GENOME 541 Section 2

Lecture 1

Introduction to Bayesian statistics

Yi Yin (yy224[at]uw.edu)
Shendure Lab
Department of Genome Sciences
University of Washington
Tuesday, April 11, 2017

Overview

- Based loosely on Peter Hoff's textbook: A First Course in Bayesian Statistical Methods
    1. Introduction to Bayesian statistics (Ch1 & 3)
    2. Gibbs sampling (Ch6)
    3. Metropolis-Hastings (Ch10)
    4. Bayesian linear models (Ch9)
- Most class notes will be presented on the board, take copious notes will help with HW.

Overview

- HW 3 (Apr. 12-Apr. 20) & HW 4 (Apr. 19-Apr. 27). LaTeX template will be provided but hand-written (clear) for derivation is OK. Generally, derive some posterior distribution, posterior predictive distribution, truncated distribution ... and simulate from it

- Familiarity of R is assumed (examples will be provided)

- Emphasis on Bayesian inference : is my coin fair? (example on board, see post-lecture note Page 1)

How do you go from $p(Y|\pi)$ to $p(\pi|Y)$?

Event A and B have probability p(A) and p(B)

$$p(A, B) = p(A)p(B|A) = p(B)p(A|B)$$
$$p(\pi)p(Y|\pi) = p(Y)p(\pi|Y)$$

$$p(\pi|Y) = \frac{p(\pi)p(Y|\pi)}{p(Y)}$$

$p(\pi)$: prior probability of π
$p(Y|\pi)$: conditional probability of event/data given parameter π
$p(\pi|Y)$: posterior probability of π after observing some data Y
$p(Y)$: marginal probability of event/data Y

$$p(\pi|Y) = \frac{p(\pi)p(Y|\pi)}{p(Y)}$$

Provides a mechanism of learning from the data Y

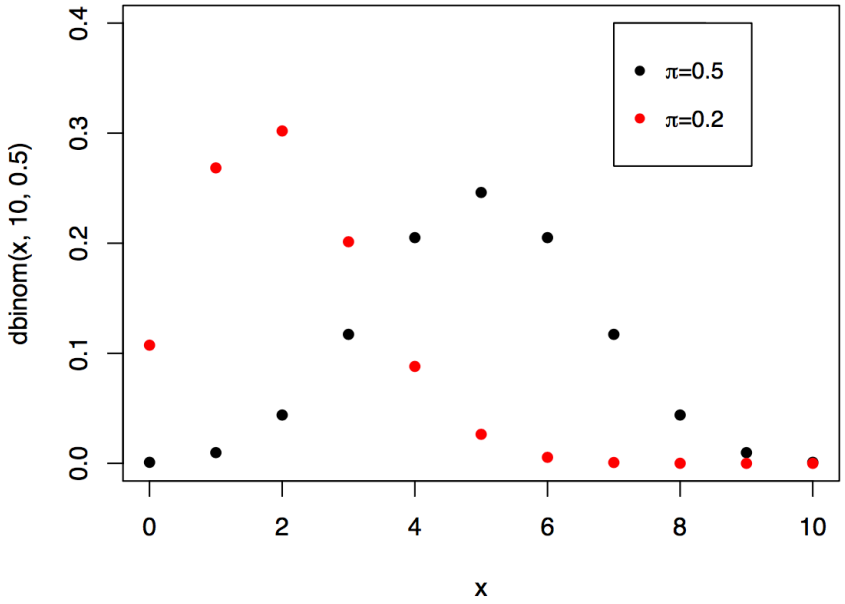
1. Prior distribution
   (for parameters in the model)

Apply Bayes' rule → Posterior Distribution

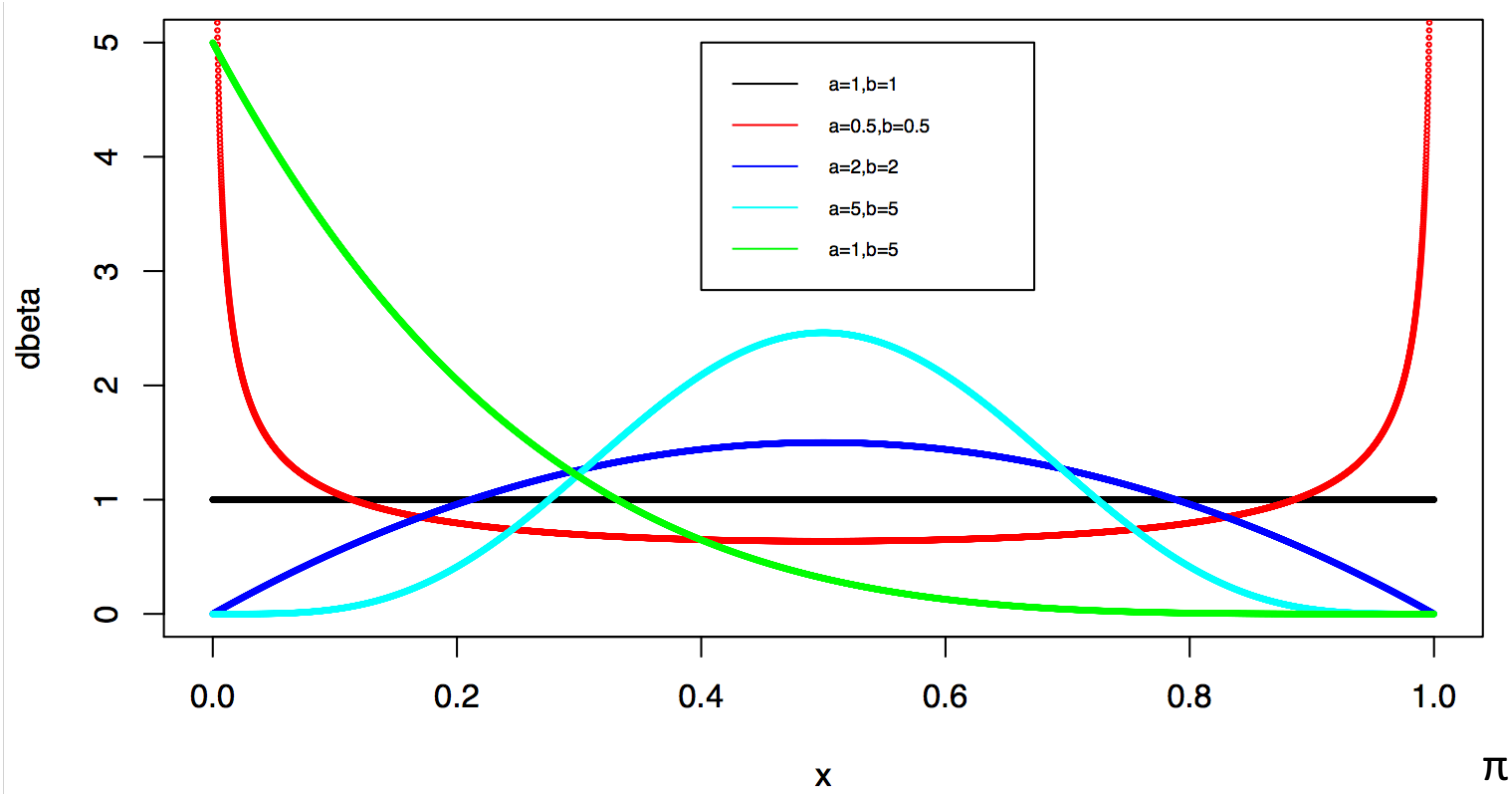
2. Likelihood function
   (for the data)

"conjugacy" in simple models: prior and posterior of the same form

Most often involves unknown normalizing constant (is it necessary to calculate this? See example on board)



Binomial (10, π)



π ~ Beta(a,b)

beta(a,b) is a probability density function on (0,1),

$$p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1},$$

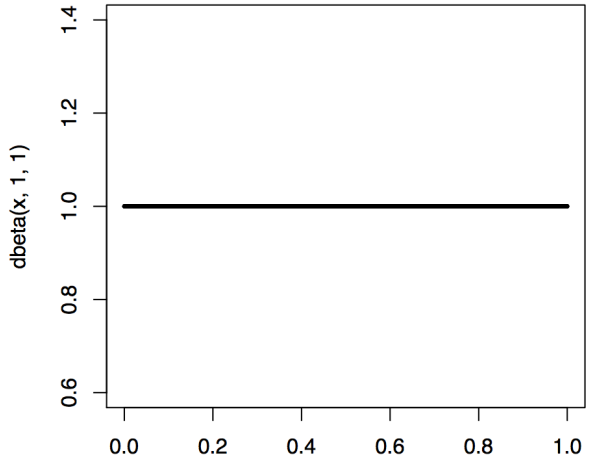
where $B(a,b)$=beta function = normalizing constant ensuring integration to one

beta(a,b) has expectation $E(\theta) = a/(a+b)$ & density more concentrated as $a+b$=prior 'sample size' increases

In R:

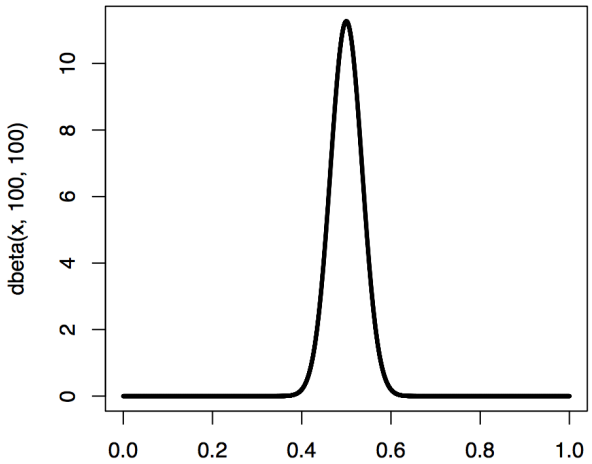
dbeta(): pdf for a point value or a vector of values
pbeta(): cdf
qbeta(): inverse cdf
rbeta(): draw random samples from a Beta distribution

"kernel" of a distribution



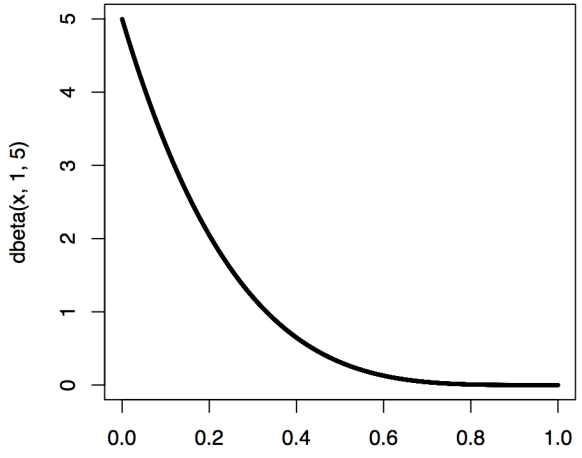
$\pi_1 \sim$ Beta (1, 1)

$E[\pi_1] = 0.5$



$\pi_2 \sim$ Beta(100,100)

$E[\pi_2] = 0.5$

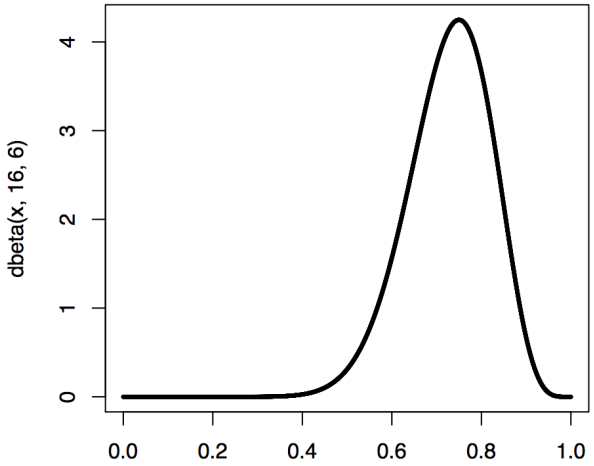


$\pi_3 \sim$ Beta(1,5)

$E[\pi_3] = 0.167$

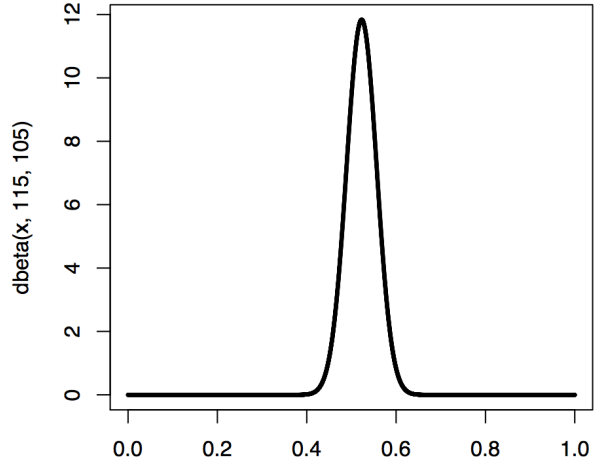
Now, we did some experiment and observed some data:

n=20, k=15; (15 head, 5 tail)



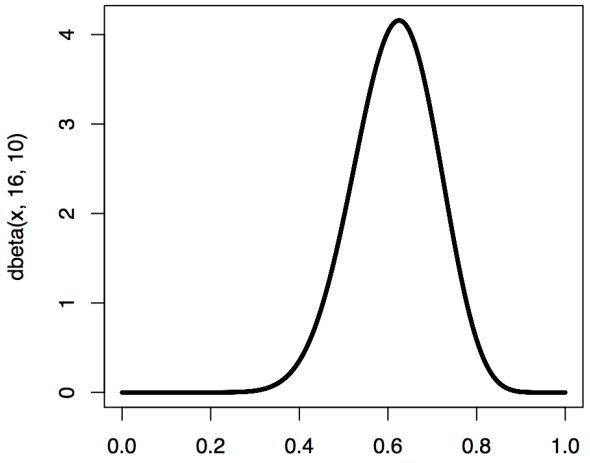
$\pi_1 | Y \sim$ Beta (16, 6)

$E[\pi_1|Y] = 0.73$



$\pi_2 | Y \sim$ Beta(115,105)

$E[\pi_2|Y] = 0.52$



$\pi_3 | Y \sim$ Beta(16,10)

$E[\pi_3|Y] = 0.615$

Let's look at the posterior mean:

Under $\pi_1 \sim$ Beta (1, 1), after observing k=15, (n-k)=5, we get:
$\pi_1|Y \sim$ Beta (16, 6), with $E[\pi_1|Y]=16/(16+6)$.

Under $\pi_1 \sim$ Beta (a, b), after observing k heads, (n-k) tails, we get:
$\pi_1|Y \sim$ Beta (a+k, b+(n-k)), with $E[\pi_1|Y]=(a+k)/(a+k+b+(n-k))=(a+k)/(a+b+n)$

$$E[\pi_1|Y] = \frac{a+k}{a+b+n} = \frac{a}{a+b} \times \frac{a+b}{a+b+n} + \frac{k}{n} \times \frac{n}{a+b+n}$$

a+b: prior sample size (2, 200 and 6 in the last example)
n: sample size in data

Posterior mean is the weighted average of prior mean and sample mean.

Always good to have enough data points to out-weight your prior.

1. Prior specification (will cover examples on the impact of eliciting different priors)(Lindley paradox…)

2. Choice of likelihood function (same as in frequentist analyses, not covered here, but will usually be provided in HWs)

3. Derive and/or simulate from posterior (focus) (Posterior needs to be a **proper** distribution!!) (HW2)

4. Inferences (based on posterior and loss function)

1. Any questions?

2. HW1: Poisson distribution with Gamma prior (understand the relationship between Poisson distribution and Negative Binomial distribution)

3. HW1 (ungraded question): think about and describe in words your ideas on how to sample from a truncated distribution (We will discuss this at the beginning of the 4^{th} lecture and in HW2, you will be asked to implement your simulation)

GENOME 541 Section 2 Lecture 1 — Introduction to Bayesian statistics

Yi Yin
Jay Shendure Lab, Department of Genome Science, University of Washington
yy224@uw.edu

1 Example 1 (Part 1): Introducing the coin tossing example

Parameters to summarize quantity of interest, for example, human heights, population mean, variance, etc.

parameter space: μ, σ^2, θ, π ... $\in \Theta$

Do some experiments or surveys to collect some data on samples.

sample space: $Y = \{y_1, y_2, ..., y_n\} \in \mathcal{Y}$

All the possible samples constitute a sample space.

Fair coin example:

$H_0 : p(H) = p(T) = 0.5$ (null hypothesis)

data: $Y = \{H, H, T, T, H\}$

$p(3H, 2T|p(H) = 0.5) = \begin{pmatrix} 5 \\ 3 \end{pmatrix} (0.5)^3(1 - 0.5)^2 = 0.3125$

If instead you toss the coin 10 times, get 9H, 1T

$p(9H, 1T|p(H) = 0.5) = \begin{pmatrix} 10 \\ 9 \end{pmatrix} (0.5)^9(1 - 0.5)^1 = 0.0098$, suggesting that you might want to throw away your null hypothesis that the coin is fair.

If we get 9H 1T, instead of saying this is unlikely if my coin is fair, can we update the parameter of interest $\pi = p(H) = 0.5$ to some bigger value and learn about p(π|data) directly?

We can formalize this a bit: before any experiment or before observing any data, we have $\pi = p(H) = 0.5$. We then observed some data under the binomial sampling scheme $p(Y|\pi) = \begin{pmatrix} 10 \\ 9 \end{pmatrix} (0.5)^9(1 - 0.5)^1 = 0.0098$, what can we say about $p(\pi|Y)$?

Now the question becomes: How do you go from p(Y|π) to p(π|Y)?

See Slide 4 and 5.

In this specific example, we have:

$p(\pi|Y) = \frac{p(\pi)p(Y|\pi)}{p(Y)} = \frac{p(\pi)p(Y|\pi)}{\int_0^1 p(Y|\pi)p(\pi)d\pi}$

This denominator looks complicated... we will leave it for now and first work on an example where this part is easy to calculate.

2 Example 2: Disease given test positive example

Suppose that you had a test for certain disease, like a CT scan, you would like to know what is the probability of having the disease, if the test is positive (since not all tests are perfect). We can write:

$p(D|T)$: p(having a disease | test is positive)

Applying Bayes rule on Slide 4 and 5, we have:

$p(D|T) = \frac{p(D)p(T|D)}{p(T)} = \frac{p(D)p(T|D)}{p(D)p(T|D)+p(noD)p(T|noD)}$

Note that this equation is a bit different from what I wrote on the board. (Thanks Anthony for pointing it out.)

This is showing how we can further factorize T (test positive) with respect to D (having disease and no disease). You can think of p(T) as having two scenarios, p(T|D) and p(T|no D), which is true positive rate associated with the test and false positive rate associated with the test, respectively. This information could be known from experience, or from the development stage of this specific test. p(D) could be the prior knowledge of this particular disease, which could be determined by how prevalent the disease is, your age, any number of factors that could affect the judgement of the physician. Intuitively, this makes sense such that whether you have the disease given test positive is dependent on the prevalence of the disease in the population as well as how accurate the test is.

3 Example 1 (Part 2): Deriving posterior distribution of the coin tossing example - Beta-Binomial conjugacy example

Now lets go back to the coin tossing example. Unlike the disease example above, its not so easy to calculate the normalizing constant. But heres a convenient fact we can use for Binomial distribution, that is, if you pick a prior distribution that has a form of the Beta distribution, your posterior distribution will also be a Beta distribution. It sounds like an arbitrary choice here but it is not a bad choice. We will first work through to see how picking a Beta prior helps.

Some basic intro about Beta distribution and Binomial distribution is on Slide 6 and 7. But below are a few equations that may help you reading through the slides.

Beta function: $\int_0^1 \pi^{a-1}(1-\pi)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Note that the notation $\Gamma(a) = (a-1)!$, but you do not have to know this...

Beta distribution has two hyper-parameters a and b. We can write $\pi \sim$ Beta(a,b). The expectation of π is a/(a+b). So if we want to pick a π that is centered on 0.5, we can pick Beta(1,1), or Beta (10, 10), as long as a/(a+b) is 0.5. You can see how these different choices of prior distribution affect our posterior inference on Slide 8 (Slide 9 provides an intuitive explanation of the relationship between prior sample size, sample size and posterior mean as a weighted average).

prior distribution:

$p(\pi) = \frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1}$

Likelihood function:

$p(Y = k|\pi) = \left(\begin{array}{c} n \\ k \end{array} \right) \pi^k(1-\pi)^{n-k}$

Note the resemblance between the kernels in the prior distribution and the likelihood (you can ignore all the constants not involving π, for sample B(a,b), and $\left(\begin{array}{c} n \\ k \end{array} \right)$.

Posterior distribution:

$p(\pi|Y = k) = \frac{p(\pi)p(Y=k|\pi)}{p(Y=k)} \propto \pi^{a-1}(1-\pi)^{b-1} \times \pi^k(1-\pi)^{n-k} = \pi^{a+k-1}(1-\pi)^{b+(n-k)-1}$

Therefore, $p(\pi|Y = k) \sim Beta(a+k, b+(n-k))$