \frac{a + \sum x_i}{n + b} \tag{24}$$

This can be rewritten as

$$\frac{a*}{b*} = \frac{a + n\bar{x}}{n + b} \tag{25}$$

Dividing both the numerator and denominator by b:

The posterior is a weighted mean of prior and likelihood

$$\frac{a*}{b*} = \frac{(a + n\bar{x})/b}{(n + b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \tag{26}$$

The posterior is a weighted mean of prior and likelihood

Since a/b is the mean m of the prior, we can rewrite this as:

$$\frac{a/b + n\bar{x}/b}{1 + n/b} = \frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \tag{27}$$

We can rewrite this as:

The posterior is a weighted mean of prior and likelihood

$$\frac{m+\frac{n}{b}\bar{x}}{1+\frac{n}{b}} = \frac{m\times 1}{1+\frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1+\frac{n}{b}} \tag{28}$$

This is a weighted average: setting $w_1 = 1$ and $w_2 = \frac{n}{b}$, we can write the above as:

$$m\frac{w_1}{w_1+w_2} + \bar{x}\frac{w_2}{w_1+w_2} \tag{29}$$

The posterior is a weighted mean of prior and likelihood

A n approaches infinity, the weight on the prior mean m will tend towards 0, making the posterior mean approach the maximum likelihood estimate of the sample.

In general, in a Bayesian analysis, as sample size increases, the likelihood will dominate in determining the posterior mean.

Regarding variance, since the variance of the posterior is:

$$\frac{a*}{b*^2} = \frac{(a + n\bar{x})}{(n + b)^2} \tag{30}$$

as n approaches infinity, the posterior variance will approach zero: more data will reduce variance (uncertainty).

Summary

We saw two examples where we can do the computations to derive the posterior using simple algebra. There are several other such simple cases. However, in realistic data analysis settings, we cannot specify the posterior distribution as a particular density. We can only specify the priors and the likelihood.

For such cases, we need to use MCMC sampling techniques so that we can sample from the posterior distributions of the parameters.

Some sampling approaches are:

- Gibbs sampling using inversion sampling
- Metropolis-Hasting
- Hamiltonian Monte Carlo

We won't discuss these in this course.