

Stats 102C - Lecture 1-1: Bayesian Basics

Miles Chen, PhD

Week 1 Monday

Section 1

Lecture 1-1: Bayesian Basics

Review of Frequentist Concepts

In frequentist statistics, probability is defined as the long-run relative frequency of an event.

The probability that coin lands heads is 0.5. This means that after an infinite number of flips, the proportion of times the coin lands heads is 0.5.

Maximum Likelihood estimation and Confidence Intervals are both ideas of frequentist statistics.

Maximum Likelihood Estimation

We begin with a probabilistic model that generates random values.

The model is defined by parameters. In frequentist statistics, the parameters are fixed but unknown values.

For example, let's say we have a box with marbles in it. Some marbles are blue while the rest are not. We want to estimate the proportion θ of marbles that are blue.

Let's say we draw 10 marbles with replacement. 6 are blue and the rest are not.

Based on this data, what is the maximum likelihood estimate for the proportion of marbles that are blue?

We draw 10 marbles with replacement. 6 are blue and the rest are not. Find the MLE for θ .

We begin with the likelihood function of the data we observed. We find the value of θ that will maximize the likelihood.

$$\hat{\theta} = \arg \max_{\theta} L(\theta | \mathbf{x} = 6) = \arg \max_{\theta} \binom{10}{6} \theta^6 (1 - \theta)^4$$

We find the maximum by taking the derivative w.r.t θ and setting it equal to zero.

Finding the maximum

Finding the derivative of a large product can get messy, so we find the log and solve for the maximum.

$$\log L = \log 210 + 6 \log \theta + 4 \log(1 - \theta)$$

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta} (\log 210 + 6 \log \theta + 4 \log(1 - \theta))$$

$$0 = 0 + \frac{6}{\theta} + \frac{4}{(1 - \theta)}(-1)$$

$$\frac{4}{(1 - \theta)} = \frac{6}{\theta}$$

$$4\theta = 6(1 - \theta)$$

$$10\theta = 6$$

$$\theta = 0.6$$

Using the ML Estimate

According to frequentist statistics and MLE, the best estimate of the proportion of blue marbles is 0.6 because we observed 10 marbles total and 6 of them were blue.

We can use this estimate for future estimates.

“What is the probability that if we draw three more marbles with replacement, what is the probability that we get exactly 2 blue marbles?”

$$\text{Answer} = \binom{3}{2} 0.6^2 (1 - 0.6)^1 = 0.432$$

Frequentist reasoning

With frequentist statistics, we can also create a confidence interval.

Let's just pretend that for a 95% CI, the margin of error is 9% points. The CI for θ goes from 0.51 to 0.69.

The interpretation is: "I am 95% confident that the proportion of blue marbles is between 51% and 69%."

In frequentist statistics, we are not allowed to say "There is a 95% probability that the proportion is between .51 and .69."

In frequentist statistics, the proportion of blue marbles is a fixed value and does not vary. The proportion is not subject to randomness, so it does not make sense to talk about probability.

The Bayesian difference

In Bayesian statistics, the unknown parameter is not a fixed value. The unknown parameter is a random variable.

Treating the unknown parameter as a random variable might not make much sense if the example is a box with marbles in it. We can at any time open the box, dump out the marbles, count them up, and find the exact proportion.

Using a random variable does make sense when the unknown parameter is a value subject to random changes. For example, what proportion of people in the United States are under the age of 18? The exact proportion is a random variable subject to people being born, people dying, and people turning 18 each day.

Using a random variable for the unknown parameter can also work if we treat **probability as a subjective belief**. Perhaps we say you are never allowed to look inside the box with marbles. The probability of drawing a blue marble can be thought of as our subjective belief of what proportion inside the box is blue.

Parameters have probability distributions

Because the parameters are treated as random variables, in Bayesian statistics, **parameters have their own probability distributions.**

Questions and calculations that depend on the parameter are now more complicated because the parameter is no longer a single value, but is now associated with a distribution.

Perhaps the Bayesian approach is more reasonable

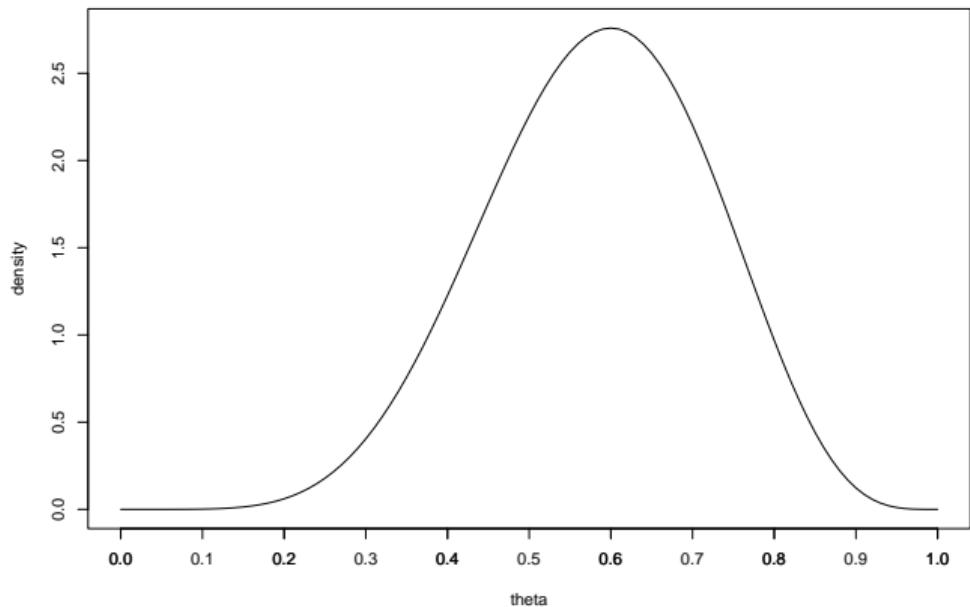
Think back to the box of marbles example. We sampled 10 marbles and 6 were blue.

In the frequentist approach, according the MLE, the proportion of blue marbles in the box is estimated to be 0.6. A confidence interval puts a margin of error around this estimate, but all future calculations will assume that the proportion is 0.6. Maybe it's just me, but I'm a bit uncomfortable about assuming the proportion is this exact value after just 10 observations.

With the Bayesian approach, the proportion of blue marbles in the box will have a probability distribution. After observing our data, the probability distribution will be tallest at 0.6, but there will also be a high probability that the proportion is 0.55 or 0.65. This acknowledges the data we observed, but also says the proportion could very well be other values.

The next slide may show what the distribution might look like.

Distribution of theta after observing 6 blue in 10 marbles



The Bayesian approach is more complicated

The trade off is that calculations are now a lot more complicated.

“What is the probability that if we draw three more marbles with replacement, what is the probability that we get exactly 2 blue marbles?”

For the Frequentist answer, we just plugged in 0.6 for θ : $\binom{3}{2}0.6^2(1 - 0.6)^1 = 0.432$

For the Bayesian answer, we have to consider all the values that θ could possibly be and weight the result by the probability that θ is that value.

Section 2

The Beta-Binomial Model

A Baseball example

Adapted from: http://varianceexplained.org/statistics/beta_distribution_and_baseball/

In baseball, one statistic that is tracked for players is batting average. It is calculated with:

$$BA = \frac{\text{hits}}{\text{at bats}}$$

In the MLB, almost all players will have BA above 0.200. Most players will have a BA below 0.300. The highest career BA of all time is 0.366. A BA of 0.330 or higher (for a single season) is considered to be very good.

Let's say a team hires a new player. We don't know what this player's batting average will be. It is unknown.

In his first 10 at bats in the MLB, the player gets 5 hits. His batting average so far is 0.500.

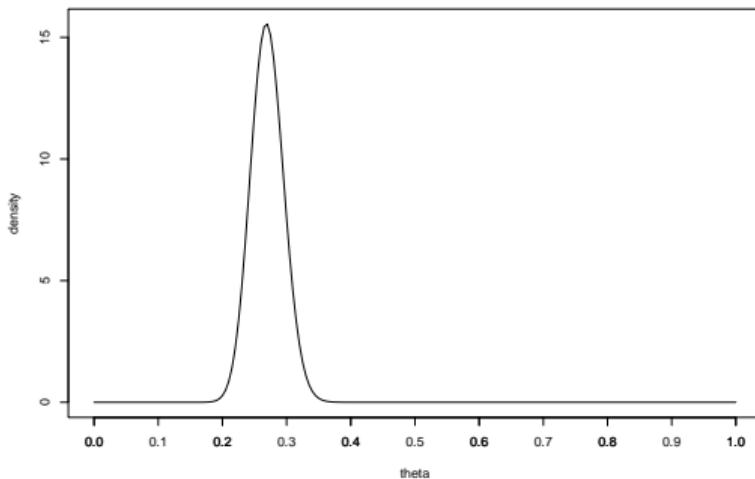
According to frequentist statistics, our estimate of the player's BA is 0.500. However, our gut and intuition tells us that by the end of the season, the player's BA is going to be much lower than 0.500.

How can we incorporate this into our calculations?

The Prior Distribution

In Bayesian stats, we will try to incorporate our knowledge about baseball into a prior distribution.

We want a distribution where almost all of the values are above 0.200 and most are below 0.300 and almost no one is above 0.400. Perhaps, the distribution can be drawn like this:



The Beta Distribution

The beta distribution is a continuous distribution defined on the interval [0,1]. It is frequently used to model a random proportion because proportions are also defined on [0,1].

If the random variable is Θ , the PDF is:

$$\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Where:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

While the value for $B(\alpha, \beta)$ looks scary, it is a constant and its only purpose is to make sure that the PDF integrates to 1.

Some intuition about beta distribution parameters

It might be helpful to think of the shape parameters in a beta distribution as how many “yes” and how many “no” values you have observed so far.

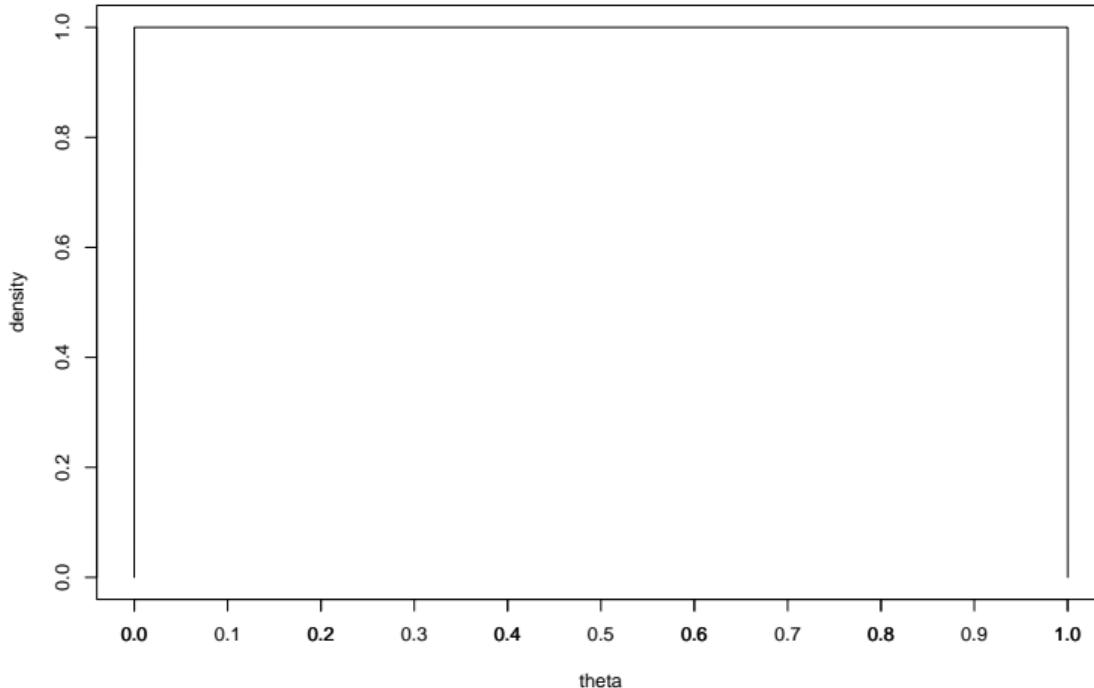
The resulting beta distribution will then reflect what is the probability of “yes.”

The beta distribution will generally have a peak around $\frac{\alpha}{\alpha+\beta}$.

The beta distribution will have more spread if the sum of α and β is low and will have less spread if the sum is high.

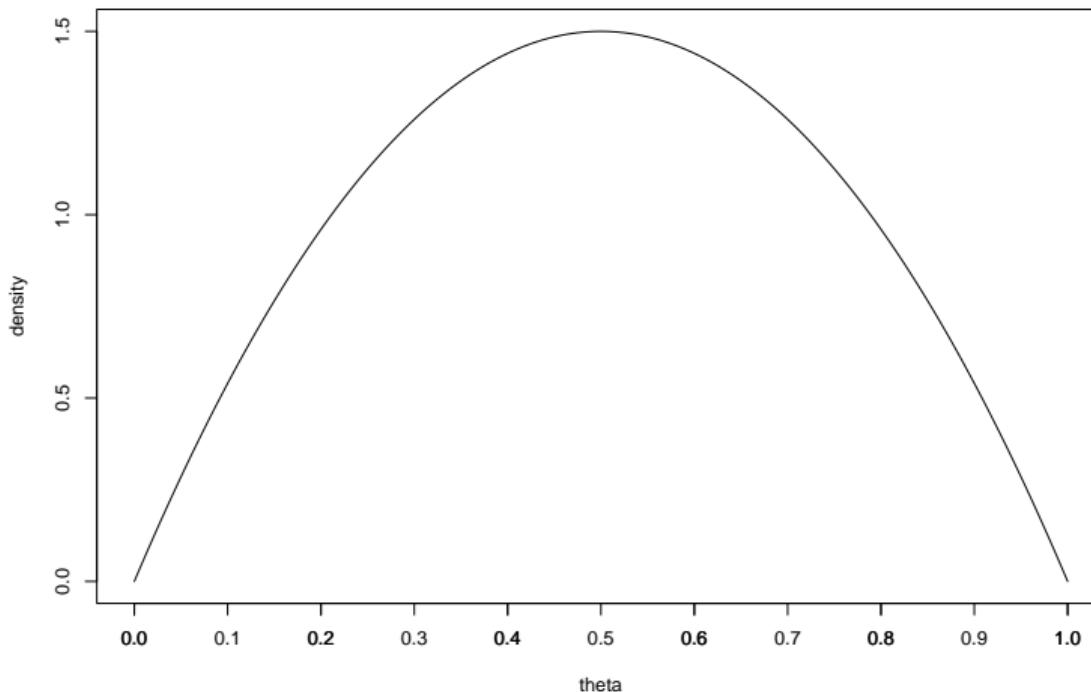
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 1 and beta = 1



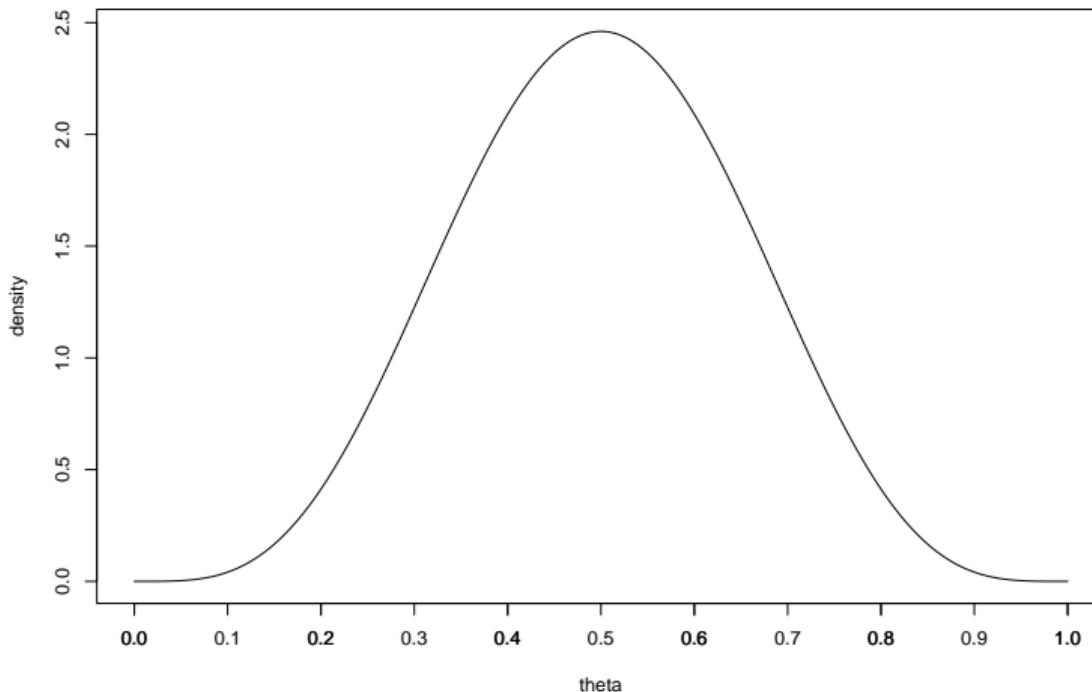
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 2 and beta = 2



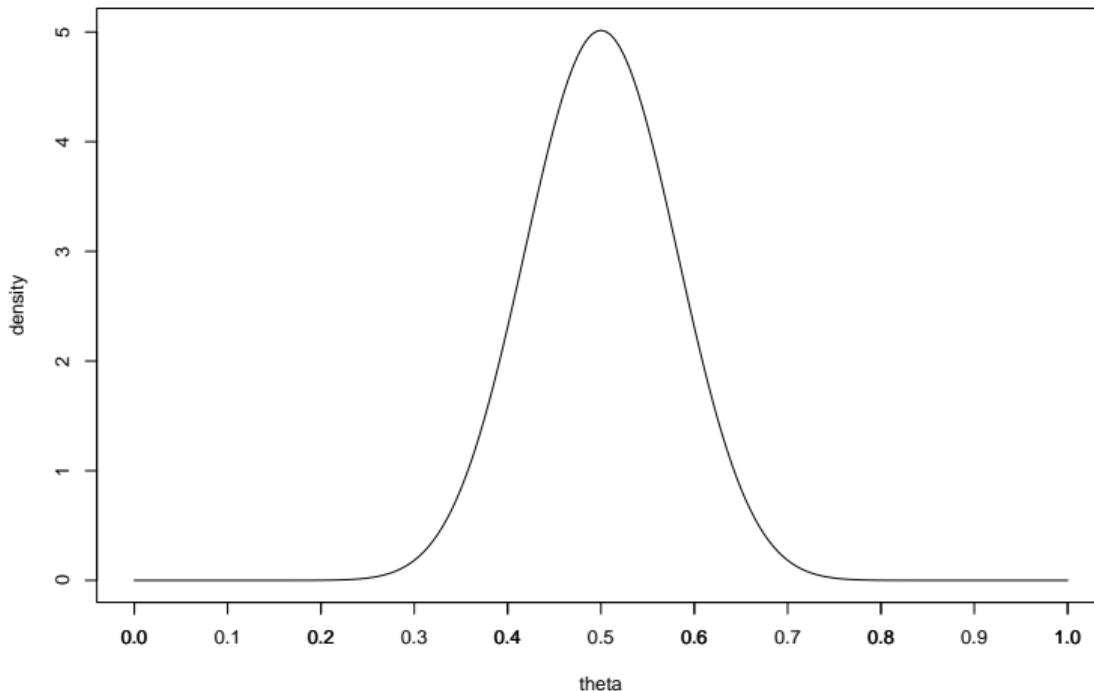
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 5 and beta = 5



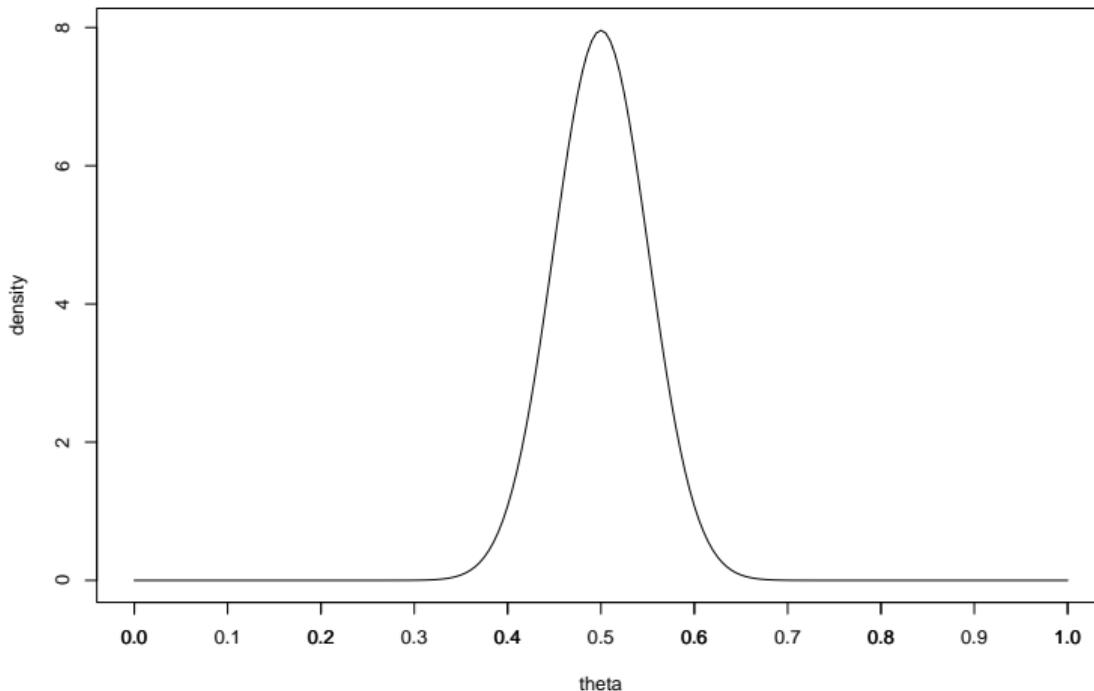
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 20 and beta = 20



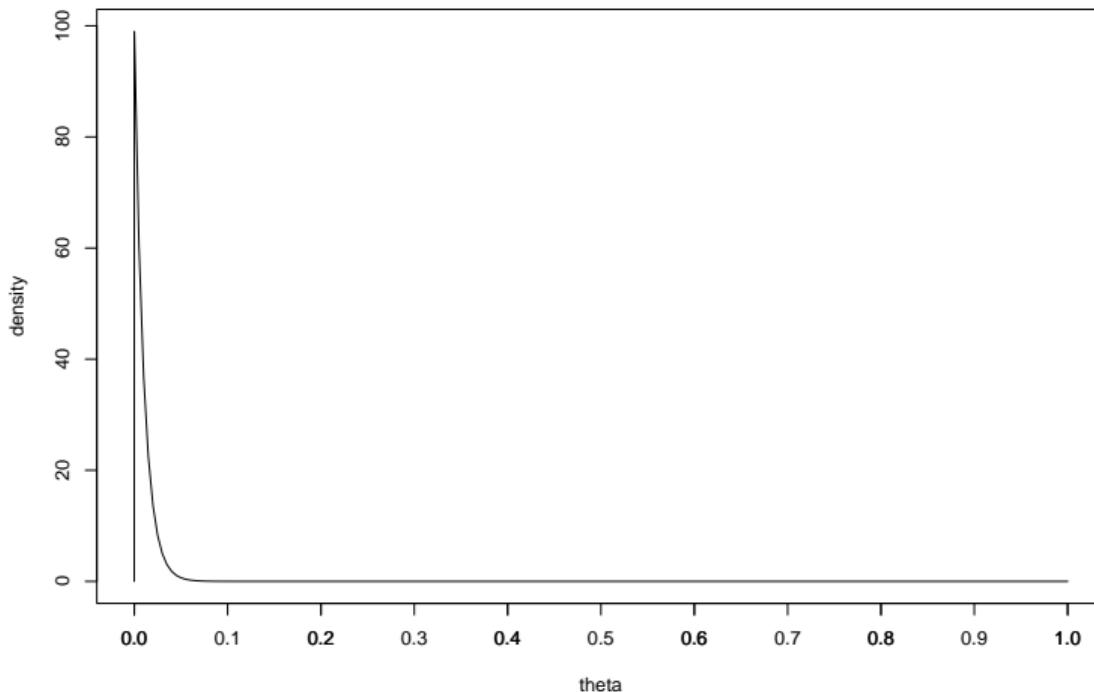
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 50 and beta = 50



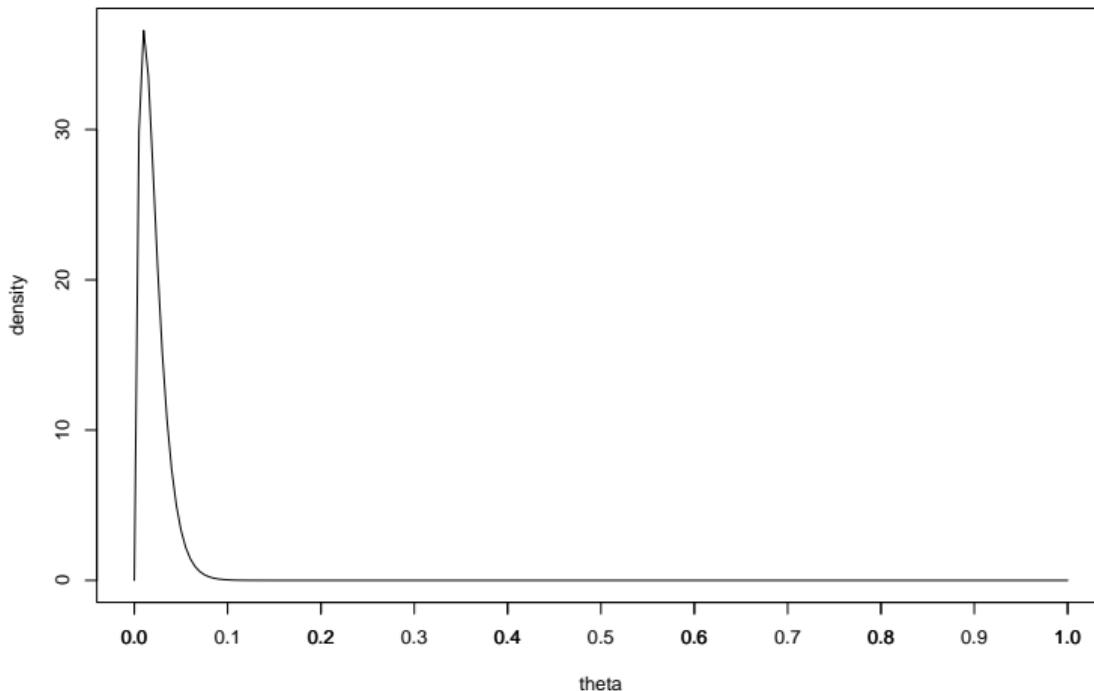
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 1 and beta = 99



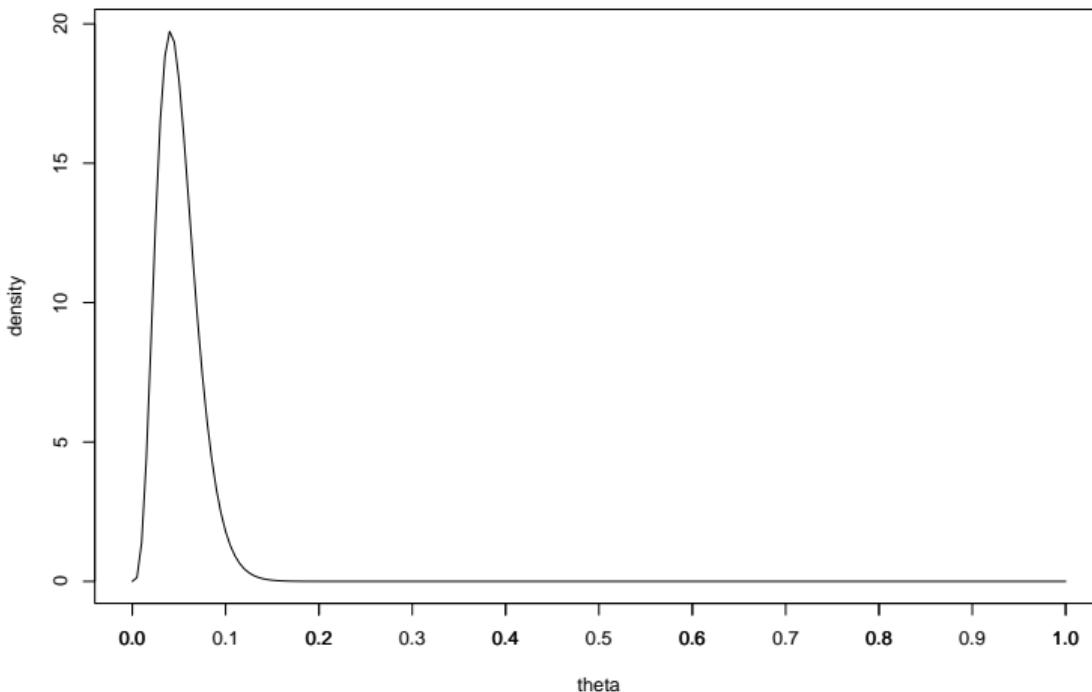
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 2 and beta = 98



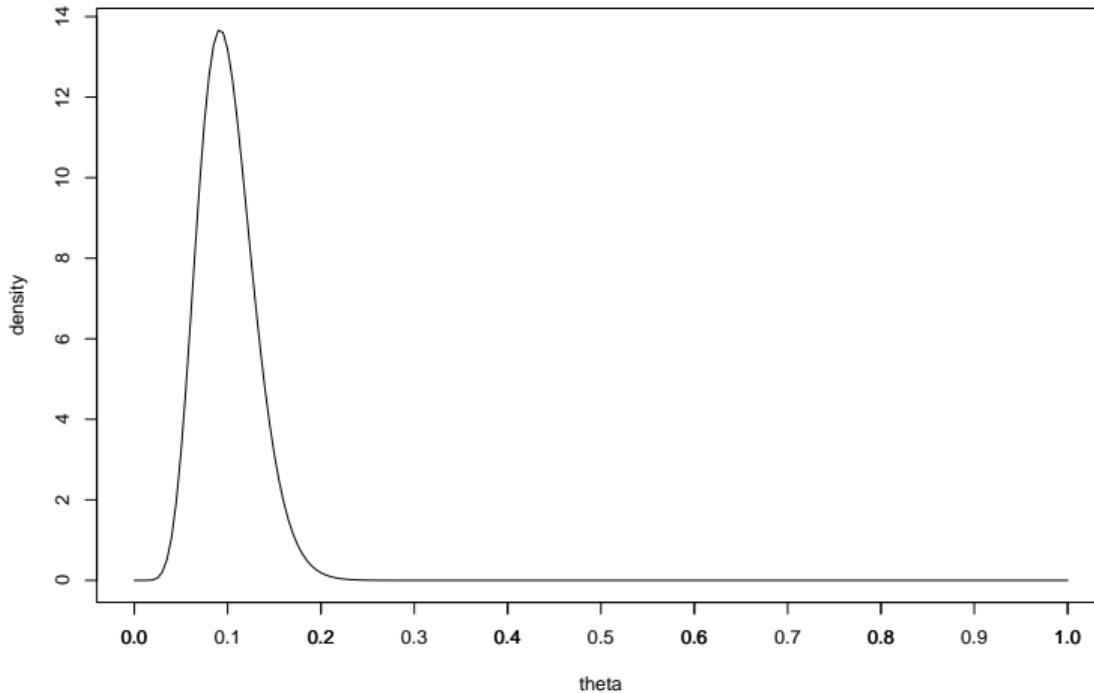
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 5 and beta = 95



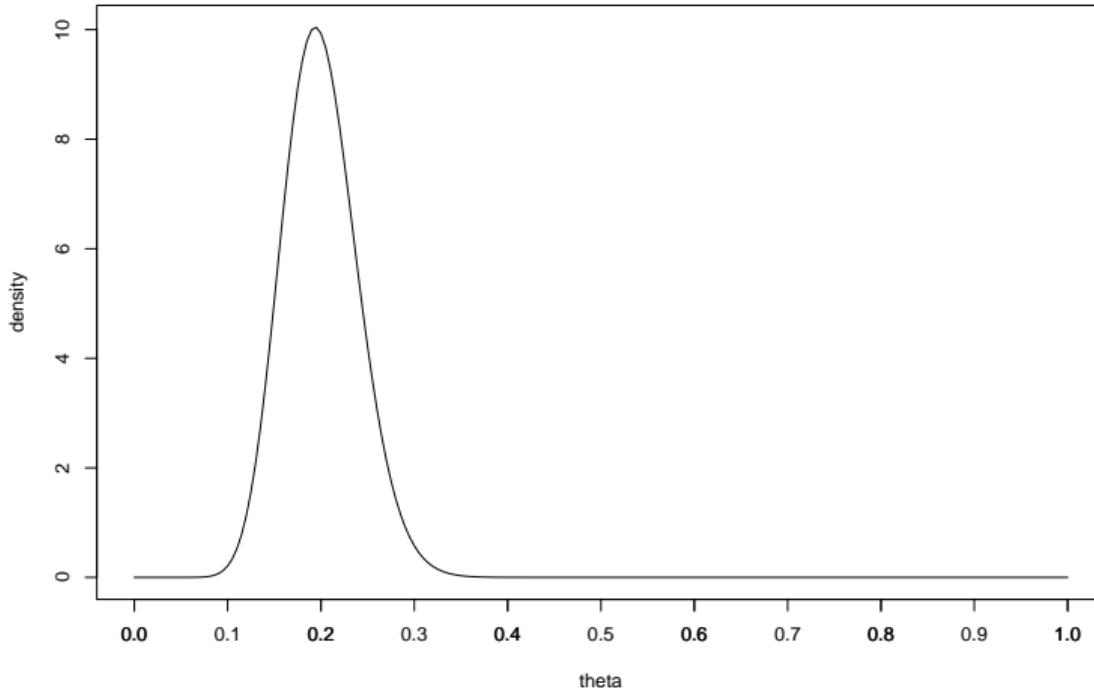
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 10 and beta = 90



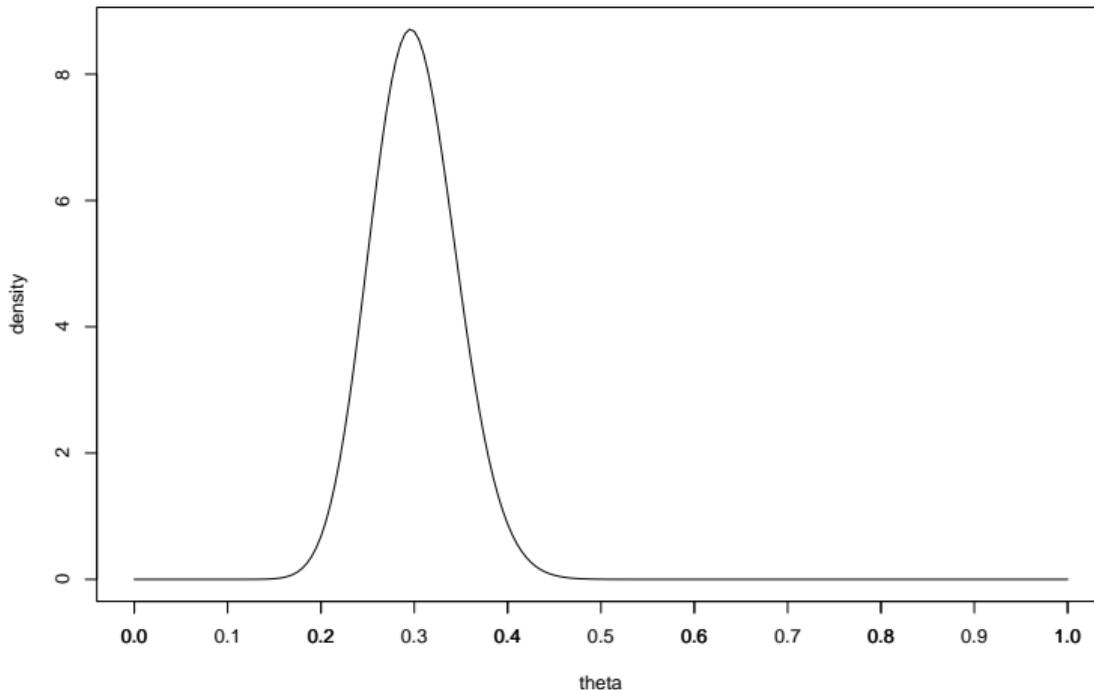
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 20 and beta = 80



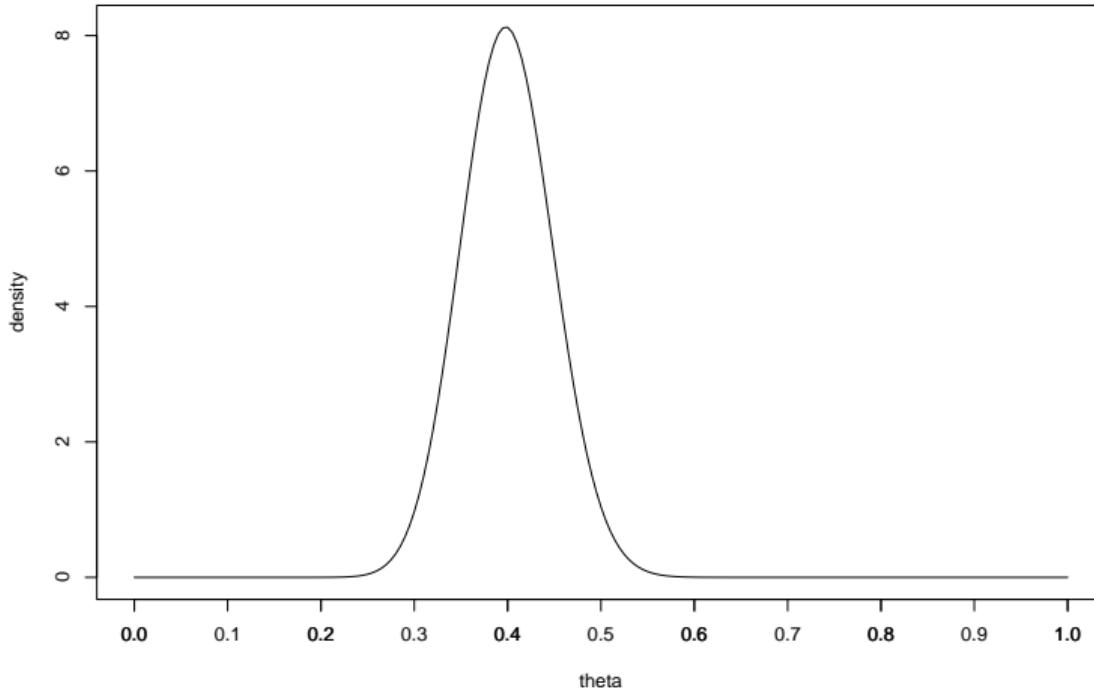
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 30 and beta = 70



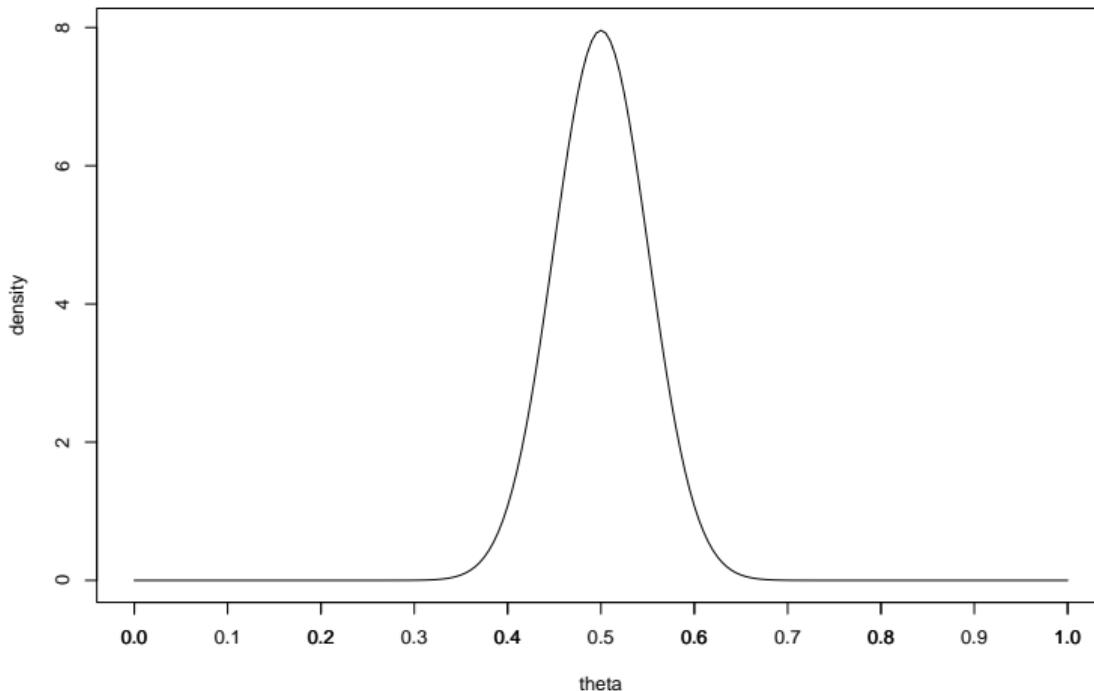
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 40 and beta = 60



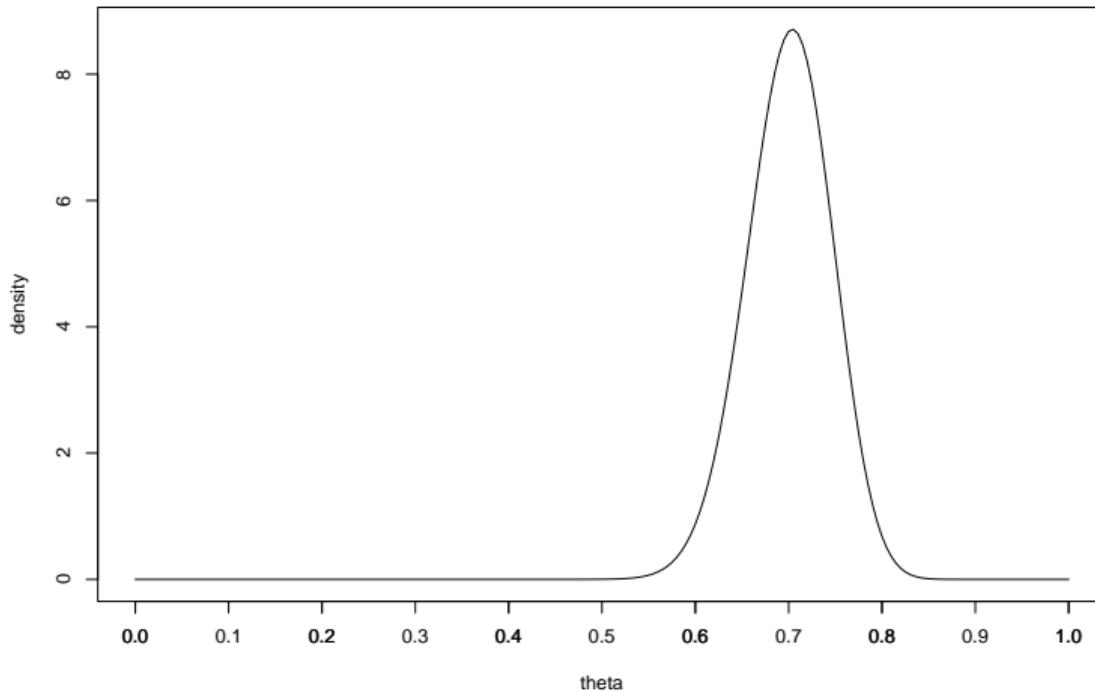
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 50 and beta = 50



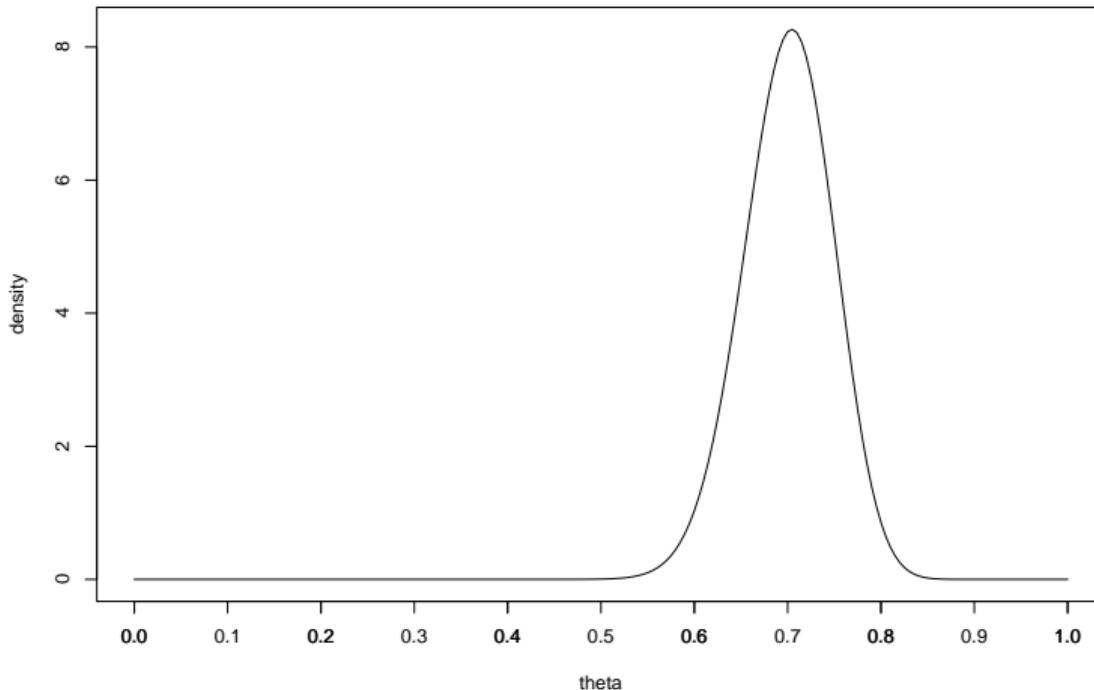
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 70 and beta = 30



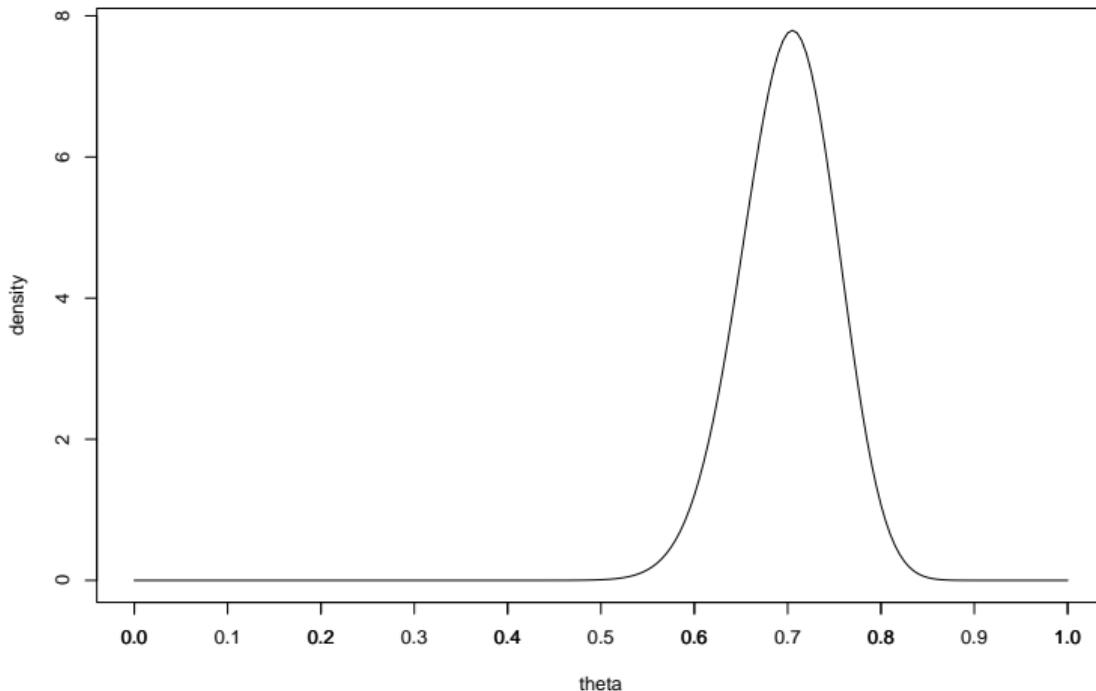
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 63 and beta = 27



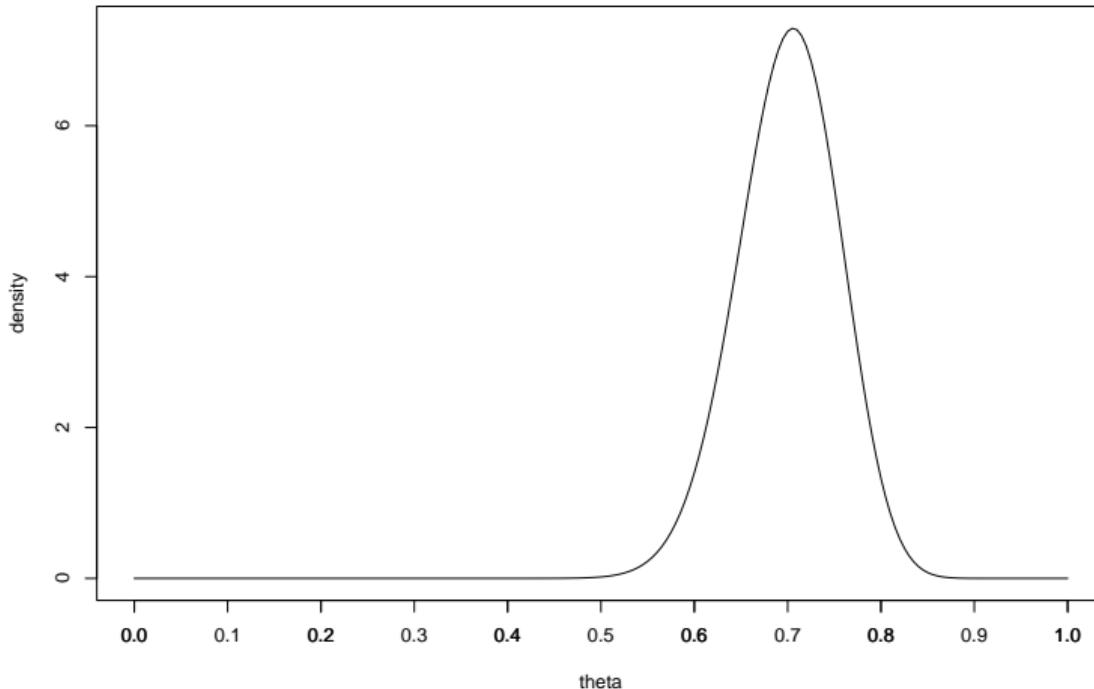
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 56 and beta = 24



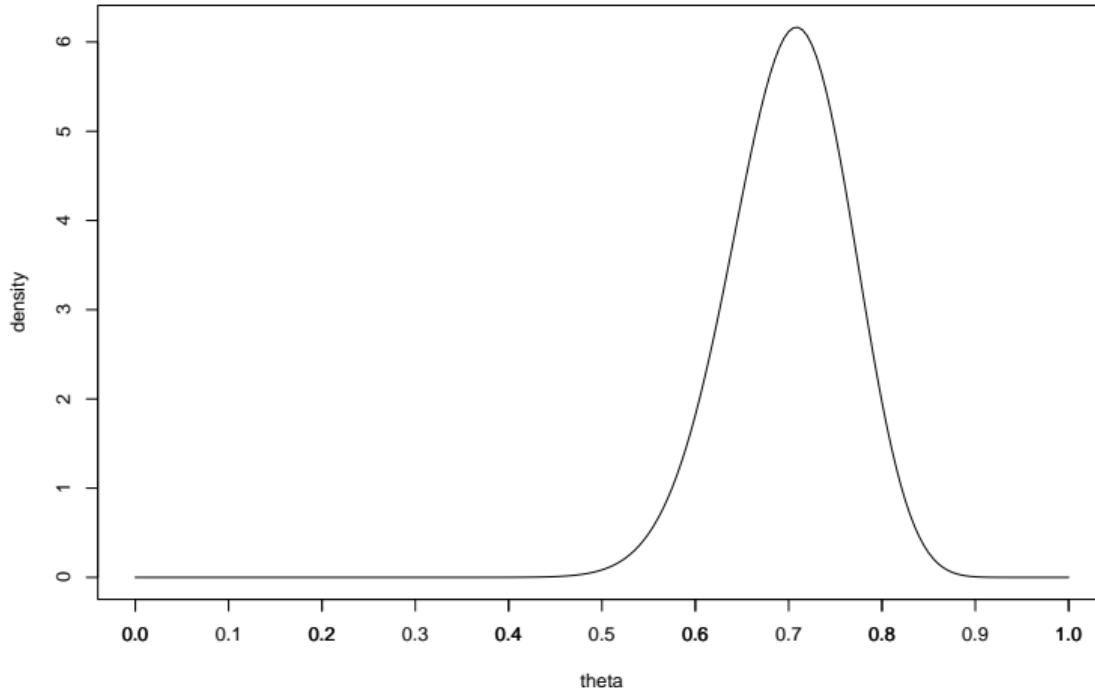
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 49 and beta = 21



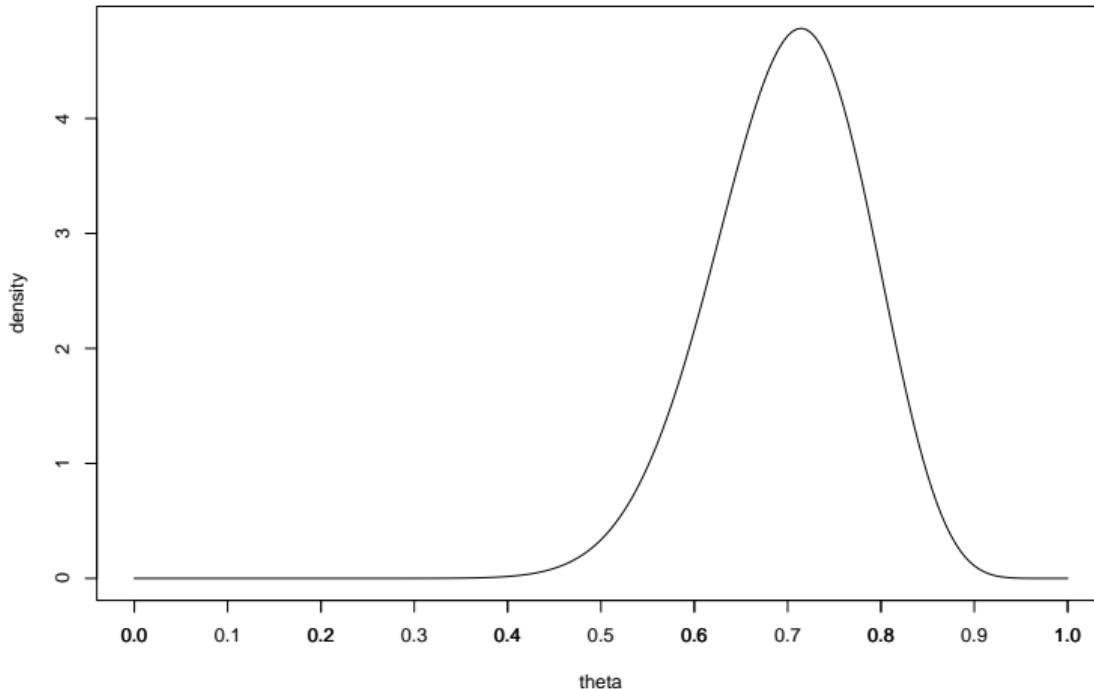
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 35 and beta = 15



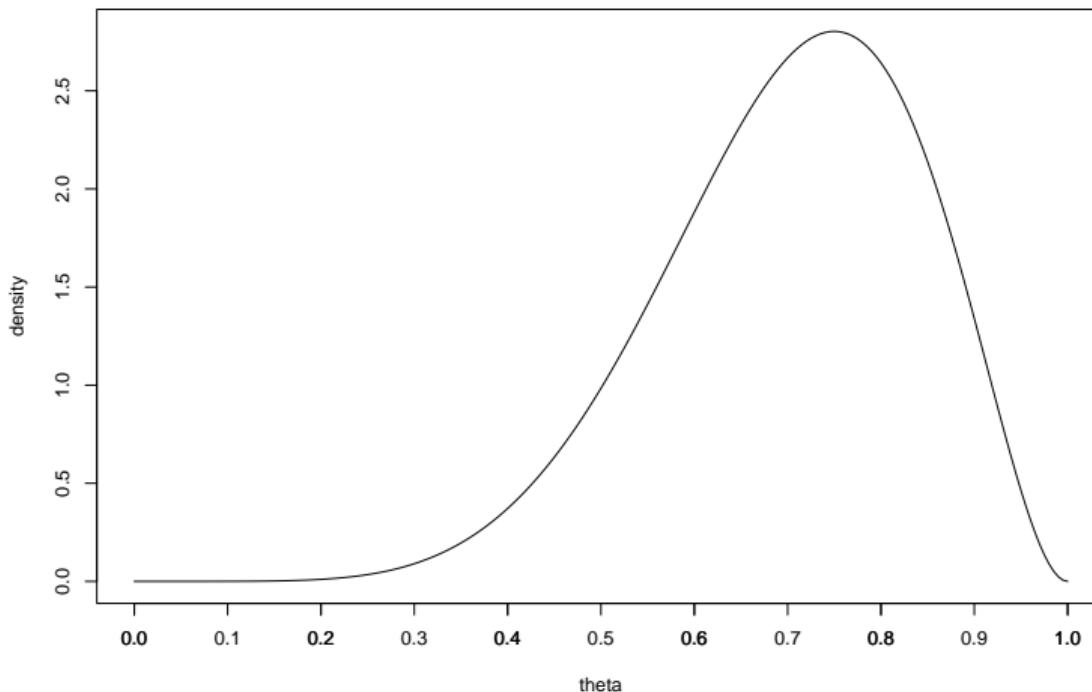
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 21 and beta = 9



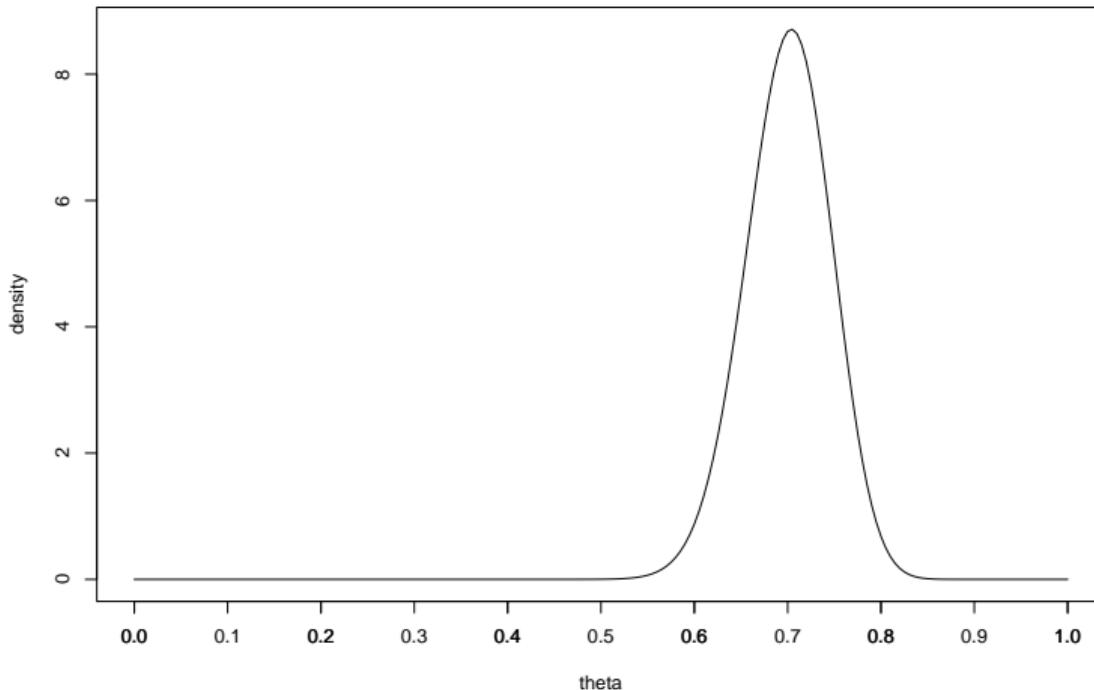
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 7 and beta = 3



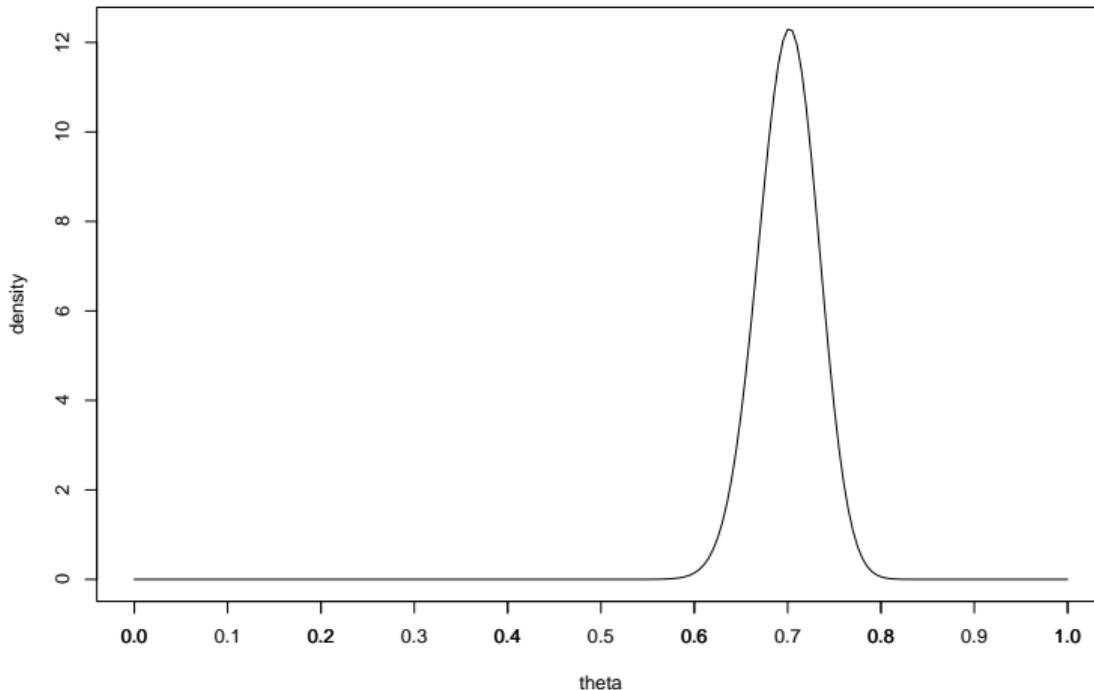
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 70 and beta = 30



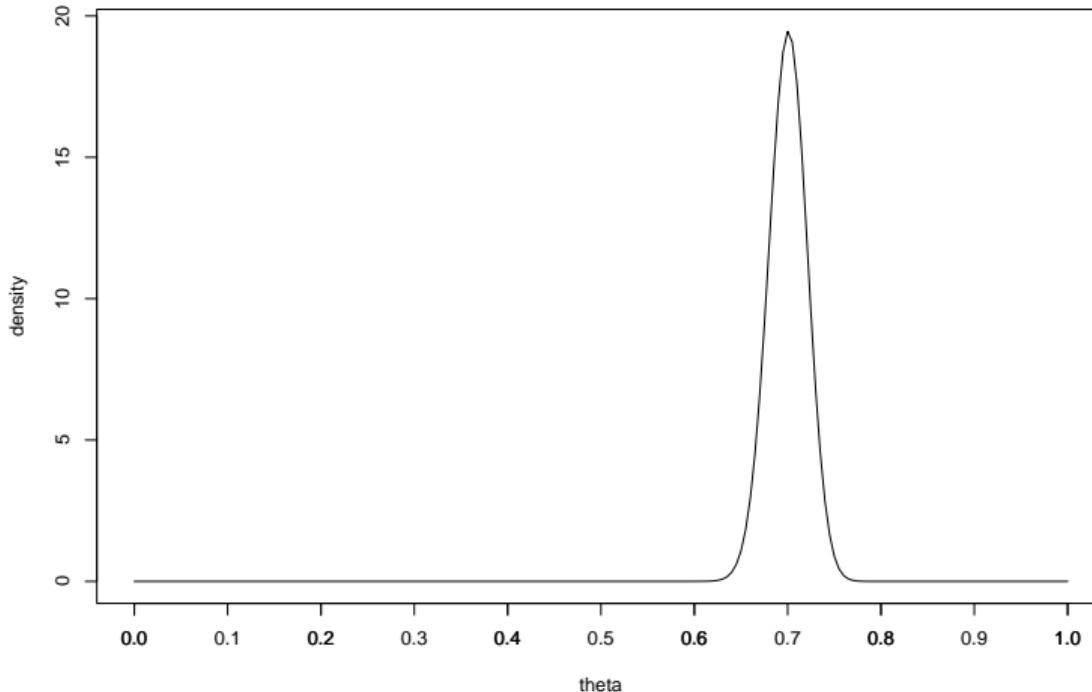
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 140 and beta = 60



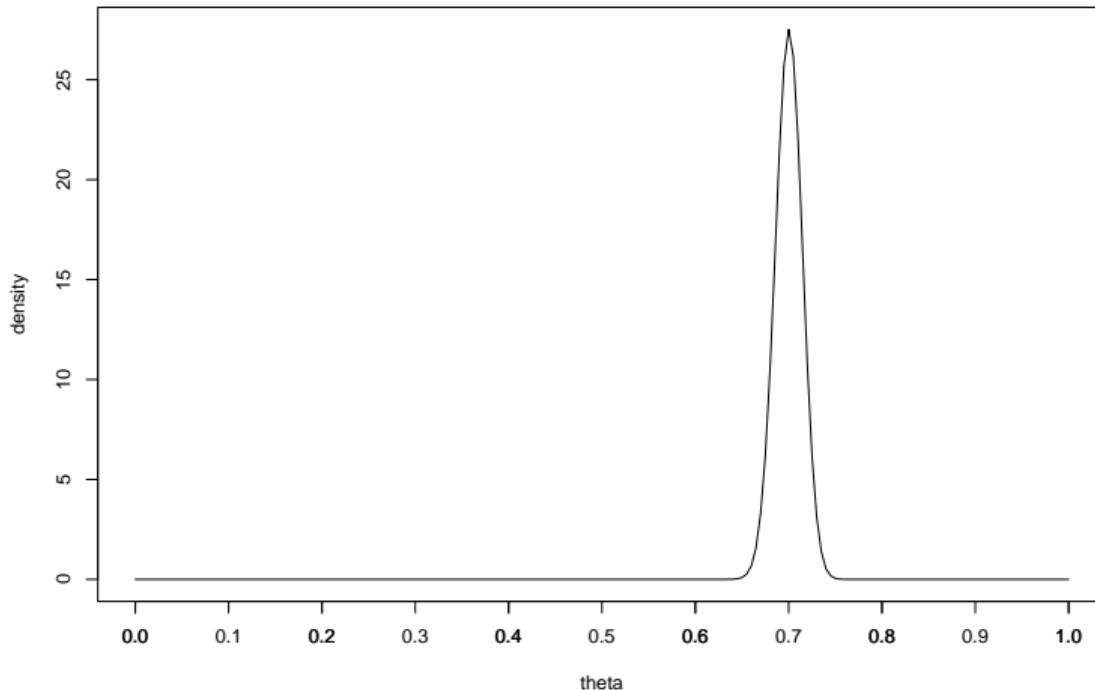
Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 350 and beta = 150



Examples of Beta Distributions:

Beta dist. w/ shape parameters alpha = 700 and beta = 300



Back to the baseball example

For our baseball example, we want our prior distribution selected so almost all of the values are above 0.200 and most are below 0.300 and almost no one is above 0.400.

Based on this, we may choose to use a Beta distribution with shape parameters $\alpha = 81$ and $\beta = 219$. These values are somewhat arbitrarily selected, but do align with our prior knowledge.

