STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 2A: The Beta Binomial Model

The Three Steps of the Bayesian Approach

- Set up a full probability model
  - A joint probability distribution for all observable and unobservable quantities in a problem.
  - The model should be consistent with knowledge about the underlying scientific problem and the data collection process
- Condition on the observed data
  - Calculate and interpret the appropriate posterior distribution
  - We are interested in the conditional probability distribution of the unobserved quantities of interest given the observed data
  - We are often interested in the marginal distribution of a subset of unobserved quantities
- Evaluate the model
  - Does the model fit the data?
  - Are substantive conclusions reasonable?
  - How sensitive are results to model assumptions?

Bayesian Framework

- Prior distribution for θ:
$$\theta \sim \pi(\theta)$$
- Sample distribution (or likelihood) of $\boldsymbol{X}$ given θ:
$$\boldsymbol{X}|\theta \sim f(\boldsymbol{x}|\theta) = \pi(\boldsymbol{x}|\theta)$$
- Joint distribution of $\boldsymbol{X}$ and θ (this is our full model):
$$f(x, \theta) = f(\boldsymbol{x}|\theta)\pi(\theta)$$
- Recall, chain rule of probability: $p(E_1 \cap E_2) = p(E_2|E_1)p(E_1)$
- Marginal distribution of $\boldsymbol{X}$:
$$m(\boldsymbol{x}) = \int_{\theta \in \Omega} f(\boldsymbol{x}, \theta) d\theta = \int_{\theta \in \Omega} f(\boldsymbol{x}|\theta)\pi(\theta) d\theta$$
- Posterior distribution of θ (conditional distribution of θ given $\boldsymbol{X}$):
$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \theta)}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \propto f(\boldsymbol{x}|\theta)\pi(\theta) \qquad \text{(Bayes' Rule)}$$

Exchangeability

- At the beginning of class 1A, we assumed the data generating process is iid. This is a typcial assumption in classical inference.
- In Bayesian inference, we will assume something weaker, *exchangeability*.
- For example, a sequence of coin tosses is exchangeable if $\pi(x_1, ..., x_n) = \pi(x_{\sigma(1)}, ..., x_{\sigma(n)})$ for every permutation σ of the order.
- Position and order is irrelevant, for any length of the sequence.
- Exchangeability is an assumption about the underlying symmetry of the sequence.
- For example: compare the probability of observing (H, H, T) with the probability of observing (H, T, H).
- Independent (Bernoulli) trials with x successes and $n - x$ failures are exchangeable: for any length, the probability is proportional to $\theta^x(1 - \theta)^{n-x}$
- Can you think of an exchangeable sequence that is not iid?

- Suppose we are interested in the outcomes of a coin toss, $X_1, ... X_n$.
- Then $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Binomial}(n, \theta) \propto \theta^x (1-\theta)^{n-x}$
- And suppose $\theta \sim \pi(\theta)$. Then,
- $\pi(1,0,0,1,1) = \int \theta^3 (1-\theta)^2 \pi\theta d\theta$
- $\pi(1,1,0,0,1) = \int \theta^3 (1-\theta)^2 \pi\theta d\theta$
- So it seems that $X_1, ... X_n$ are exchangeable.

- Claim: if $\theta \sim \pi(\theta)$ and $X_1, ...X_n$ are conditionally iid given θ, then $X_1, ...X_n$ are exchangeable.
- Proof:

$$\begin{aligned} \pi(x_1, ...x_n) &= \int \pi(x_1, ...x_n|\theta)\pi(\theta)d\theta && \text{(def'n of } m(x)) \\ &= \int \left[\prod_{i=1}^{n} \pi(x_i|\theta)\right]\pi\theta d\theta && (X_1, ..., X_n \text{ is cond iid}) \\ &= \int \left[\prod_{i=1}^{n} \pi(x_{\sigma(i)}|\theta)\right]\pi\theta d\theta && \text{(product does not depend on order)} \\ &= \pi(x_{\sigma(1)}, ..., x_{\sigma(n)}) && \text{(def of } m(x)) \end{aligned}$$

- But what about the other way around? if $X_1, ...X_n$ are exchangeable, does it follow that we can write our model as:

$$\pi(x_1, ...x_n) = \int \pi(x_1, ...x_n|\theta)\pi(\theta)d\theta$$

DeFinetti's Theorem

- Let $X_i \in \mathcal{X}$ for all $i \in \{1, 2, ...\}$. Suppose that, for any n, our belief model for $X_1, ..., X_n$ is exchangeable:

  $$\pi(x_1, ...x_n) = \pi(x_{\sigma(1)}...x_{\sigma(n)})$$

  for all permutations σ of $\{1, ...n\}$.
- Then our model can be written as:

  $$\pi(x_1, ...x_n) = \int \left[\prod_{i=1}^{n} \pi(x_i|\theta) \right] \pi(\theta) d\theta$$

- This θ, whose existence is guaranteed for exchangeable sequences, can be interpreted as the Bayesian prior. But notice that it is not imposed into the problem. Its existence follows from a symmetry assumption weaker than iid.

Problem

A Curious Coin

You have come across a curious coin. It seems (you suspect) bent in a way that biases it toward landing on heads. You will give this coin to your trusty RA, and ask them to perform an experiment (i.e., toss it repeatedly) in order to help you decide whether the coin is biased.

Some Data

A Curious Coin

Your RA reports having tossed the coin 12 times, with the following results:

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

Inference

- This is an instance of a problem about inference for a proportion, $\theta \in [0, 1]$.
- Some questions you might want to ask include:

  What is a good point estimate of θ?

  How confident am I that θ is in some range $[a, b]$?

  Can I reject the hypothesis that $\theta \leq 0.5$?
- Before you can answer these questions you have to ask:

  What is the prior distribution of θ?

  What is the full joint probability model?

  What is the posterior distribution for θ?

Bayesian Approach to the Coin Problem

- Let $X = 1$ denote the coin landing on heads.
- In our problem, we know that the likelihood of $\boldsymbol{X}$ given θ is Bernoulli in θ:

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} [\theta^{x_i}(1-\theta)^{1-x_i}] = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

- Or: Since $X \sim \text{Bernoulli}(\theta)$, we know that $\sum_{i=1}^{n} X_i$ is Binomial in (n, θ):

$$f(x|\theta, n) = \binom{n}{\sum x_i}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

- Now we need to identify a prior distribution for θ.
- A flexible prior distribution for the unknown parameter of a Bernoulli/Binomial process is a beta distribution with parameters α and β:

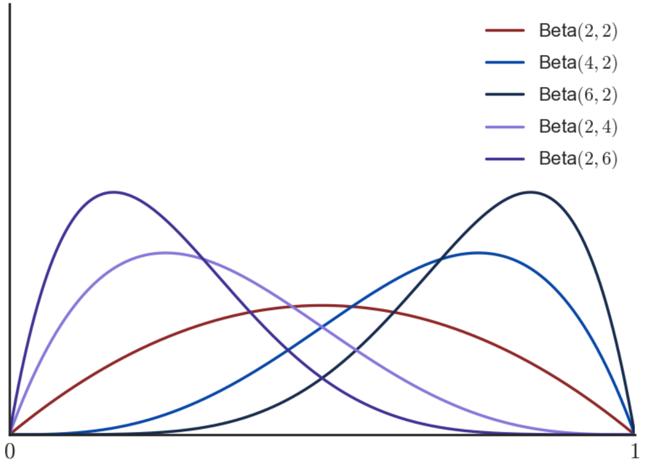
$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  where $\Gamma(x)$ is the complete Gamma function, $\int_0^\infty t^{x-1}e^{-t}dt$ and for positive integers n, $\Gamma(n) = (n-1)!$.
- Note that

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Note also that $\Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ is not a function of θ. It is the normalizing constant for this distribution. Often we can ignore it and renormalize after updating.

Bayesian Approach to the Coin Problem



Bayesian Approach to the Coin Problem

- The posterior distribution for θ is

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto f_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\pi(\theta)\\ &= \prod_{i=1}^{n}\{\theta^{x_i}(1-\theta)^{1-x_i}\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \prod_{i=1}^{n}\{\theta^{x_i}(1-\theta)^{1-x_i}\}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{\sum_{i=1}^{n} x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^{n} x_i+\beta-1}\end{aligned}$$

- Using the fact from the previous slide, we know that

$$\begin{aligned}&\int_0^1 \theta^{\sum_{i=1}^{n} x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^{n} x_i+\beta-1}d\theta\\ &= \frac{\Gamma(\sum_{i=1}^{n} x_i+\alpha)\Gamma(n-\sum_{i=1}^{n} x_i+\beta)}{\Gamma\big(\big[\sum_{i=1}^{n} x_i+\alpha\big]+\big[n-\sum_{i=1}^{n} x_i+\beta\big]\big)}\end{aligned}$$

Bayesian Approach to the Coin Problem

- Let $\alpha^* = \sum_{i=1}^n x_i + \alpha$. Let $\beta^* = n - \sum_{i=1}^n x_i + \beta$. Then our posterior distribution for θ is
$$\pi(\theta|\boldsymbol{x}) = \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$$
- In other words, the posterior distribution is still of the beta form, except that our new α corresponds to the initial α plus the number of successes/heads and the new β corresponds to the initial β plus the number of failures/tails.
- This lends itself to a natural interpretation: The initial α value corresponds to the number of pseudo tosses that came up heads, whereas the initial β value corresponds to the number of pseudo tosses that came up tails.
- Bayesian updating is easily accomplished by adding the pseudo heads to the observed heads and pseudo tails to observed tails.

Bayesian Approach to the Coin Problem

Conjugate family

Let $\mathcal{F}$ denote the class of distributions for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$ and all priors $\pi \in \Pi$, and all $x \in \mathcal{X}$.

- What we have seen so far is that the beta distribution is conjugate to the Bernoulli process. This is sometimes called the “beta-binomial” family.
- In the classes to follow, we will look at other commonly used conjugate families of distributions.

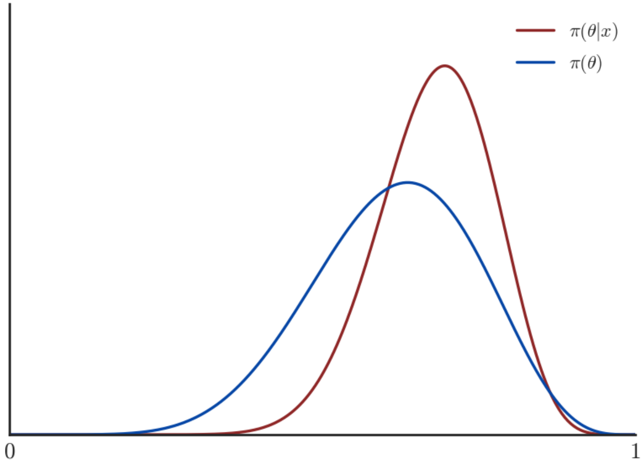
Some properties of the Beta distribution

- Let $X \sim \text{Binomial}(n, \theta)$
- Prior mean: $\alpha/(\alpha+\beta)$
- Posterior mean: $(\alpha+x)/(\alpha+\beta+n)$
- Prior variance:
$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
- Posterior variance:
$$\frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$
- Prior mode: $(\alpha-1)/(\alpha+\beta-2)$
- Posterior mode: ?
- Note that the posterior mean is a compromise between prior information and data:
$$\text{E}[\theta|x] = \frac{\alpha+x}{\alpha+\beta+n} = \frac{n}{\alpha+\beta+n}\left(\frac{x}{n}\right) + \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)$$
- This is: p(MLE) + 1 - p (prior mean) where p corresponds to the weight of the data and 1-p corresponds to the weight of the prior (or “pseudo” data).

Bayesian Approach to the Coin Problem

- Now we can tackle our problem.
- The coin looked bent in a way that made it biased toward heads (i.e., $\theta > 0.5$).
- What are reasonable values for α and β?
- Suppose $\alpha = 8$ and $\beta = 5$.
- The posterior distribution is beta with $\alpha = 9 + 8 = 17$ and $\beta = 3 + 5 = 8$.

Posterior Distribution



Inferences on θ

- Any statements that we wish to make about θ can be easily computed from the posterior distribution.
- The posterior distribution describes all our beliefs about θ after viewing the data.
- For example, we way want to make a point estimate using the posterior mean.

  This is given by $\alpha/(\alpha+\beta)$.

  Before seeing the data, this was $8/(8+5) \approx 0.61$.

  After seeing the data, this is $17/(17+8) \approx 0.68$.

  Note that the sample mean is 0.75. The data has nudged our prior toward a stronger belief in the coin's bias toward heads.

  But is is not as strong as the MLE because (remember) the posterior mean is a weighted average of the MLE and the prior mean.
- We may also want the mode, which is the value we think most likely. This is $(\alpha-1)/(\alpha+\beta-2) = (17-1)/(17+8-2) \approx 0.69$.

Bayesian Hypothesis Tests

- Recall that what we really wanted to know was a simple question: is the coin biased toward heads?
- Now we can answer it directly:

$$\begin{aligned}\Pr(\theta > 0.5) &= \int_{0.5}^{1} \pi(\theta|\boldsymbol{x})d\theta \\ &= 1 - CDF(\theta|\boldsymbol{x})|_{\theta=0.5} \\ &= 1 - 0.03 \\ &= 0.97\end{aligned}$$

- R code: `1 - pbeta(0.5, 17, 8)`
- We are 97% confident that the coin is biased toward heads.
- We now have an answer to a one-sided hypothesis test:

$$H_0 : \theta \leq 0.5 \qquad\qquad H_1 : \theta > 0.5$$

- But instead of accepting/rejecting the null hypothesis, we make probabilistic statements about the research hypothesis from the posterior distribution.

Bayesian Two-Sided Hypothesis Tests

- But what if we want to know whether the coin is fair or not? That is,

$$H_0 : \theta = 0.5 \qquad\qquad H_1 : \theta \neq 0.5$$

- On the picture developed so far, we cannot do this.
- The probability that θ takes on any specific value is 0. Thus the posterior probability for any such H_0 will be 0.
- We will see how to make binary decisions in the Bayesian framework once we introduce the notions of loss and Bayes risk.

Credible Intervals

- However, we can calculate a $(1 - \alpha)100\%$ credible interval for θ. For example, a 95% credible interval for θ is,

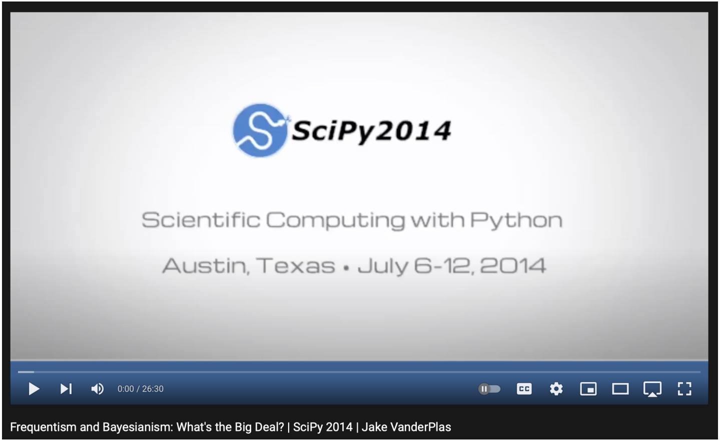
$$\Pr(a < \theta < b) = \int_a^b \pi(\theta|\boldsymbol{x}) d\theta = 0.95$$

- In our case, $a = 0.49$ and $b = 0.84$.
- R code: qbeta(c(0.025,0.975),17,8)
- We can also compute the probability that θ is in any desired region of the posterior distribution. This gives us a probabilistic statement about a small region around a point null hypothesis. For example:

$$\begin{aligned}\Pr(0.4 < \theta < 0.6) &= \int_{0.4}^{0.6} \pi(\theta|\boldsymbol{x}) d\theta \\ &= CDF(\theta|\boldsymbol{x})|_{\theta=0.6} - CDF(\theta|\boldsymbol{x})|_{\theta=0.4} \\ &= 0.19\end{aligned}$$

- R code: pbeta(0.6, 17, 8) - pbeta(0.4, 17, 8).
- We are about 20% confident that θ is between 0.4 and 0.6.
- Compare this to the confidence interval from Class 1A. This is now a probabilistic statement about θ, treated as a random quantity. And not a probabilistic statement about X and the proportion of cases in which it will cover θ if sampled repeatedly!

Video


SciPy2014
Scientific Computing with Python
Austin, Texas • July 6-12, 2014
0:00 / 26:30
CC

Frequentism and Bayesianism: What's the Big Deal? | SciPy 2014 | Jake VanderPlas

https://youtu.be/KhAUfqhLakw?t=1231

Practice

- Estimate the chance θ of recidivism based on a study in which there were $n = 43$ individuals released from jail and $x = 15$ committed another crime within 36 months of release.
- Using a $\text{Beta}(2, 8)$ prior for θ, find the posterior mean, standard deviation, and 95% credible interval
- Recall: $\int_a^b p(\theta|x)d\theta = x$ is the $(x \times 100)$% credible interval.
- $\theta|x = 15 \sim \text{Beta}(17, 36)$
- $\text{E}[\theta|x = 15] = 17/53 = 0.32$
- $\text{Sd}(\theta|x = 15) = [(17 \times 36)/(53^2 \times 54)]^{1/2} = 0.06$
- $\text{CI}(\theta|x = 15) = (0.20, 0.45)$
- R code for CI: qbeta(c(0.025,0.975),17,36)