Lecture 2 Notes: Simple Bernoulli Modeling

January 23, 2017

Today we will begin to build our first model. As a motivating example, we will borrow from the example in section 2.2 of McElreath (2016)'s *Statistical Rethinking* book using globe tosses. This is because I sincerely hate coin flip examples.

We have a globe representing the planet Earth. The globe is itself a model of the actual planet that we live on.

Let's say we are interested in the proportion of the surface covered in water. Therefore we have N people toss the globe in the air. Upon catching it, pick a finger and record whether or not that finger is touching water ($y = 1$) or land ($y = 0$).

For example, suppose we have 5 people toss the globe and we get

W L W W L

where W is water and L is land.

```
N <- 5 # number of globe tosses
obs <- c("W", "L", "W", "W", "L") # observed data
y <- ifelse(obs == "W", 1, 0) # convert to 1's and 0's
```

1 The Bernoulli likelihood function

We now have a data story that we need to translate into a formal probability model by restating the data story as a sampling process:

1. The true proportion of water covering the globe is θ.
2. A single toss has a probability θ of producing a water (W) observation and probability $(1 - \theta)$ of producing a land (L) observation.
3. Each toss is independent of the others.

Based on these descriptions of the sampling process, we can describe the probability of each outcome using the Bernoulli distribution:

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

Here each data value y is fixed by an observation and θ is a continuous variable.

Equation 1 specifies the probability of fixed y's as a function of candidate values of θ, and different values of θ yield different probabilities of y.

- Eq. 1 is the *likelihood function* of θ.

As mentioned above, we assume that each globe toss y_i is independent. For a set of outcomes Y, the probability of the set is the multiplicative product of the individual outcome probabilities.

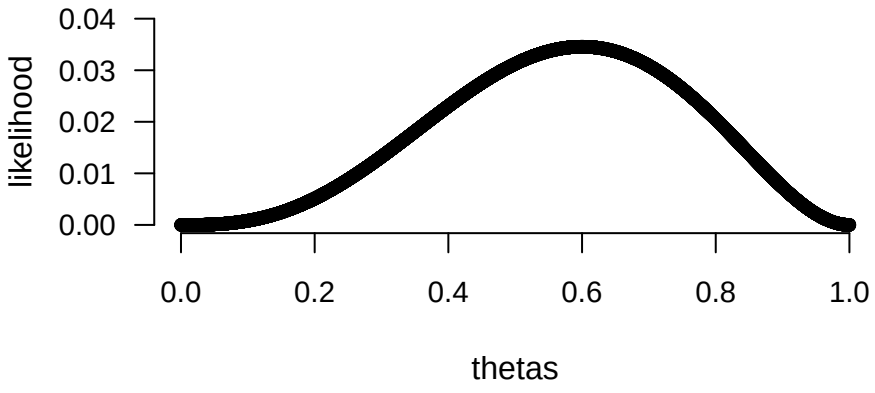
- If we denote the number of W's as $z = \sum_i y_i$ and the number of L's as $N - z = \sum_i (1 - y_i)$, then

$$\begin{aligned} p(Y|\theta) &= \prod_i p(y_i|\theta) \\ &= \theta^z (1-\theta)^{N-z}. \end{aligned} \tag{2}$$

Now that we have our likelihood function, we can calculate the likelihoods of our fixed data for a range of θ values from 0–1.

```
thetas <- seq(0, 1, length=10000) # sequence of candidate theta values

# vector of Likelihoods for the fixed data for each candidate theta
liks <- numeric(length = length(thetas))

for (i in 1:length(thetas)) {
  temp <- dbinom(y, 1, thetas[i])
  liks[i] <- prod(temp)
}
plot(thetas, liks, pch=1, xlab="thetas", ylab="likelihood",
  las=1, cex=0.8, cex.axis=0.9, ylim=c(0,0.04), frame.plot=FALSE)
```

We can then plot out the probabilities of the data as a function of each of those candidate θ values:



We can estimate the maximum likelihood using R:

```
maxLik <- thetas[liks == max(liks)]
```

or calculate it empirically as $z/N \approx 0.6$.

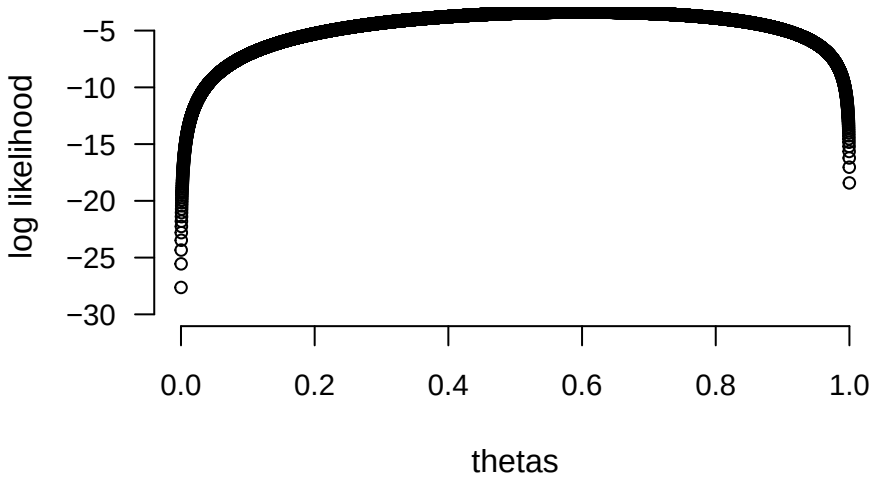
In almost all cases, it is easier and more computationally efficient to work with likelihoods on the log scale.

- Log-likelihoods are added rather than multiplied
- Prevents numerical underflow (floating-point processor errors as the values get REALLY close to zero)
-

Under the hood, the Hamiltonian monte carlo (HMC) sampler that Stan uses works on log-probability gradients.

We can easily calculate the likelihood of our data given a range of thetas on a log scale:

```
  logLiks <- numeric(length = length(thetas)) #vector of log-likelihoods

for (i in 1:length(thetas)) {
  temp <- dbinom(y, 1, thetas[i], log=TRUE)
  logLiks[i] <- sum(temp)
}
plot(thetas, logLiks, pch=1, xlab="thetas", ylab="log likelihood",
  las=1, cex=0.8, cex.axis=0.9, frame.plot=FALSE, ylim=c(-30,-4))
```



This "flatter" surface makes it much easier to explore parameter space relative as compared to raw likelihoods.

2 Specifying a prior

One of the fundamental differences of Bayesian vs. frequentist statistics is that parameters (e.g, θ) are themselves random variables.

- i.e., variables whose values are unknown until observed or sampled and are drawn from some underlying probability distribution.

Thus, a key requirement of Bayesian statistics is to define the *prior* distribution that describes the parameters.

But how do we translate prior information about the real world into a mathematical probability distribution? How do we pick a prior?

Historically, *conjugate priors* were selected that played nicely with likelihood functions, so that $p(y|\theta)$ and $p(\theta)$ combine such that the posterior distribution $p(\theta|y)$ has the same functional form as the prior $p(\theta)$.

- For a Bernoulli or binomial distribution, the conjugate prior for θ is a *beta distribution*:

$$\theta \sim \text{Beta}(a, b) \tag{3}$$

Parameterizing a beta distribution

The beta distribution has two parameters, a & b.

I conceptualize the beta distribution using a balloon metaphor:

Imagine a balloon in a box. If you put pressure on the left side, the density inside the balloon shifts to the right.

- This is what happens when the parameter a is increased. The probability density shifts to higher values.
- The same thing happens when b is increased except that the probability density shifts to the left.
- The higher the value of a or b, the harder you are pushing down on the balloon. If both values are high, a lot of the balloon's shape will be concentrated as a tall peak in the middle.

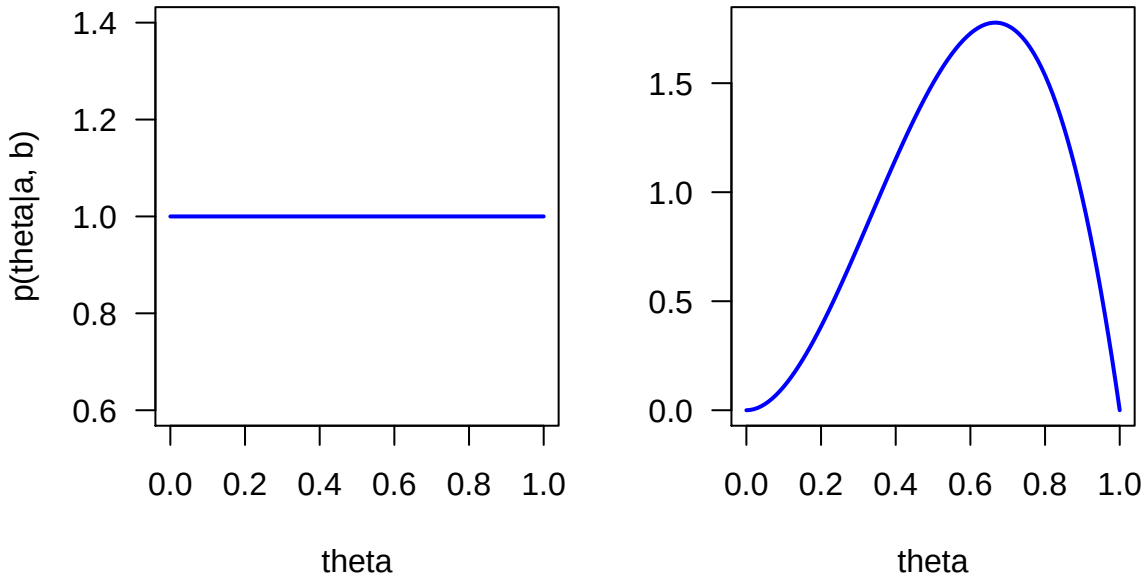
We can also think about a and b in terms of previously observed data, in which there were a W's and b L's in a total of $a + b = N$ globe tosses.

If we had no information about the planet other than the knowledge that the Earth has both land and water, that is equal to observing one W and one L and $\theta \sim \text{Beta}(1, 1)$.

a uniform distribution where all $\theta's$ are equally probable.

Alternatively, say the data we originally collected (W, L, W, W, L) was being used to inform our prior. Then $\theta \sim \text{Beta}(3, 2)$.

```
par(mfrow=c(1,2))
par(mar=c(4,4,0.1,0.5))
# beta(1,1)
curve(dbeta(x, shape1=1, shape2=1),las=1, ylab="p(theta|a, b)",
xlab="theta", col="blue", lwd=2)
# beta(3,2)
curve(dbeta(x, shape1=3, shape2=2),las=1, ylab="",
xlab="theta", col="blue", lwd=2)
```



Another way to parameterize a beta distribution is in terms of central tendency and our confidence in that central tendency. For example, we might think that the earth is 70% covered in water, but are a bit uncertain about that estimate.

- e.g., observing $N = 10$ globes previously

We can think about a beta distribution in terms of its mean (μ), mode (ω), and concentration (κ).

- When $a = b$, the mean and mode are 0.5.
- When $a > b$, the mean and mode are greater than 0.5.
- When $a < b$, the mean and mode are less than 0.5.
- The spread of the beta distribution is related to the *concentration* $\kappa = a + b$.
  - As κ gets larger, the distribution becomes more concentrated.

To parameterize the distribution in terms of the mean,

$$a = \mu\kappa \text{ and } b = (1-\mu)\kappa. \tag{4}$$

To parameterize the distribution in terms of the mode,

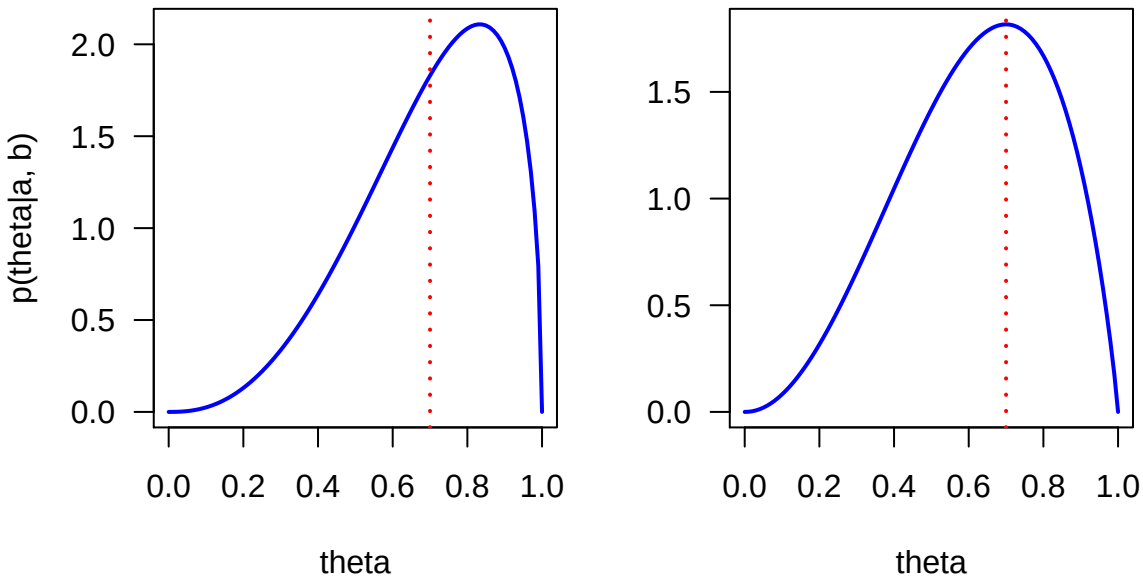
$$a = \omega(\kappa - 2) + 1 \text{ and } b = (1-\omega)(\kappa - 2) + 1 \tag{5}$$

for $\kappa > 2$.

We can think about κ as the amount of information or data needed to change our prior beliefs about μ or ω.

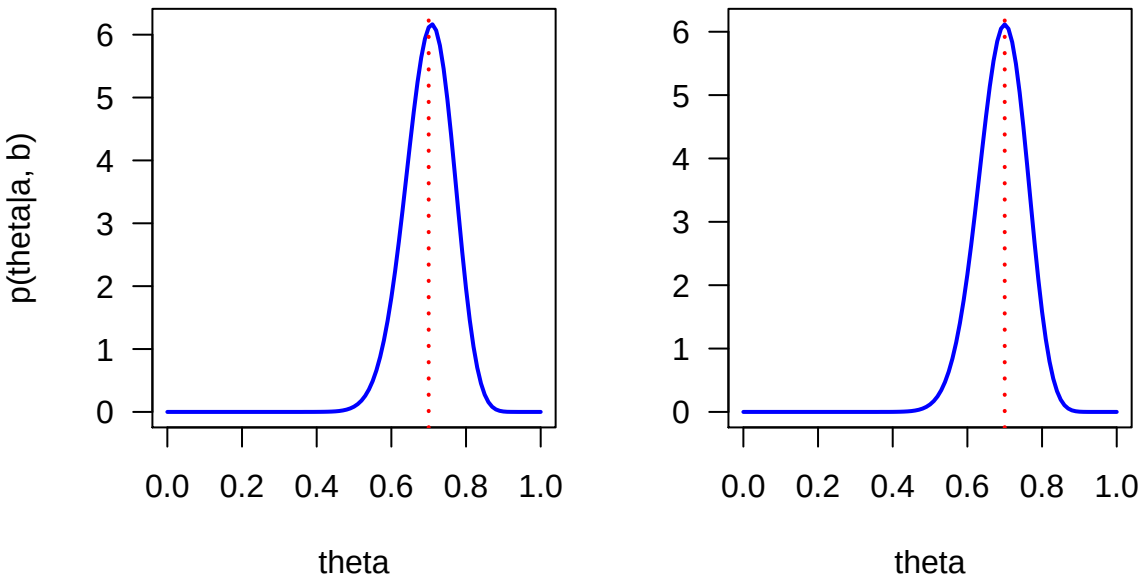
- If we are not very confident in the proportion of water (say 70%) covering the earth, we might only need a few globe flips (e.g., 5) and thus a small κ. For example,

```
par(mfrow=c(1,2))
par(mar=c(4,4,0.1,0.5))

mu <- 0.7 # expected mean proportion of water
omega <- 0.7 # proportion of water as the mode
kappa <- 5 # low confidence in central measures

# plot in terms of mean:
a <- mu * kappa
b <- (1 - mu) * kappa

curve(dbeta(x, shape1=a, shape2=b),las=1, ylab="p(theta|a, b)",
  xlab="theta", col="blue", lwd=2)
abline(v = mu, col="red", lwd=2, lty=3)

# Plotting in terms of mode:
a <- omega * (kappa - 2) + 1
b <- (1 - omega) * (kappa - 2) + 1

curve(dbeta(x, shape1=a, shape2=b),las=1, ylab="",
  xlab="theta", col="blue", lwd=2)
abline(v=omega, col="red", lwd=2, lty=3)
```



Conversely, if we are very confident in that proportion, we might need $\kappa = 50$ or more globe flips.

```
par(mfrow=c(1,2))
par(mar=c(4,4,0.1,0.5))

mu <- 0.7 # expected mean proportion of water
omega <- 0.7 # proportion of water as the mode
kappa <- 50 # low confidence in central measures

# plot in terms of mean:
a <- mu * kappa
b <- (1 - mu) * kappa

curve(dbeta(x, shape1=a, shape2=b),las=1, ylab="p(theta|a, b)",
  xlab="theta", col="blue", lwd=2)
abline(v = mu, col="red", lwd=2, lty=3)

# Plotting in terms of mode:
a <- omega * (kappa - 2) + 1
b <- (1 - omega) * (kappa - 2) + 1

curve(dbeta(x, shape1=a, shape2=b),las=1, ylab="",
  xlab="theta", col="blue", lwd=2)
abline(v=omega, col="red", lwd=2, lty=3)
```



For skewed distributions, the mode ω can be more intuitive. The mode is where the curve reaches it's greatest height, whereas the mean is somewhere away from the mode along the longer tail.

- This is apparent from the dotted line in each of the plots, especially for the first set of plots with low κ.