STAT7630: Bayesian Statistics
Lecture Slides # 3

Chapter 3 - The Beta-Binomial Bayesian Model

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: August, 2024

Outline

Notation

- Denote our data as the $n \times k$ matrix $\mathbf{Y}$.
- Denote the parameter(s) of interest (possibly multidimensional) as the vector $\boldsymbol{\theta}$.
- The posterior distribution for $\boldsymbol{\theta}$ is denoted by $p(\boldsymbol{\theta}|\mathbf{Y})$.

Likelihood Function

- The likelihood function $L(\boldsymbol{\theta}|\mathbf{Y})$ is a function of $\boldsymbol{\theta}$ that shows how "likely" various parameter values $\boldsymbol{\theta}$ are to have produced the observed data $\mathbf{Y}$.
- In classical statistics, the specific value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{Y})$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.
- For large sample sizes n, $L(\boldsymbol{\theta}|\mathbf{Y})$ is often unimodal in $\boldsymbol{\theta}$.
- Unlike $p(\boldsymbol{\theta}|\mathbf{Y})$, $L(\boldsymbol{\theta}|\mathbf{Y})$ does not necessarily obey the usual laws for probability distributions.

Mathematical Formulation

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

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

where $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are the n data vectors.

The Likelihood Principle

- The Likelihood Principle of Birnbaum states that, given the data, all evidence about $\boldsymbol{\theta}$ is contained in the likelihood function.
- It implies that two experiments yielding equal likelihoods should produce equivalent inference about $\boldsymbol{\theta}$.

The Bayesian Framework

- Suppose we observe an iid sample of data $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$.
- Now $\mathbf{Y}$ 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:

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

The Bayesian Framework

- Then by Bayes' Rule, our posterior distribution is:

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

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

The Bayesian Framework

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

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

- 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.

Outline

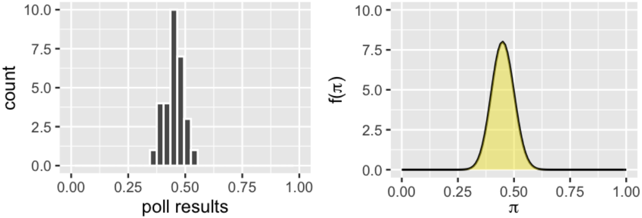
Outline

Examples of the Beta-Binomial Model

- **Example (Kasparov vs Deep Blue):**
- Recall the model for 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 $\pi \in [0, 1]$.
- **Example (Candidate Running for Office):**
- The book gives the example of a candidate (Michelle) 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$(50, \pi)$.

Introduction to Michelle's Election Model

- Michelle is running for president, and you've conducted 30 different polls in Minnesota.
- Michelle's support has varied between 35% and 55%, with an average of 45%.
- The results of these polls can be organized into a continuous prior probability model for π, the proportion of Minnesotans supporting Michelle.
- The left plot shows a histogram of poll results, and the right plot shows a density plot for π.



Outline

Building the Prior Model

- Elections are dynamic, but past polls can provide prior information about π.
- We construct a continuous prior probability model of π based on the polls.
- This prior model allows π to take any value between 0 and 1, most likely around 0.45.
- 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).

A Prior Distribution for π

- A reasonable prior is represented by a Beta distribution, which we'll explore further in this chapter.
- The Beta distribution is defined by two shape parameters, α and β, which determine the distribution's shape.
- The formula for the pdf of a Beta prior distribution for π is:

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

  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, meaning 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.

Expected Value of the Beta

- The expected value of a Beta(α, β) random variable is:

$$\mathbf{E}[\pi] = \frac{\alpha}{\alpha + \beta}.$$

- If our prior belief is that π is closer to 0 than to 1, we should choose our hyperparameters $\alpha < \beta$.
- 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:

$$\text{Mode} = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

Variance of the Beta

- The variance of a Beta(α, β) random variable is:

$$\text{Var}(\pi) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

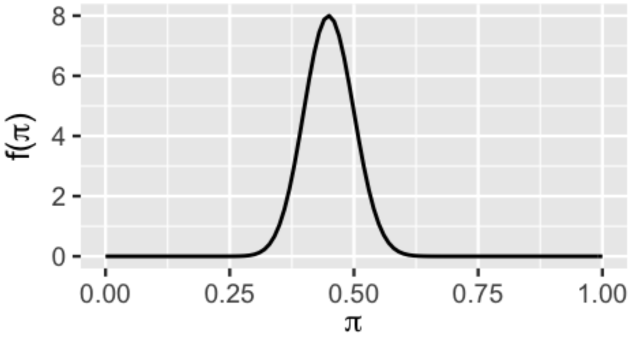
- The standard deviation is the square root of this variance.
- If our prior belief is strong that π is near a certain value, we can pick α and β so that this variance is small.
- If our prior belief is less certain, we can pick α and β so that this variance is large.

Selecting the Hyperparameters of the Beta Distribution

- The `plot_beta` function in the `bayesrules` package allows us to experiment with different values of α and β to find the best fit for our prior beliefs.
- For example, if we believe that π is around 0.45, various combinations of α and β could be used to achieve $\mathbf{E}[\pi] = 0.45$.
- Some options include $\alpha = 9$ and $\beta = 11$, $\alpha = 18$ and $\beta = 22$, or $\alpha = 45$ and $\beta = 55$.
- Plotting the Beta(45, 55) probability density function (pdf) indicates that π is most likely between 0.3 and 0.6.
- For the Beta(45, 55) distribution, the standard deviation is approximately 0.05, meaning the interval (0.3, 0.6) shows the region which is within three standard deviations of the mean.

Tuning the Beta Prior

- We tune the Beta prior model to reflect our understanding of Michelle's support.
- For example, $\alpha = 45$, $\beta = 55$ reflects an average support around 45%.
- The resulting Beta(45,55) prior captures the typical outcomes and variability observed in the polls.



Modeling with the Binomial Distribution

- Recall the poll which is conducted with 50 randomly selected voters in Minnesota, and the number of supporters of Michelle (denoted as $Y \mid \pi$) is modeled as a Binomial(50, π) random variable:

$$Y \mid \pi \sim \text{Binomial}(50, \pi)$$

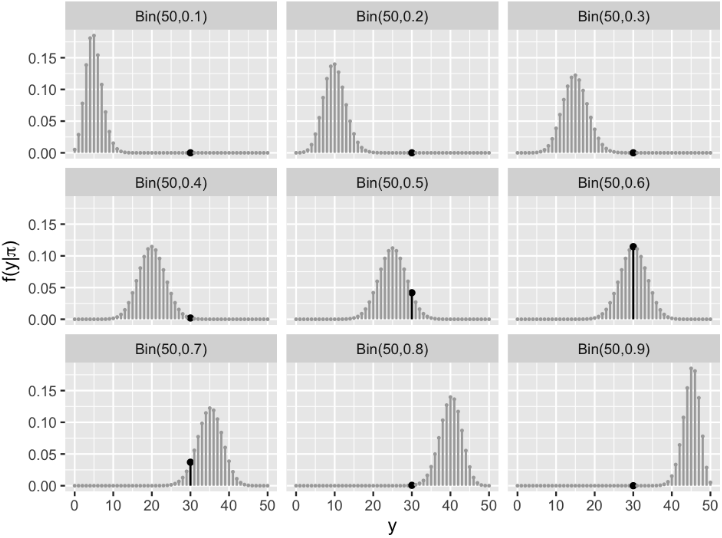
- The probability mass function (pmf) for this binomial distribution is given by:

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

- This pmf answers the question: Given a success probability π, what is the probability that exactly y out of the 50 voters support the candidate?
- The likelihood function describes the probability of observing the data given different values of π.

Outline

The Binomial Data Model



Calculating the Likelihood Using the Binomial Model

- Suppose we find that 30 out of the 50 voters support Michelle. We can then compute the likelihood of π given $y = 30$:

$$L(\pi \mid y = 30) = \binom{50}{30} \pi^{30} (1 - \pi)^{20}.$$

- This likelihood function tells us: Given that 30 voters were supportive, what is the likelihood of any particular binomial probability π?
- For instance, the likelihood that $\pi = 0.6$ given $y = 30$ is approximately 0.115, while the likelihood that $\pi = 0.5$ given $y = 30$ is approximately 0.042.

Maximizing the Likelihood with the Binomial Model

- Through calculus, it can be demonstrated that the likelihood function is maximized when $\pi = 0.6$.
- Therefore, the estimate $\hat{\pi} = 0.6$, which corresponds to the sample proportion $30/50$, is known as the maximum likelihood estimate (MLE) of π for this data set.
- It is important to note that this maximum likelihood estimation method relies solely on the information from the sample data and does not incorporate any prior information about π.

Outline

The Beta Posterior Model

- The prior distribution provides information about π based on our prior knowledge.
- Example: We might believe that π is close to 0.45 before observing any data.
- The likelihood function, on the other hand, reflects the information from the observed data.
- Example: Based on the data, we might estimate that π is close to 0.6.
- The posterior distribution combines the prior information with the data, updating our belief about π.
- You can use R plots to visually compare the posterior distribution with the prior and the likelihood.

Mathematical Development of the Posterior

- The posterior density function is denoted by $f(\pi \mid y)$. According to Bayes' Rule:

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

- The denominator $f(y)$ is simply a normalizing constant, ensuring that the posterior distribution integrates to 1.
- We can simplify this by noting that the posterior is proportional to the product of the prior and the likelihood:

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

- Example: For Michelle:

$$f(\pi \mid y) \propto \pi^{74}(1-\pi)^{74}$$

Complete Derivation of Beta-Binomial Bayesian Model

- Suppose we observe n independent Bernoulli(π) random variables $X_1, \ldots, X_n$.
- We wish to estimate the "success probability" π via the Bayesian approach.
- We will use a Beta(α, β) prior for π and show this is a conjugate prior.
- Consider the random variable $Y = \sum_{i=1}^{n} X_i$, which has a Binomial(n, π) distribution.
- First, write the joint density of Y and π (using $f(\cdot)$ to denote densities, not $p(\cdot)$, to avoid confusion with the parameter π).

Derivation of Beta-Binomial Model

$$
\begin{aligned}
f(y, \pi) &= f(y|\pi) f(\pi) \\
&= \binom{n}{y} \pi^y (1-\pi)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1} \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{y+\alpha-1} (1-\pi)^{n-y+\beta-1}
\end{aligned}
$$

Derivation of Beta-Binomial Model

- Although it is not really necessary, let's derive the marginal density of Y (this pdf is called the Beta-Binomial(n, α, β) distribution):

$$
\begin{aligned}
f(y) &= \int_0^1 f(y, \pi)\, d\pi \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}\, d\pi \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \\
&\times \int_0^1 \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}\, d\pi \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}
\end{aligned}
$$

Derivation of Beta-Binomial Model

- Then, the posterior $p(\pi \mid y) = f(\pi \mid y)$ is

$$
\begin{aligned}
f(\pi|y) &= \frac{f(y,\pi)}{f(y)} \\
&= \frac{\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}}{\frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}, \quad 0 \leq \pi \leq 1.
\end{aligned}
$$

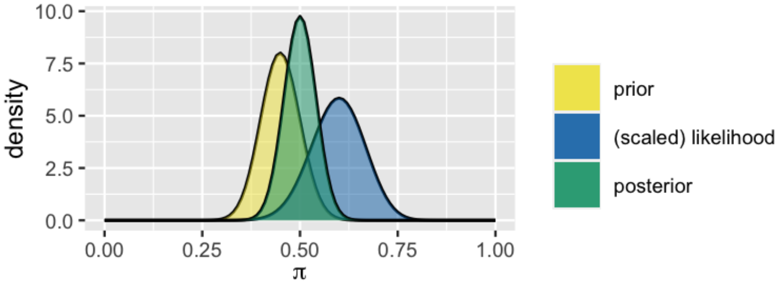
- Clearly, this posterior is a $\text{Beta}(\alpha + y, \beta + n - y)$ distribution.

The Beta Posterior Model

- Combining the prior and data, we construct the posterior model:

$$\pi \mid \mathbf{Y} = 30 \sim \text{Beta}(75, 75)$$

- The posterior model reflects the updated belief about π after incorporating the new poll results.
- The posterior strikes a balance between the prior and the data.



Focusing on the Kernel of the Posterior

- It is important to note that we can disregard all normalizing constants in both the likelihood and the prior.
- By doing so, we are left with only the **kernel** of the posterior distribution.
- In this case, we identify the kernel as corresponding to a Beta(75, 75) distribution for π.
- Thus, the posterior distribution of π is Beta(75, 75).

General Formula for the Beta Posterior

- Generally, if $Y \mid \pi \sim \text{Bin}(n, \pi)$ (data model) and $\pi \sim \text{Beta}(\alpha, \beta)$ (prior model), then the posterior distribution is:

$$\pi \mid y \sim \text{Beta}(\alpha + y, \beta + n - y).$$

- The posterior expected value (i.e. mean) is:

$$\mathbf{E}[\pi \mid y] = \frac{\alpha + y}{\alpha + \beta + n}.$$

- The posterior mode is:

$$\text{Mode}[\pi \mid y] = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}.$$

- The posterior variance is:

$$\text{Var}[\pi \mid y] = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}.$$

Choosing Point Estimators Based on the Posterior

- Both the posterior mean (expected value) and the posterior mode can serve as estimators for π. Posterior mean is also called Bayes estimator.
- An estimator derived from the posterior distribution takes into account both prior information and the observed data.

Outline

Conjugate Prior

- A **conjugate prior** is a prior distribution where the posterior distribution belongs to the same family (has the same functional form) as the prior, but with updated parameters.
- For instance, in the Beta-binomial model, the prior distribution is a Beta distribution, and the posterior distribution also remains a Beta distribution—hence, this is a conjugate prior.
- The parameters of the prior distribution represent our initial beliefs (through α and β), while the parameters of the posterior distribution incorporate both the prior beliefs and the data (through α, β, y, and n).

Inference with the Beta-Binomial Model

- Consider using the Bayesian point estimate $\hat{\pi}_B$, which is the posterior mean of π.
- The posterior mean of the Beta distribution is given by:

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

- This can also be expressed as a combination:

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

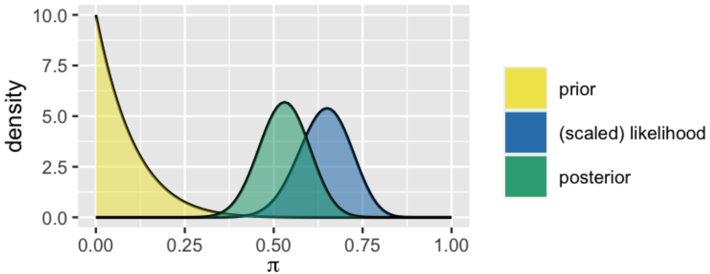
- where the first term is related to the sample data and the second term to the prior information.

Inference with the Beta/Binomial Model

- The Bayesian estimator $\hat{\pi}_B$ is essentially a weighted average of the frequentist estimator (sample mean or proportion of successes) and the prior mean.
- As the sample size n increases, the sample data have more influence, while the prior information becomes less influential.
- In general, with Bayesian estimation, as the sample size grows, the likelihood increasingly dominates the prior.
- For a practical illustration, see the R example using credit card debt data.

Milgram's Behavioral Study

- Milgram's study investigated the propensity to obey authority, even when it might harm others.
- We can analyze the study using the Beta-Binomial framework.
- The prior model $\pi \sim \text{Beta}(1, 10)$ reflects the psychologist's belief that a small proportion of people would obey authority.
- After observing the data, where 26 out of 40 participants administered the most severe shock, the posterior model is $\pi \mid \mathbf{Y} = 26 \sim \text{Beta}(27, 24)$.



Chapter Summary

- The Beta-Binomial model is a powerful tool for modeling proportions π between 0 and 1.
- The model combines prior information with new data to update beliefs about π.
- The posterior distribution is a Beta distribution with updated parameters reflecting both prior beliefs and observed data.
- This model is applicable in various settings where proportions are of interest, such as election polling, social behavior studies, and more.