Stats 102C - Lecture 1-2: Beta-Binomial Model

Miles Chen, PhD

Week 1 Wednesday

Copyright Miles Chen. For personal use only. Do not distribute.

Section 1

Lecture 1-2: Beta-Binomial Model

Copyright Miles Chen. For personal use only. Do not distribute.

The Beta Distribution

The PDF of the beta distribtuion is:

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

Where:

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

The purpose of this constant is so that the integral of the entire PDF is equal to 1.

Copyright Miles Chen. For personal use only. Do not distribute.

The Beta Distribution

The Beta distribution is a PDF defined on [0,1]. We know that all PDFs integrate to 1.

$$\int_0^1 \frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = 1$$

$$\frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = 1$$

Which means that the value of the function $B(\alpha,\beta)$ can be expressed as:

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \mathrm{B}(\alpha,\beta)$$

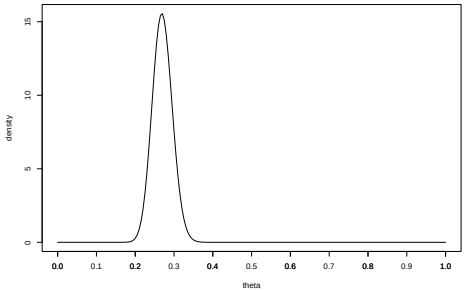
This relation will be useful later.

Copyright Miles Chen. For personal use only. Do not distribute.

Our baseball example

We model the prior distribution of a player's batting average with a Beta distribution with shape parameters $\alpha = 81$ and $\beta = 219$. These values are somewhat aribtrarily selected, but do align with our prior knowledge.

The result is a distribution where almost of the values are above 0.200, most are below 0.300, and almost no one is above 0.400.



Copyright Miles Chen. For personal use only. Do not distribute.

Our baseball example

For our baseball example, with $\alpha = 81$ and $\beta = 219$, the prior distribution is:

$$\frac{1}{\mathrm{B}(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \frac{1}{\mathrm{B}(81, 219)}\theta^{81-1}(1-\theta)^{219-1}$$

θ represents the player's batting average and can be thought of the probability of getting a hit.

Copyright Miles Chen. For personal use only. Do not distribute.

The likelihood function of the data

Let's say the new player has 10 at bats and earns 5 hits in the following sequence (H for Hit, x = 1; O for out, x = 0)

O H H O H O O H H O; or $\mathbf{x} = [0, 1, 1, 0, 1, 0, 0, 1, 1, 0]$

The probability of a hit is the player's batting average: θ.

The likelihood of this sequence of data is:

$$\theta^5(1-\theta)^5$$

And more generally, if you have a total of N observations, the likelihood of one sequence with a total of z successes will be:

$$\theta^z(1-\theta)^{N-z}$$

Copyright Miles Chen. For personal use only. Do not distribute.

Bayes' Rule

Bayes' rule:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

The numerator is equal to $\Pr(B \cap A)$

Copyright Miles Chen. For personal use only. Do not distribute.

Bayes' Rule

$$\Pr(\theta|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\theta)\Pr(\theta)}{\Pr(\mathbf{x})}$$

- Posterior: $\Pr(\theta|\mathbf{x})$ is the probability distribution of the parameter θ that we want to find.
- Likelihood: $\Pr(\mathbf{x}|\theta)$ is the probability of observing the observed data for a given value of θ.
- Prior: $\Pr(\theta)$ is prior probability distribution of θ.
- Marginal probability: $\Pr(\mathbf{x})$ is the probability of observing the values in our data regardless of the value of θ. This is equal to the integral of the numerator across all possible values of θ. This is a constant.

Because the marginal probability $\Pr(\mathbf{x})$ is a constant, the Posterior distribution of θ is proportional to the likelihood function times the prior distribution.

$$\Pr(\theta|\mathbf{x}) \propto \Pr(\mathbf{x}|\theta)\Pr(\theta)$$

Copyright Miles Chen. For personal use only. Do not distribute.

Bayes' Rule for our Baseball example

The Likelihood function for observing a sequence of 5 hits in 10 at bats is:

$$\Pr(\mathbf{x}|\theta) = \theta^5(1-\theta)^5$$

The prior distribution of θ is a beta distribution with shape parameters $\alpha = 81$ and $\beta = 219$. This has PDF:

$$\Pr(\theta) = \frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \frac{1}{\mathrm{B}(81,219)}\theta^{81-1}(1-\theta)^{219-1}$$

Because I'm expressing the posterior distribution as a function that is proportional to the true posterior distribution, we can ignore the constant $\mathrm{B}(81,219)^{-1}$.

$$\Pr(\theta|\mathbf{x}) \propto \Pr(\mathbf{x}|\theta)\Pr(\theta) = \theta^5(1-\theta)^5\theta^{81-1}(1-\theta)^{219-1}$$

Copyright Miles Chen. For personal use only. Do not distribute.

Posterior Distribution

$$\Pr(\theta|\mathbf{x}) \propto \Pr(\mathbf{x}|\theta)\Pr(\theta) = \theta^5(1-\theta)^5\theta^{81-1}(1-\theta)^{219-1}$$

$$\Pr(\theta|\mathbf{x}) \propto \theta^{85}(1-\theta)^{223}$$

Now we must find the normalizing constant to make it a PDF.

Copyright Miles Chen. For personal use only. Do not distribute.

Beta function

Earlier, we established:

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \mathrm{B}(\alpha,\beta)$$

Our posterior distribution is proportional to the numerator of Bayes Rule:

$$\Pr(\theta|\mathbf{x}) \propto \theta^{85}(1-\theta)^{223}$$

If we integrate this numerator across all possible values of θ (from 0 to 1), we get:

$$\int_0^1 \theta^{85}(1-\theta)^{223}d\theta = \mathrm{B}(86,224)$$

Thus the PDF of the posterior distribution can be expressed as:

$$\Pr(\theta|\mathbf{x}) = \frac{1}{\mathrm{B}(86,224)}\theta^{85}(1-\theta)^{223}$$

Copyright Miles Chen. For personal use only. Do not distribute.

Posterior Distribution

The PDF of the posterior distribution is:

$$\Pr(\theta|\mathbf{x}) = \frac{1}{\mathsf{B}(86, 224)}\theta^{85}(1-\theta)^{223}$$

Which is the beta distribution with $\alpha = 86$ and $\beta = 224$

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

Copyright Miles Chen. For personal use only. Do not distribute.

Generalizing the Beta-Binomial relationship

Let's say you have a prior distribution for θ with α and β:

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

The data comes from a binomial distribution (or is equivalent to a sequence of N Bernoulli draws). The likelihood for seeing a total of z successes is:

$$\Pr(\mathbf{x}|\theta) = \theta^{z}(1-\theta)^{N-z}$$

The PDF of the posterior distribution is:

$$\Pr(\theta|\mathbf{x}) = \frac{1}{\mathrm{B}(z+\alpha, N-z+\beta)}\theta^{z+\alpha-1}(1-\theta)^{N-z+\beta-1}$$

Which is the beta distribution with $\alpha = z + \alpha$ and $\beta = N - z + \beta$

Copyright Miles Chen. For personal use only. Do not distribute.

Conjugate Priors

In today's examples, the likelihood function of the binomial / Bernoulli data multiplies "nicely" with the Beta distribution.

When the likelihood function and the PDF of the prior distribution "multiply nicely," we can say they are conjugate priors.

There are quite a few distributions that are conjugate priors for different types of data. The beta distribution and binomial distribution is one such example.

https://en.wikipedia.org/wiki/Conjugate_prior

Copyright Miles Chen. For personal use only. Do not distribute.

A note about the following slides

In the following slides, I will plot:

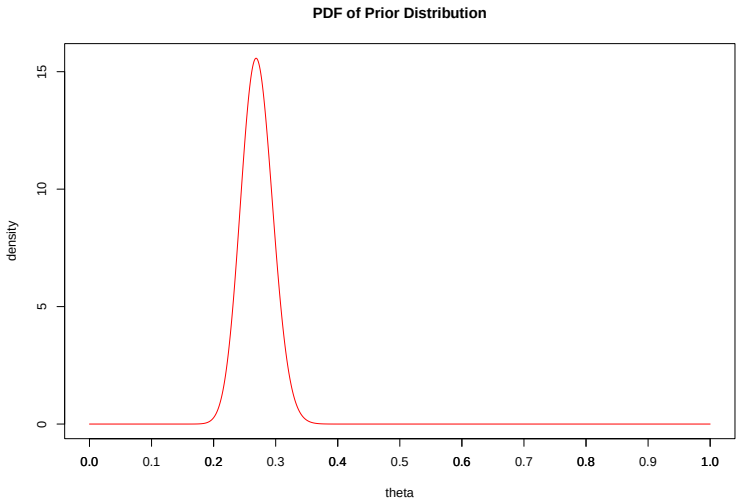
- the PDF of the prior distribution (red)
- the likelihood function (blue)
- the PDF of the posterior distribution (black)
- a plot with all three functions

In the plot with all three functions, I artificially amplify the likelihood function by multiplying it by a constant. The true value of the likelihood function will be much smaller and would be "too flat" to be seen on the same scale as the density functions. For these plots, I plot a beta distribution instead of the likelihood function to amplify the shape of the likelihood function. The likelihood function of a binomial probability for N observations with z successes is **proportional to, but not equal to** a beta distribution with $\alpha = z + 1$ and $\beta = N - z + 1$.

The likelihood function does not integrate to 1. The density functions do integrate to 1.

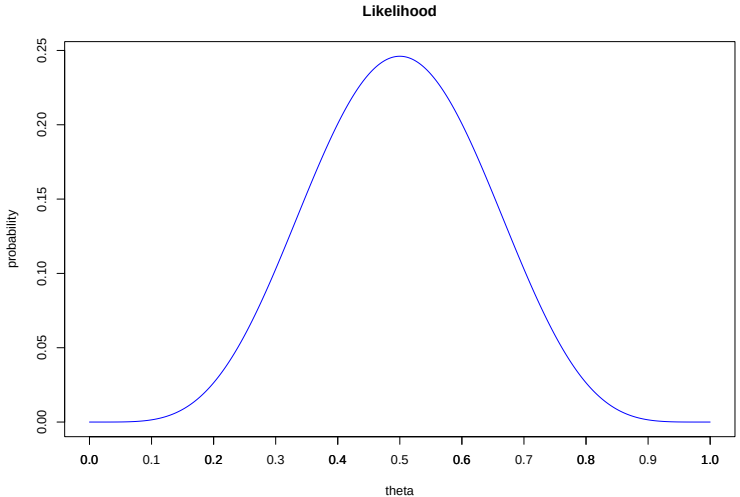
Copyright Miles Chen. For personal use only. Do not distribute.

Prior distribution of the player's batting average: almost everyone has a BA over 0.200, most are below 0.300, and everyone is below 0.400



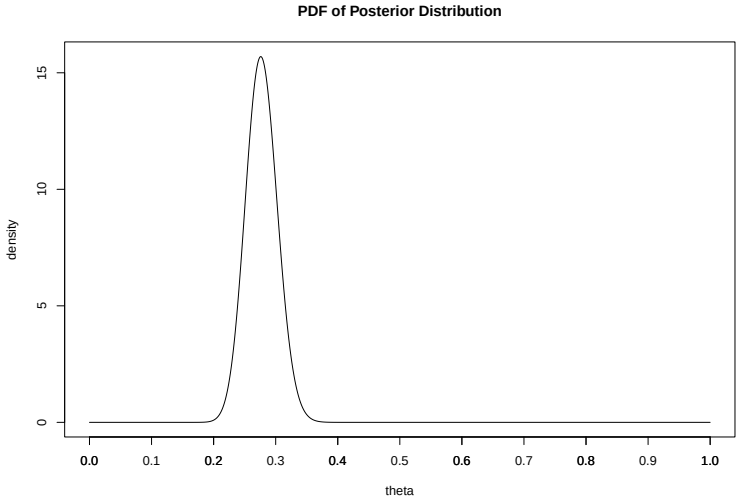
Copyright Miles Chen. For personal use only. Do not distribute.

Likelihood of the player's data: 5 hits in 10 at bats. The likelihood is maximized at 0.5, but our intuition tells us that this kind of batting average cannot be sustained for a very long time.



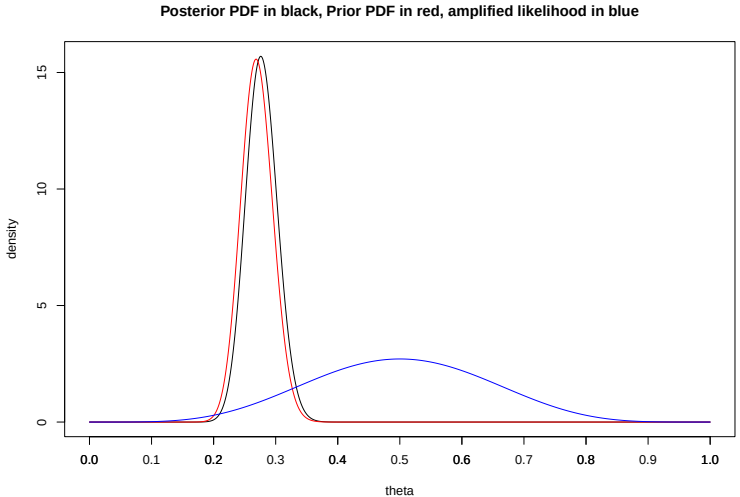
Copyright Miles Chen. For personal use only. Do not distribute.

The posterior is a compromise between the likelihood and prior.



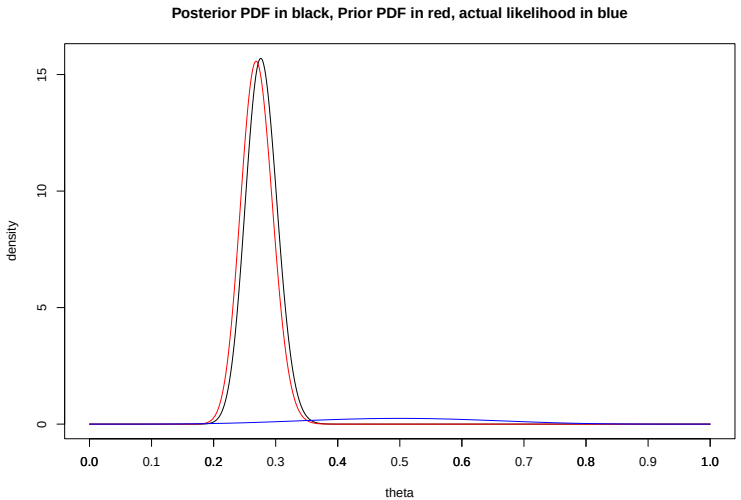
Copyright Miles Chen. For personal use only. Do not distribute.

The posterior is a compromise between the likelihood and prior. The prior has a mean around 0.27 while the posterior has a mean that has been shifted slightly higher to 0.28



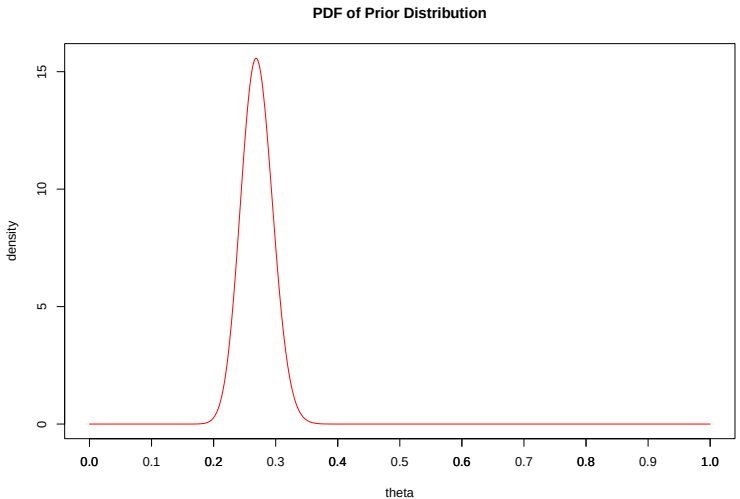
Copyright Miles Chen. For personal use only. Do not distribute.

Just for comparison, I plot the actual likelihood function instead of the "amplified" version. You can see that the likelihood function is much flatter. In fact, as N grows, the likelihood function gets even more flat.



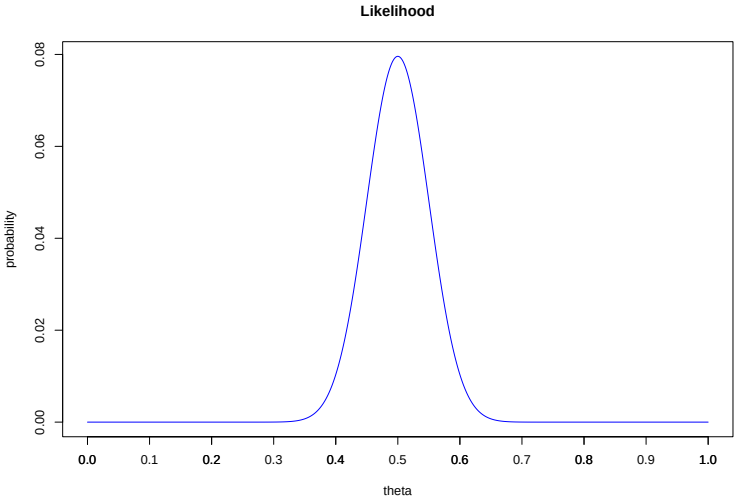
Copyright Miles Chen. For personal use only. Do not distribute.

Prior distribution of the player's batting average: almost everyone has a BA over 0.200, most are below 0.300, and everyone is below 0.400

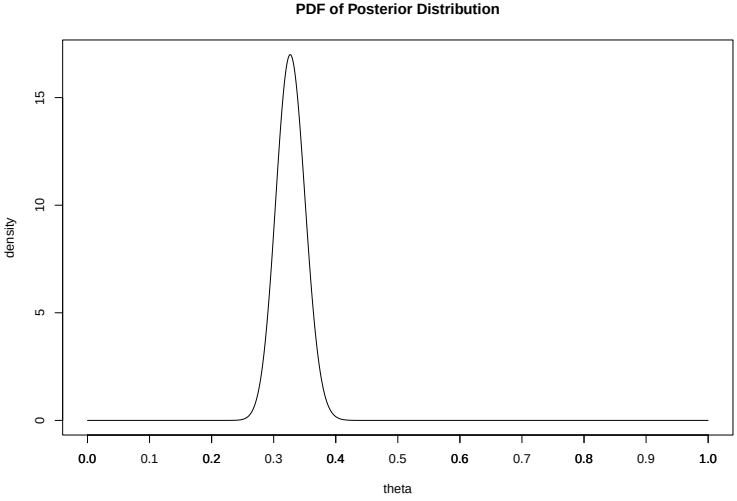


Copyright Miles Chen. For personal use only. Do not distribute.

Let's say the player is able to sustain this batting average over more at bats: 50 hits in 100 at bats. The likelihood is maximized at 0.5. We still doubt that this can be sustained for a very long time, but there is now more evidence showing this player has a high batting average.

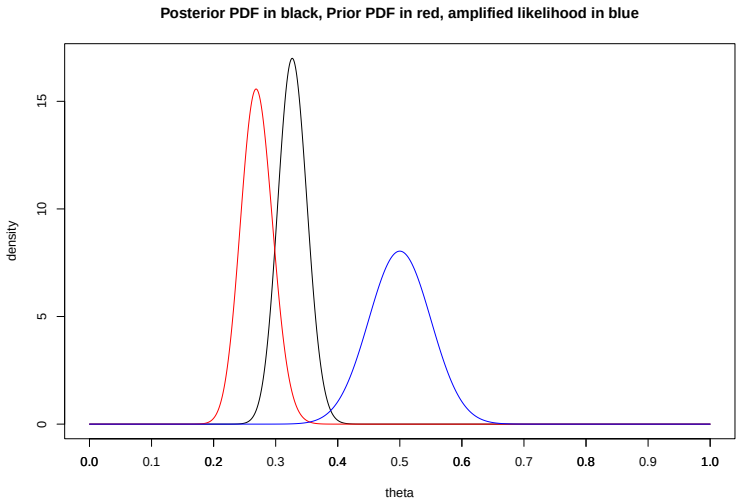


Copyright Miles Chen. For personal use only. Do not distribute.



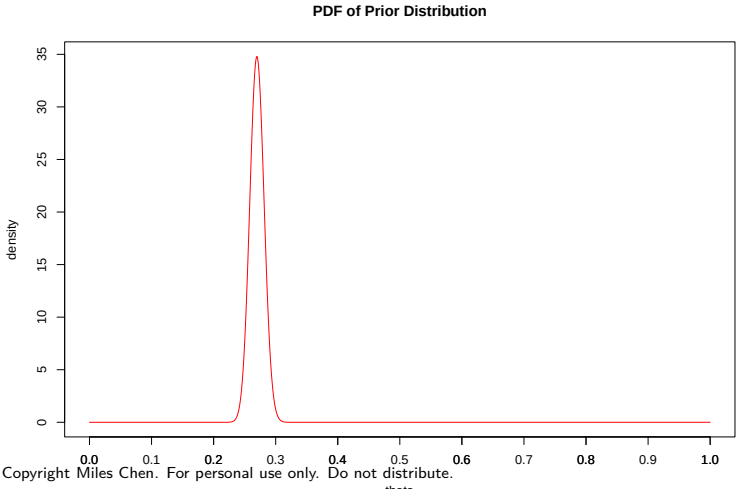
Copyright Miles Chen. For personal use only. Do not distribute.

The compromise between the prior and likelihood is visible. With more data (100 observations) to back up the likelihood, the posterior is shifted more than when we had only 10 observations. The posterior distribution now has a mean around 0.33.



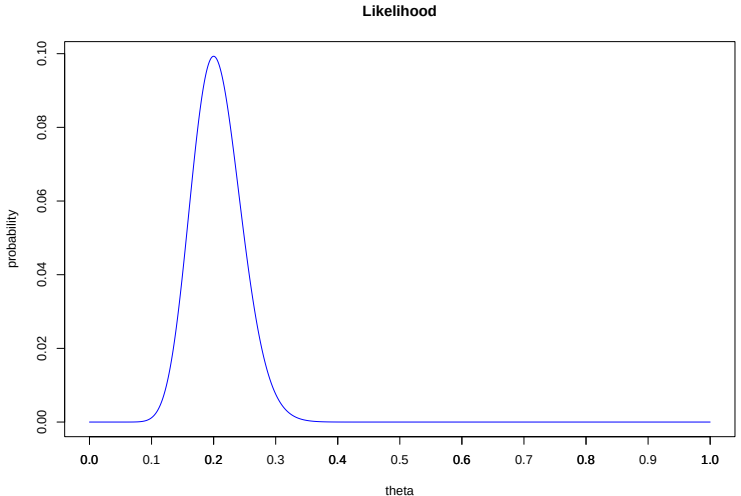
Copyright Miles Chen. For personal use only. Do not distribute.

This distribution represents the prior distribution of a player who has had many at bats. I'm using parameters $\alpha = 405$ and $\beta = 1095$. The distribution has a mean of 0.27. The prior distribution has much less spread indicating that because we have lots of data from this player, we are very confident that his true batting average is close to this value.

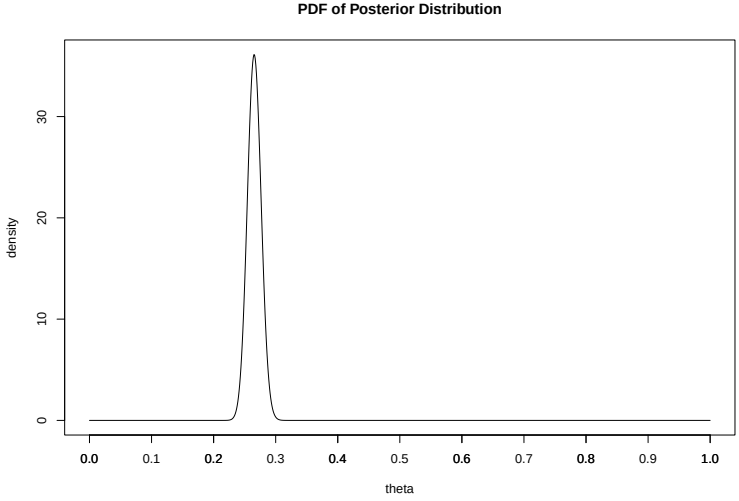


Copyright Miles Chen. For personal use only. Do not distribute.

Let's say the player has had 100 at bats and has only 20 hits. This might be considered a "slump" for the player.

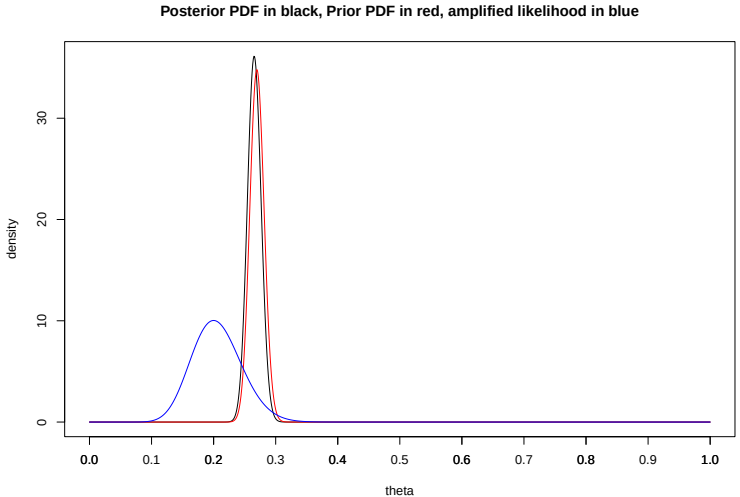


Copyright Miles Chen. For personal use only. Do not distribute.



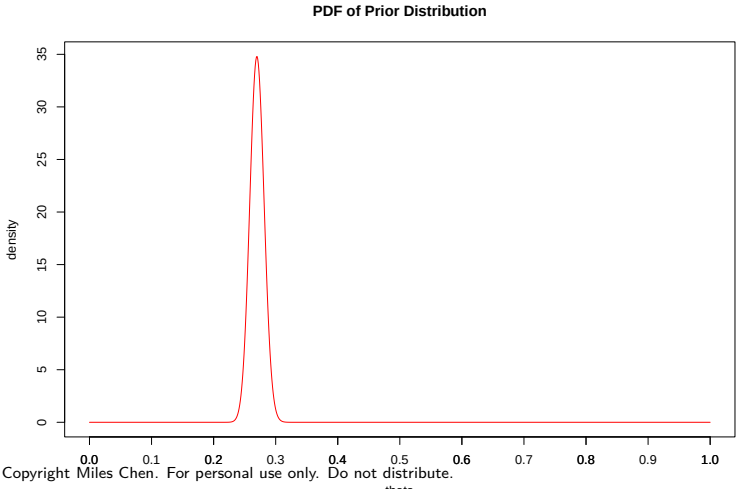
Copyright Miles Chen. For personal use only. Do not distribute.

The posterior distribution is only slightly affected because the prior distribution was established based on many observations.



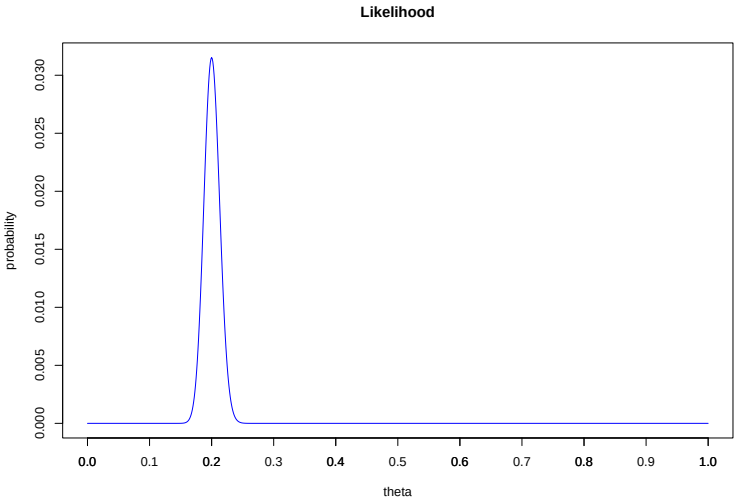
Copyright Miles Chen. For personal use only. Do not distribute.

This distribution represents the prior distribution of a player who has had many at bats. I'm using parameters $\alpha = 405$ and $\beta = 1095$. The distribution has a mean of 0.27. The prior distribution has much less spread indicating that because we have lots of data from this player, we are very confident that his true batting average is close to this value.

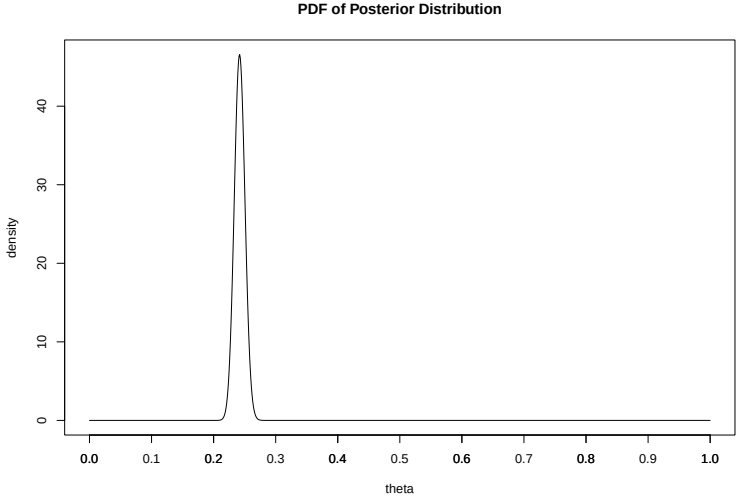


Copyright Miles Chen. For personal use only. Do not distribute.

This time the “slump” is sustained for many at bats.

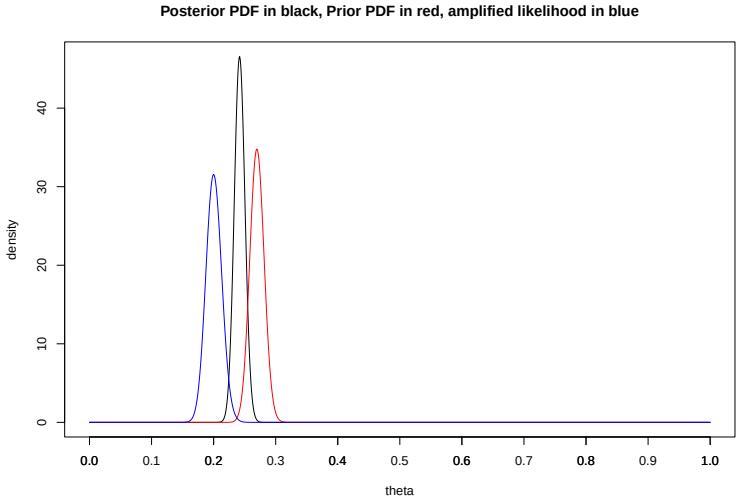


Copyright Miles Chen. For personal use only. Do not distribute.



Copyright Miles Chen. For personal use only. Do not distribute.

The posterior distribution is affected more.



Copyright Miles Chen. For personal use only. Do not distribute.