

STA 360/602L: MODULE 2.1

CONJUGACY; BETA-BERNOULLI AND BETA-BINOMIAL MODELS

DR. OLANREWAJU MICHAEL AKANDE



OUTLINE

- Conjugacy
- Kernels
- Bernoulli data
- Binomial data

BAYESIAN INFERENCE

- Once again, given data y and an unknown population parameter θ , estimate θ .
- As a Bayesian, you update some prior information for θ with information in the data y , to obtain the posterior density $p(\theta|y)$.
- Personally, I prefer being able to obtain posterior densities that describe my parameter, instead of estimated summaries (usually measures of central tendency) along with confidence intervals.
- Bayes' theorem - reminder:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int_{\Theta} p(\tilde{\theta})p(y|\tilde{\theta})d\tilde{\theta}} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

COMMENTS ON THE POSTERIOR DENSITY

- The posterior density is more concentrated than the prior & quantifies learning about θ .
- In fact, this is the optimal way to learn from data - see discussion in Hoff chapter 1.
- As more & more data become available, posterior density will converge to a normal (Gaussian) density centered on the MLE (Bayes central limit theorem).
- In finite samples for limited data, the posterior can be highly skewed & noticeably non-Gaussian.

CONJUGACY

- Starting with an arbitrary prior density $p(\theta)$ & sampling density $p(y|\theta)$ we may encounter problems in calculating the posterior density $p(\theta|y)$.
- In particular, you may notice the denominator in the Bayes' rule:

$$p(y) = \int_{\Theta} p(\theta)p(y|\theta)d\theta.$$

This integral may not be analytically tractable!

- When the prior is **conjugate** however, the marginal likelihood can be calculated analytically.
- **Conjugacy** \Rightarrow the posterior density (or mass) function has the same form as the prior density (or mass) function.
- Conjugate priors make calculations easy but may not represent our prior information well.

KERNELS

- In Bayesian statistics, the **kernel** of a pdf or pmf omits any multipliers that do not depend on the random variable or parameter we care about.
- For many distributions, the kernel is in a simple form but the normalizing constant complicates calculations.
- If one recognizes the kernel as that matching a known distribution, then the normalizing factor can be reinstated. This is a very MAJOR TRICK we will use to calculate posterior distributions.
- For example, the normal density is given by

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

but the kernel is just

$$p(y|\mu, \sigma^2) \propto e^{-\frac{(y-\mu)^2}{2\sigma^2}} .$$

BERNOULLI DATA

- Back to our example: suppose $\theta \in (0, 1)$ is the population proportion of individuals with diabetes in the US.
- Suppose we take a sample of n individuals and record whether or not they have diabetes (as binary: 0,1).
- Then we can use the Bernoulli distribution as the sampling distribution.
- Also, we already established that we can use a beta prior for θ .

BERNOULLI DATA

- Generally, it turns out that if
 - $p(y_i|\theta) : y_i \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ for $i = 1, \dots, n$, and
 - $\pi(\theta) : \theta \sim \text{Beta}(a, b)$,

then the posterior distribution is also a beta distribution.

- Can we derive the posterior distribution and its parameters? Let's do some work on the board!
- Updating a beta prior with a Bernoulli likelihood leads to a beta posterior - we have conjugacy!
- Let $y = (y_1, \dots, y_n)$. Specifically, we have.

$$p(\theta|y) = \text{Beta} \left(a + \sum_{i=1}^n y_i, b + n - \sum_{i=1}^n y_i \right).$$

- This is the **beta-Bernoulli model**. More generally, this is actually the **beta-binomial model**.

BETA-BINOMIAL IN MORE DETAIL

- Suppose the sampling density of the data is

$$p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}.$$

- Suppose also that we have a $\text{Beta}(a, b)$ prior on the probability θ .
- Then the posterior density then has the beta form

$$\pi(\theta|y) = \text{Beta}(a + y, b + n - y).$$

- The posterior has expectation

$$\mathbb{E}(\theta|y) = \frac{a + y}{a + b + n} = \frac{a + b}{a + b + n} \times \text{prior mean} + \frac{n}{a + b + n} \times \text{sample mean}.$$

- For this specification, **sometimes a and b are interpreted as "prior data" with a interpreted as the prior number of 1's, b as the prior number of 0's, and $a + b$ as the prior sample size.**
- As we get more and more data, the majority of our information about θ comes from the data as opposed to the prior.

BINOMIAL DATA

- For example, suppose you want to find the Bayesian estimate of the probability θ that a coin comes up heads.
- Before you see the data, you express your uncertainty about θ through the prior $p(\theta) = \text{Beta}(2, 2)$
- Now suppose you observe 10 tosses, of which only 1 was heads.
- Then, the posterior density $p(\theta | y)$ is $\text{Beta}(3, 11)$.

BINOMIAL DATA

- Recall that the mean of $\text{Beta}(a, b)$ is $\frac{a}{a+b}$.
- So, before you saw the data, you thought the mean for θ was $\frac{2}{2+2} = 0.50$.
- However, after seeing the data, you believe it is $\frac{3}{3+11} = 0.21$.
- The variance of $\text{Beta}(a, b)$ is $\frac{ab}{(a+b)^2(a+b+1)}$.
- So before you saw data, your uncertainty about θ , in terms of the standard deviation, was $\sqrt{\frac{4}{4^2 \times 5}} = 0.22$.
- However, after seeing 1 Heads in 10 tosses, your standard deviation gets updated to $\sqrt{\frac{33}{14^2 \times 15}} = 0.11$.
- Clearly, as the number of tosses goes to infinity, your uncertainty goes to zero.

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!