STAT 535: Chapter 3:
The Beta-Binomial Bayesian Model

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

Some General Notation

- **Notation**: We hereby denote our data as the $n \times k$ matrix $\boldsymbol{Y}$.
- We denote the parameter(s) of interest (possibly multidimensional) to be the vector $\boldsymbol{\theta}$.
- We will denote our posterior distribution for $\boldsymbol{\theta}$ using $p(\boldsymbol{\theta}|\boldsymbol{Y})$.

Likelihood Theory

- The likelihood function $L(\theta|\boldsymbol{Y})$ is a function of $\boldsymbol{\theta}$ that shows how "likely" are various parameter values $\boldsymbol{\theta}$ to have produced the data $\boldsymbol{Y}$ that **were observed**.
- In classical statistics, the specific value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\boldsymbol{Y})$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.
- In many common probability models, when the sample size n is large, $L(\boldsymbol{\theta}|\boldsymbol{Y})$ is unimodal in $\boldsymbol{\theta}$.
- **Note**: Unlike $p(\theta|\boldsymbol{Y})$, $L(\theta|\boldsymbol{Y})$ does **not necessarily** obey the usual laws for probability distributions.
- Also, in the classical framework, all the randomness within $L(\theta|\boldsymbol{Y})$ is attached to $\boldsymbol{Y}$, not to $\boldsymbol{\theta}$.

Likelihood Theory

- Mathematically, if the data $\boldsymbol{Y}$ represent iid observations from probability distribution $p(\boldsymbol{Y}|\boldsymbol{\theta})$, then

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \prod_{i=1}^{n} p(\boldsymbol{Y}_i|\boldsymbol{\theta})$$

  (where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are the n data vectors).
- The **Likelihood Principle** of Birnbaum states that (given the data) all of the evidence about $\boldsymbol{\theta}$ is contained in the likelihood function.
- Likelihood Principle implies: Two experiments that yield equal (or proportional) likelihoods should produce equivalent inference about $\boldsymbol{\theta}$.

The Bayesian Framework

- Suppose we observe an iid sample of data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$.
- Now $\boldsymbol{Y}$ is considered fixed and known.
- We also **must** specify $p(\boldsymbol{\theta})$, the prior distribution for $\boldsymbol{\theta}$, based on any knowledge we have about $\boldsymbol{\theta}$ **before** observing the data.
- Our model for the distribution of the data will give us the likelihood

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \prod_{i=1}^{n} p(\boldsymbol{Y}_i|\boldsymbol{\theta}).$$

The Bayesian Framework

- Then by Bayes' Rule, our posterior distribution is

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{Y})}{p(\boldsymbol{Y})} = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{Y})}{\int_{\Theta} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{Y})\, d\boldsymbol{\theta}}$$

- Note that the **marginal distribution** of $\boldsymbol{Y}$, $p(\boldsymbol{Y})$, is simply the joint density $p(\boldsymbol{\theta}, \boldsymbol{Y})$ (i.e., the numerator) with $\boldsymbol{\theta}$ integrated out.
- With respect to $\boldsymbol{\theta}$, it is simply a **normalizing constant** that ensures that $p(\boldsymbol{\theta}|\boldsymbol{Y})$ integrates to 1.

The Bayesian Framework

- Since $p(\boldsymbol{Y})$ carries no information about $\boldsymbol{\theta}$, for conciseness we may drop it and write

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{Y}).$$

- Often we can calculate the posterior distribution by multiplying the prior by the likelihood and **then** normalizing the posterior at the **last** step, by including the necessary constant.
- Having presented the Bayesian framework in general, we now look at a specific example of a very common Bayesian model.

Examples of the Beta-Binomial Model

- Recall the model for, say, Y, the number of games (out of 6) that Kasparov would win in the tournament against Deep Blue.
- We model Y as binomial with parameters $n = 6$ and success probability $\pi \in [0, 1]$.
- The book gives the example of a candidate running for office. If the probability of a randomly selected voter supporting the candidate is π, then the number of voters in a random sample of 50 voters who support her is binomial$(50, \pi)$.

A Prior Distribution for π

- Since the parameter π is restricted to be between 0 and 1, we should choose a prior distribution with **support** on $[0, 1]$.
- Let $f(\pi)$ denote the prior probability density function (pdf) for π.
- Note $f(\pi)$ has the usual properties of a pdf: It is non-negative everywhere, and it integrates to 1 over its support (which is $[0, 1]$ in this example).
- The formula for the pdf of a **Beta** prior distribution for π is:

$$f(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}, \; 0 \le \pi \le 1,$$

  where $\alpha > 0$ and $\beta > 0$ are the **hyperparameters** of this prior model.
- Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$.

Properties of the Beta Distribution

- In a real problem, we need to specify the values of our hyperparameters α and β of our prior.
- Ideally our choices of α and β should reflect our prior beliefs about π.
- If we have no prior idea what π is, we could set $\alpha = \beta = 1$, which corresponds to a Uniform$(0, 1)$ prior for π: completely flat, so that all values of π are equally likely *a priori*.
- If we have more informative prior beliefs about the value of π, we could choose α and β to reflect that.
- Plots of the Beta pdf for various values of α and β can help inform the prior specification (see R examples).

Expected Value of the Beta

- The expected value of a Beta(α, β) r.v. is
$$\frac{\alpha}{\alpha + \beta}.$$
- So if our prior belief is that π is closer to 0 than to 1, we should choose our hyperparameters α and β such that $\alpha < \beta$.
- If our prior belief is that π is closer to 1 than to 0, we should set $\alpha > \beta$.
- The mode (location where the pdf reaches its maximum) for the Beta(α, β) pdf is
$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

Variance of the Beta

- The variance of a Beta(α, β) r.v. is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

  and the standard deviation is the square root of this.
- So if our prior belief is strong that π is near a certain value, we can pick α and β so that this variance is **small**.
- If our prior belief is less certain, we can pick α and β so that this variance is **large**.

Choosing the Hyperparameters of the Beta

- The `plot_beta` function in the `bayesrules` package can help us pick α and β by trial and error.
- Example: If we believe the value of π is around 0.45, we could choose many sets of α and β that would yield $E(\pi) = 0.45$.
- For example, $\alpha = 9$ and $\beta = 11$; $\alpha = 18$ and $\beta = 22$; $\alpha = 45$ and $\beta = 55$.
- Plotting the Beta$(45, 55)$ pdf shows that this choice of priors indicates we believe that π is very likely to be between 0.3 and 0.6.
- Check: For a Beta$(45, 55)$ distribution, the standard deviation is 0.05.
- So the interval $(0.3, 0.6)$ is within three standard deviations of the mean.

The Binomial Model for the Data

- Political candidate example: Suppose we plan to conduct a poll of 50 randomly selected voters and count how many of these 50 voters support our candidate.
- Given π, the number of the 50 voters who support her (denote this as $Y|\pi$) is a binomial$(50, \pi)$ random variable with probability mass function (pmf):

$$f(y|\pi) = P(Y = y|\pi) = \binom{50}{y} \pi^y (1-\pi)^{50-y}.$$

- This pmf tells us: If the success probability is π, what is the probability that the total number of supportive voters Y equals some value y?

The Likelihood Using the Binomial Model

- Suppose we take the sample and find that $Y = 30$ of the 50 sampled voters support her.
- We could calculate the **likelihood** of π given $y = 30$:

$$L(\pi | y = 30) = \binom{50}{30} \pi^{30} (1 - \pi)^{50-30}.$$

- This likelihood tells us: Given that $y = 30$ of the 50 voters were supportive, what is the likelihood of any particular binomial probability π?
- Some examples: The likelihood that $\pi = 0.6$ given $y = 30$ is

$$L(\pi = 0.6 | y = 30) = \binom{50}{30} 0.6^{30} (0.4)^{20} \approx 0.115.$$

- The likelihood that $\pi = 0.5$ given $y = 30$ is

$$L(\pi = 0.5 | y = 30) = \binom{50}{30} 0.5^{30} (0.5)^{20} \approx 0.042.$$

Maximizing the Likelihood with the Binomial Model

- Using calculus, you can show that the likelihood here is maximized when $\pi = 0.6$.
- So $\hat{\pi} = 0.6$ (which is just the sample proportion $30/50$ here) is called the maximum likelihood estimate (MLE) of π for this data set.
- Note that this **maximum likelihood estimation** approach does not use the prior information to help estimate π; it only uses the information in the sample data.

The Beta Posterior Model

- The prior tells us information about the value of π, based on our prior knowledge.
- Candidate example: We believe *a priori* that the value of π is near 0.45.
- The likelihood tells us information about the value of π, based on information in our data.
- Candidate example: We believe *based on the data* that the value of π is near 0.6.
- The **posterior distribution** balances the information in the prior and the data.
- The **posterior** uses the data information to **update** the prior information.
- See the R plots to visually assess the position of the posterior relative to the prior and the likelihood.

Mathematical Development of the Posterior

- The posterior density function is denoted $f(\pi|y)$ and by Bayes' Rule, this is

$$f(\pi|y) = \frac{f(\pi)f(y|\pi)}{f(y)} = \frac{f(\pi)L(\pi|y)}{f(y)}$$

- The denominator $f(y)$ is just a normalizing constant and we don't actually have to calculate it.
- We can use the fact that the posterior is **proportional to** the prior times the likelihood, i.e.,

$$f(\pi|y) \propto f(\pi) \times L(\pi|y)$$

- Candidate example:

$$\begin{aligned} f(\pi|y) &\propto \pi^{45-1}(1-\pi)^{55-1}\pi^{30}(1-\pi)^{20} \\ &= \pi^{74}(1-\pi)^{74} \end{aligned}$$

We Only Need the Kernel of the Posterior

- Notice that we can ignore **all** of the normalizing constants in the likelihood and the prior.
- This leaves us with only the **kernel** of the posterior distribution.
- but we recognize this as the kernel of a Beta$(75, 75)$ distribution for π.
- So the posterior distribution of π is Beta$(75, 75)$.

General Formula for the Beta Posterior

- In general, if $Y|\pi \sim Bin(n, \pi)$ (data model) and $\pi \sim Beta(\alpha, \beta)$ (prior model), then the posterior model will be:
$$\pi|y \sim Beta(\alpha + y, \beta + n - y).$$

- So the posterior expected value is
$$E(\pi|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

- The posterior mode is
$$Mode(\pi|y) = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}$$
and the posterior variance is
$$Var(\pi|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

Possible Point Estimators Based on the Posterior

- Either the posterior mean (expected value) or the posterior mode could be used as an estimator of π.
- An estimator based on the **posterior** would take into account **both** the prior information and the data information.

Conjugate Prior

- A **conjugate prior** is one for which the prior distribution and the posterior distribution have the same family (same functional form), just with different (updated) parameters.
- For example, in the Beta-binomial model, the prior is a Beta and the posterior is also a Beta, so this was a conjugate prior.
- Again, the prior's parameters reflect only our prior knowledge (via α and β) whereas the posterior's parameters reflect both the prior and the data (via α, β, y, and n).

Inference with Beta-Binomial Model

- Consider letting the Bayesian point estimate of π be $\hat{\pi}_B =$ the posterior mean.
- The mean of the (posterior) beta distribution is:

$$\hat{\pi}_B = \frac{y+\alpha}{y+\alpha+n-y+\beta} = \frac{y+\alpha}{\alpha+\beta+n}$$

$$\begin{aligned} \text{Note } \hat{\pi}_B &= \frac{y}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta+n} \\ &= \left[\frac{n}{\alpha+\beta+n}\right]\left(\frac{y}{n}\right) + \left[\frac{\alpha+\beta}{\alpha+\beta+n}\right]\left(\frac{\alpha}{\alpha+\beta}\right) \end{aligned}$$

Inference with Beta/Binomial Model

- So the Bayes estimator $\hat{\pi}_B$ is a weighted average of the usual frequentist estimator (sample mean, i.e., the sample proportion of "successes" here) and the prior mean.
- As $n \uparrow$, the **sample data** are weighted **more** heavily and the **prior** information **less** heavily.
- In general, with Bayesian estimation, as the sample size increases, the **likelihood dominates the prior**.
- See R example with credit card debt data.