

STAT 238 - Bayesian Statistics

Lecture Eight

Spring 2026, UC Berkeley

Aditya Guntuboyina

06 Feb 2026

The main goal today is to go over Bayesian inference of a parameter $\theta \in [0, 1]$ from observation $X \sim \text{Bin}(n, \theta)$ (n is also observed) with a Beta distribution prior on θ .

Let us start by recalling the Beta distribution.

1 Beta Distribution

The Beta distribution with parameters a and b is given by the density:

$$f_{\theta}(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} I\{0 \leq \theta \leq 1\}$$

where the normalizing constant $B(a, b)$ is the Beta function given by

$$B(a, b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta.$$

For the Beta density to be proper, we need both a and b to be strictly positive. Here are some basic properties of the Beta distribution:

1. When $a = 1$ and $b = 1$, we get the uniform distribution on $(0, 1)$. We consider the uniform distribution as the simplest example of a Beta distribution.
2. The mean of the Beta distribution is

$$\text{mean} = \frac{a}{a+b}.$$

For example, if a is much smaller compared to b , then the mean will be small.

3. The variance of the Beta distribution is

$$\text{variance} = \frac{a}{a+b} \frac{b}{a+b} \frac{1}{a+b+1}.$$

An implication is that, when $a + b$ is large, the variance tends to be small, so the Beta density will look skinny.

4. More generally, any moment $\mathbb{E}(\theta^k)$ (for positive integers k) can be written in closed form as (here $\theta \sim \text{Beta}(a, b)$)

$$\mathbb{E}(\theta^k) = \frac{a(a+1)\dots(a+k-1)}{(a+b)(a+b+1)\dots(a+b+k-1)} \quad \text{for every } k \geq 1.$$

In Bayesian inference, improper priors are also sometimes used. In the case of Beta priors, it is common to use the distributions $Beta(0, 0)$ as a prior. This corresponds to the density

$$f_{Beta(0,0)}(\theta) \propto \frac{I\{0 < \theta < 1\}}{\theta(1-\theta)}.$$

This improper density cannot be normalized so we don't put any normalizing constants on the right hand side.

If one wants to interpret $Beta(0, 0)$ as a probability distribution, the correct answer is the discrete distribution taking the values 0 and 1 with equal probability:

$$Beta(0, 0) = Bernoulli(0.5).$$

One way to formalize this equivalence is to argue that the distribution $Beta(\epsilon, \epsilon)$ converges to $Bernoulli(0.5)$ as $\epsilon \downarrow 0$ in the usual sense of weak convergence. The easiest way to see this is that the moments of $Beta(\epsilon, \epsilon)$ converge to the moments of $Bernoulli(0.5)$.

2 Beta-Binomial Inference

The goal is to infer θ from observations X (and n) where $X \sim Bin(n, \theta)$. One can also formulate this problem as that of estimating θ from n i.i.d Bernoulli observations $Z_1, \dots, Z_n \stackrel{\text{i.i.d}}{\sim} Ber(\theta)$. The likelihood is given by:

$$f_{X|\theta}(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \propto \theta^x (1-\theta)^{n-x}.$$

The frequentist estimator (MLE) for θ is x/n . For Bayesian inference, we will use the $Beta(a, b)$ prior:

$$f_{\theta}(\theta) \propto \theta^{a-1} (1-\theta)^{b-1} I\{0 \leq \theta \leq 1\}$$

leading to the posterior:

$$\begin{aligned} f_{\theta|x}(\theta) &\propto \theta^{a-1} (1-\theta)^{b-1} I\{0 \leq \theta \leq 1\} \theta^x (1-\theta)^{n-x} \\ &= \theta^{x+a-1} (1-\theta)^{n-x+b-1} I\{0 \leq \theta \leq 1\}. \end{aligned}$$

Thus

$$\theta | x \sim Beta(x + a, n - x + b).$$

A natural point estimate of θ is the posterior mean:

$$\hat{\theta} = \mathbb{E}(\theta | x) = \frac{x + a}{n + a + b}.$$

Here are two things to note about the posterior mean:

1. It is a convex combination of the MLE and the prior mean $a/(a+b)$:

$$\frac{x + a}{n + a + b} = \left(\frac{n}{a + b + n} \right) \frac{x}{n} + \left(\frac{a + b}{a + b + n} \right) \frac{a}{a + b}.$$

So if n is much larger compared to $a + b$, then the posterior mean will be very close to the MLE. Conversely if n is much smaller compared to $a + b$, then the posterior mean will be very close to the prior mean.

- When $a = b = 0$, then the posterior mean exactly coincides with the MLE. Thus if we use the $Beta(0, 0)$ prior, then posterior inference will be close to frequentist inference. Note again that this prior is improper.
- If we use the $Beta(0, 0)$ prior and if we observe $n = x = 1$ (i.e., we only have data on one Bernoulli trial which resulted in a heads). Then the posterior is $Beta(x+a, n-x+b) = Beta(1, 0)$. $Beta(1, 0)$ is also improper but it can be interpreted as the point mass at 1:

$$Beta(1, 0) = Bernoulli(1) = \delta_{\{1\}}.$$

This can be proved, for example, by computing the moments of $Beta(1, \epsilon)$ and showing that they all approach 1 (which are the moments of $\delta_{\{1\}}$) as $\epsilon \rightarrow 0$. Similarly for $x = 0$ and $n = 1$, the posterior becomes $Beta(0, 1)$ which should be interpreted as $\delta_{\{0\}}$. If $n = 2$ and $x = 1$ (i.e., we observe one head and one tail), the posterior is the uniform distribution on $(0, 1)$.

- The process of going from the $Beta(0, 0)$ prior to the $Beta(1, 0)$ posterior when $n = x = 1$ is described as intuitive in the following situation by Jaynes [1, Section 12.4.3]: *...in a chemical laboratory we find a jar containing an unknown and unlabeled compound. We are at first completely ignorant as to whether a small sample of this compound will dissolve in water or not. But, having observed that one small sample does dissolve, we infer immediately that all samples of this compound are water soluble, and although the conclusion does not carry quite the force of deductive proof, we feel strongly that the inference was justified.*

3 Multiple Replications of the Same Problem

Consider the problem of estimating θ_i from (X_i, n_i) for each $i = 1, \dots, N$, with $X_i \sim Bin(n_i, \theta_i)$. For a concrete example, take all counties in the United States with X_i denoting the number of deaths due to kidney cancer (in the period 1980 – 89) and n_i denoting the average population (during the same period) for the i -th county.

It is easy to see that the naive frequentist estimate X_i/n_i can be very bad for θ_i if n_i is small. We will therefore use Bayes estimation using the prior:

$$\theta_i \stackrel{\text{i.i.d.}}{\sim} Beta(a, b).$$

The choice of a and b is crucial here. Since we have a large dataset, it makes sense to learn good values of a and b from the observed data. We will look at the following three ways of doing this.

- Method One:** We place a threshold on n_i and filter out i for which n_i exceeds the threshold. We then select a and b by fitting the $Beta(a, b)$ density to the data $\{X_i/n_i : n_i \geq \text{threshold}\}$. The Beta density fitting can be done by just matching mean and variances. Let m and V denote the mean and variance of $\{X_i/n_i : n_i \geq \text{threshold}\}$. Then we obtain a and b by solving:

$$\frac{a}{a+b} = m \quad \text{and} \quad \frac{a}{a+b} \frac{b}{a+b} \frac{1}{a+b+1} = V$$

which gives

$$\hat{a} = m \left(\frac{m(1-m)}{V} - 1 \right) \quad \text{and} \quad \hat{b} = (1-m) \left(\frac{m(1-m)}{V} - 1 \right).$$

One drawback of this method is that the selection of threshold can be arbitrary. In the kidney cancer example, we choose 300000 as the threshold, but we could have also chosen some other threshold such as 350K or 400K.

2. **Method Two:** We consider the marginal likelihood of X_i given a, b . Note that

$$\begin{aligned}\mathbb{P}\{X_i = x \mid a, b\} &= \int_0^1 \mathbb{P}\{X_i = x \mid \theta_i = \theta\} f_{\theta_i \mid a, b}(\theta) d\theta \\ &= \int_0^1 \binom{n_i}{x} \theta^x (1 - \theta)^{n_i - x} \frac{\theta^{a-1} (1 - \theta)^{b-1}}{B(a, b)} d\theta \\ &= \binom{n_i}{x} \frac{1}{B(a, b)} \int_0^1 \theta^{x+a-1} (1 - \theta)^{n_i - x + b - 1} d\theta \\ &= \binom{n_i}{x} \frac{B(x + a, n_i - x + b)}{B(a, b)}\end{aligned}$$

As a result, the marginal likelihood of the data given in terms of a, b is:

$$\prod_{i=1}^N \binom{n_i}{X_i} \frac{B(X_i + a, n_i - X_i + b)}{B(a, b)}$$

So the maximum likelihood estimates of a, b are given by:

$$(\hat{a}_{\text{MLE}}, \hat{b}_{\text{MLE}}) = \underset{a, b}{\operatorname{argmax}} \prod_{i=1}^N \binom{n_i}{X_i} \frac{B(X_i + a, n_i - X_i + b)}{B(a, b)}.$$

This maximization can be done numerically by taking a grid of a and b values.

3. **Method Three:** This is the fully Bayes solution where we simply place priors on a and b . It is sensible to reparametrize as:

$$\mu = \frac{a}{a + b} \quad \text{and} \quad \kappa = a + b.$$

Now we take the priors:

$$\mu \sim \text{uniform}(0, 1) \quad \text{and} \quad \log \kappa \sim \text{uniform}(-\infty, \infty).$$

Inference under this fully Bayesian model is nontrivial from a computational point of view. We shall discuss this in the next lecture.

References

- [1] Jaynes, E. T., (2003) *Probability theory: the logic of science*. Cambridge University Press, 2003.