Prior vs Likelihood vs Posterior
Posterior Predictive Distribution
Poisson Data

Statistics 220

Spring 2005

Choosing the Likelihood Model

While much thought is put into thinking about priors in a Bayesian Analysis, the data (likelihood) model can have a big effect.

Choices that need to be made involve

- Independence vs Exchangable vs More Complex Dependence
- Tail size, e.g. Normal vs t_{df}
- Probability of events

Example: Probability of God's Existance

Two different analyses - both using the prior $P[\text{God}] = P[\text{No God}] = 0.5$

Likelihood Ratio Components:

$$D_i = \frac{P[\text{Data}_i|\text{God}]}{P[\text{Data}_i|\text{No God}]}$$

Evidence (Data_i)	D_i - Unwin	D_i - Shermer
Recognition of goodness	10	0.5
Existence of moral evil	0.5	0.1
Existence of natural evil	0.1	0.1
Intranatural miracles (prayers)	2	1
Extranatural miracles (resurrection)	1	0.5
Religious experiences	2	0.1

$P[\text{God}|\text{Data}]$:

- Unwin: $\frac{2}{3}$
- Shermer: 0.00025

So even starting with the same prior, the difference beliefs about what the data says gives quite different posterior probabilities.

This is based on an analysis published in an July 2004 Scientific American article. (Available on the course web site on the Articles page.)

Stephen D. Unwin is a risk management consultant who has done work in physics on quantum gravity. He is author of the book *The Probability of God*.

Michael Shermer is the publisher of *Skeptic* and a regular contributor to *Scientific American* as author of the column *Skeptic*.

Note in the article, Bayes' rule is presented as

$$P[\text{God}|\text{Data}] = \frac{P[\text{God}]D}{P[\text{God}]D + P[\text{No God}]}$$

See if you can show that this is equivalent to the normal version of Bayes' rule, under the assumption that the components of the data model are independent.

Prior vs Likelihood vs Posterior

The posterior distribution can be seen as a compromise between the prior and the data

In general, this can be seen based on the two well known relationships

$$E[\theta] = E[E[\theta|y]] \quad (1)$$

$$\mathrm{Var}(\theta) = E[\mathrm{Var}(\theta|y)] + \mathrm{Var}(E[\theta|y]) \quad (2)$$

The first equation says that our prior mean is the average of all possible posterior means (averaged over all possible data sets).

The second says that the posterior variance is, on average, smaller than the prior variance. The size of the difference depends on the variability of the posterior means.

This can be exhibited more precisely using examples

- Binomial Model - Conjugate prior

$$\begin{aligned} \pi &\sim Beta(a, b) \\ y|\pi &\sim Bin(n, \pi) \\ \pi|y &\sim Beta(a + y, b + n - y) \end{aligned}$$

Prior mean:

$$E[\pi] = \frac{a}{a+b} = \tilde{\pi}$$

MLE:

$$\hat{\pi} = \frac{y}{n}$$

Then the posterior mean satisfies

$$E[\pi|y] = \frac{a+b}{a+b+n}\tilde{\pi} + \frac{n}{a+b+n}\hat{\pi} = \bar{\pi}$$

a weighted average of the prior mean and the sample proportion (MLE)

Prior variance:

$$\mathrm{Var}(\pi) = \frac{\tilde{\pi}(1-\tilde{\pi})}{a+b+1}$$

Posterior variance:

$$\mathrm{Var}(\pi|y) = \frac{\bar{\pi}(1-\bar{\pi})}{a+b+n+1}$$

So if n is large enough, the posterior variance will be smaller than the prior variance.

- Normal Model - Conjugate prior, fixed variance

$$\begin{aligned} \theta &\sim N(\mu_0, \tau_0^2) \\ y_i|\theta &\overset{iid}{\sim} N(\theta, \sigma^2); \quad i = 1, \ldots, n \\ \theta|y &\sim N(\mu_n, \tau_n^2) \end{aligned}$$

Then

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

So the posterior mean is a weighted average of the prior mean and a sample mean of the data.

The posterior mean can be thought of in two other ways

$$\begin{aligned}\mu_n &= \mu_0 + (\bar{y} - \mu_0)\frac{\tau_0^2}{\frac{\sigma^2}{n} + \tau_0^2} \\ &= \bar{y} - (\bar{y} - \mu_0)\frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau_0^2}\end{aligned}$$

The first case has μ_n as the prior mean adjusted towards the sample average of the data.

The second case has the sample average $shrunk$ towards the prior mean.

In most problems, the posterior mean can be thought of as a shrinkage estimator, where the estimate just based on the data is shrunk toward the prior mean. The form of the shrinkage may not be able to be written out in as quite a nice form for more general problems.

In this example the posterior variance is never bigger than the prior variance as

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \geq \frac{1}{\tau_0^2} \qquad \text{and} \qquad \frac{1}{\tau_n^2} \geq \frac{n}{\sigma^2}$$

The first part of this is thought of as

$$\text{Posterior Precision} = \text{Prior Precision} + \text{Data Precision}$$

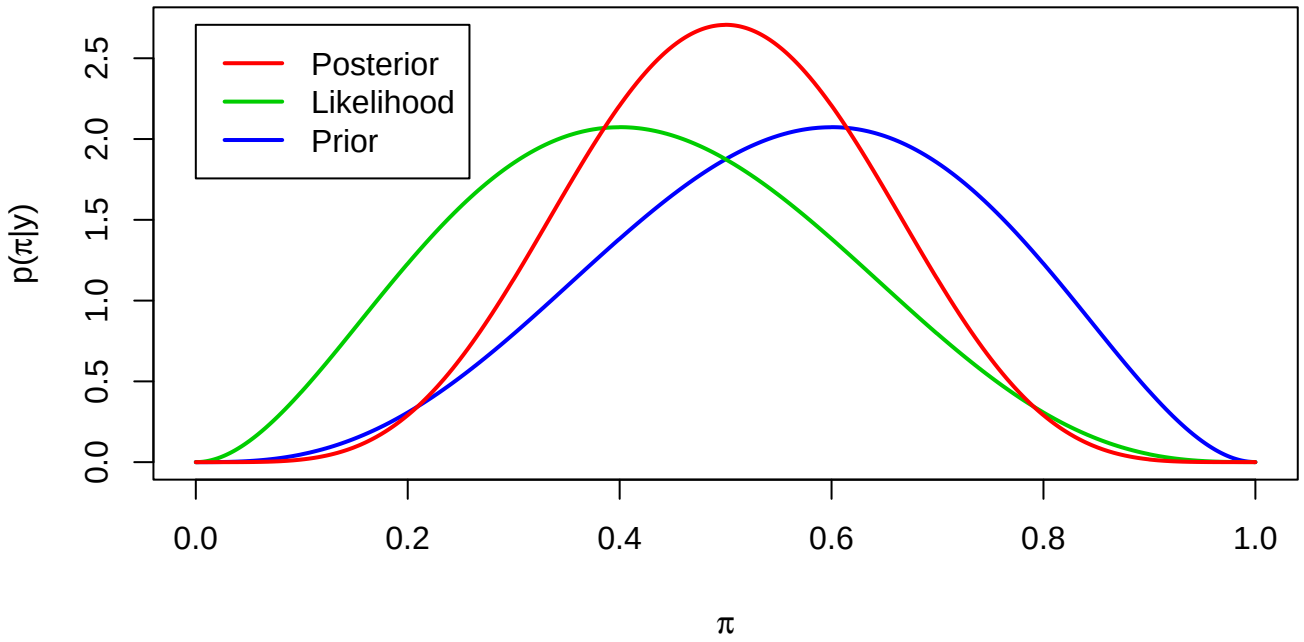
The first inequality gives

$$\tau_n^2 \leq \tau_0^2$$

The second inequality gives

$$\tau_n^2 \leq \frac{\sigma^2}{n}$$

$n = 5, y = 2, a = 4, b = 3$

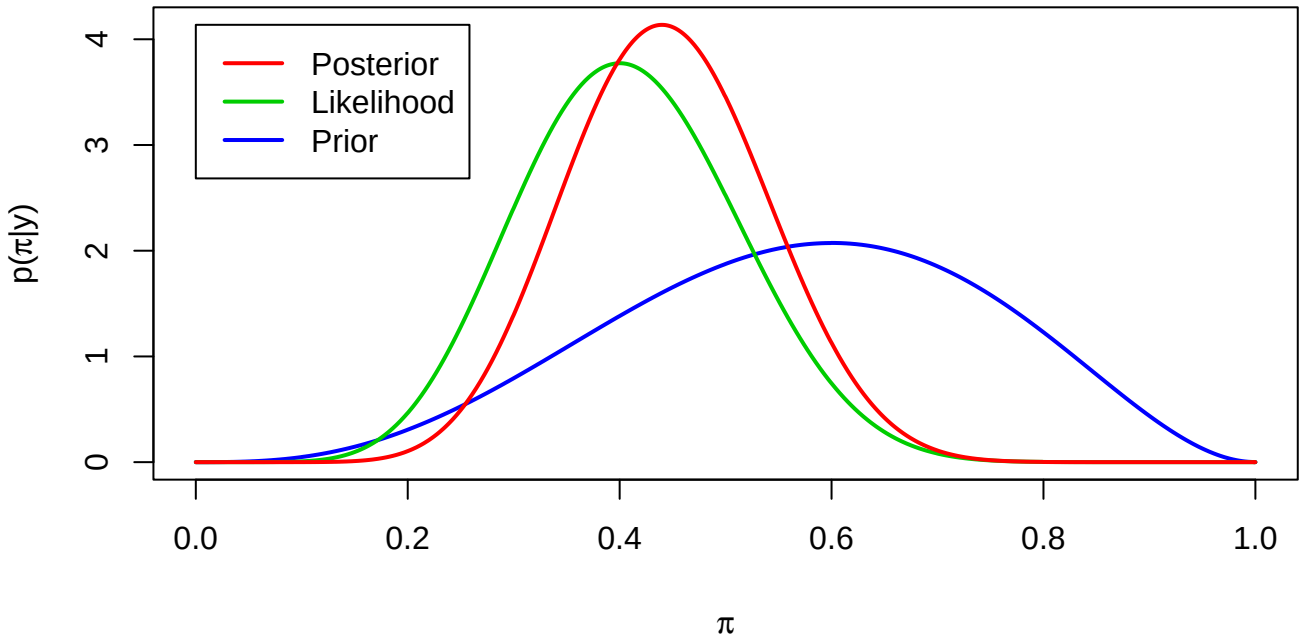


$$E[\pi] = \frac{4}{7} \qquad \hat{\pi} = \frac{2}{5} \qquad E[\pi|y] = \frac{6}{12} = 0.5$$

$$\text{Var}(\pi) = \frac{3}{98} = 0.0306 \qquad \text{Var}(\pi|y) = \frac{1}{52} = 0.0192$$

$$\text{SD}(\pi) = 0.175 \qquad \text{SD}(\pi|y) = 0.139$$

$n = 20, y = 8, a = 4, b = 3$

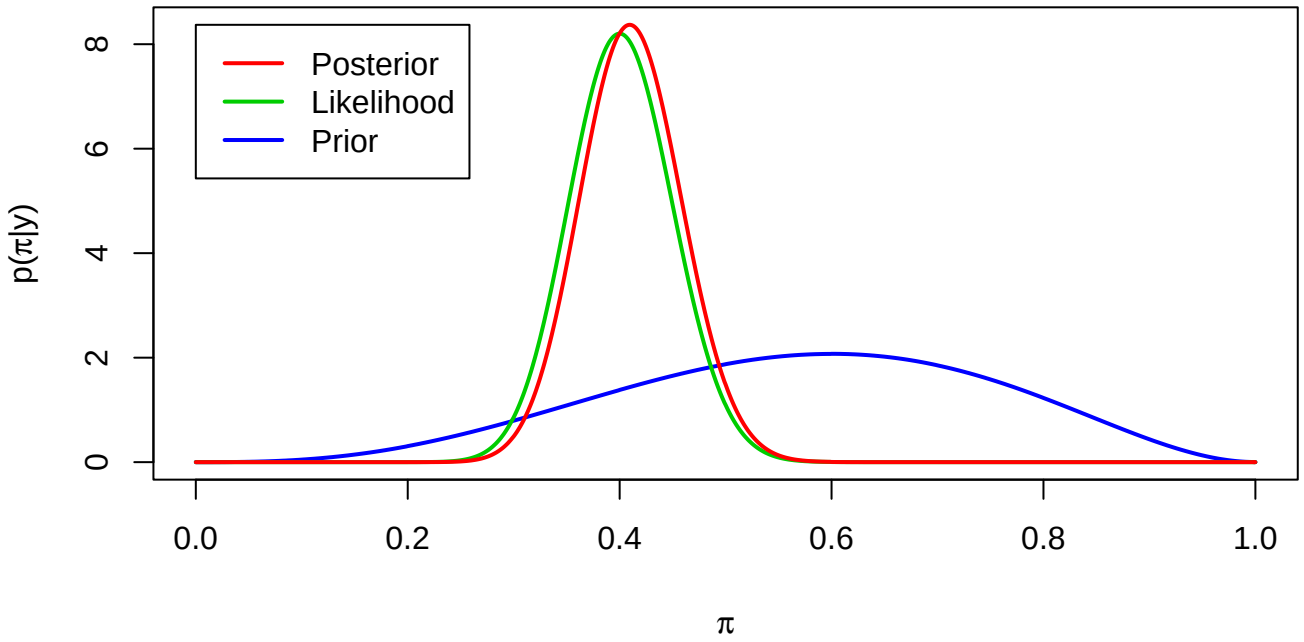


$$E[\pi] = \frac{4}{7} \qquad \hat{\pi} = \frac{2}{5} \qquad E[\pi|y] = \frac{12}{27} = 0.444$$

$$\text{Var}(\pi) = \frac{3}{98} = 0.0306 \qquad \text{Var}(\pi|y) = 0.0088$$

$$\text{SD}(\pi) = 0.175 \qquad \text{SD}(\pi|y) = 0.094$$

$n = 100, y = 40, a = 4, b = 3$

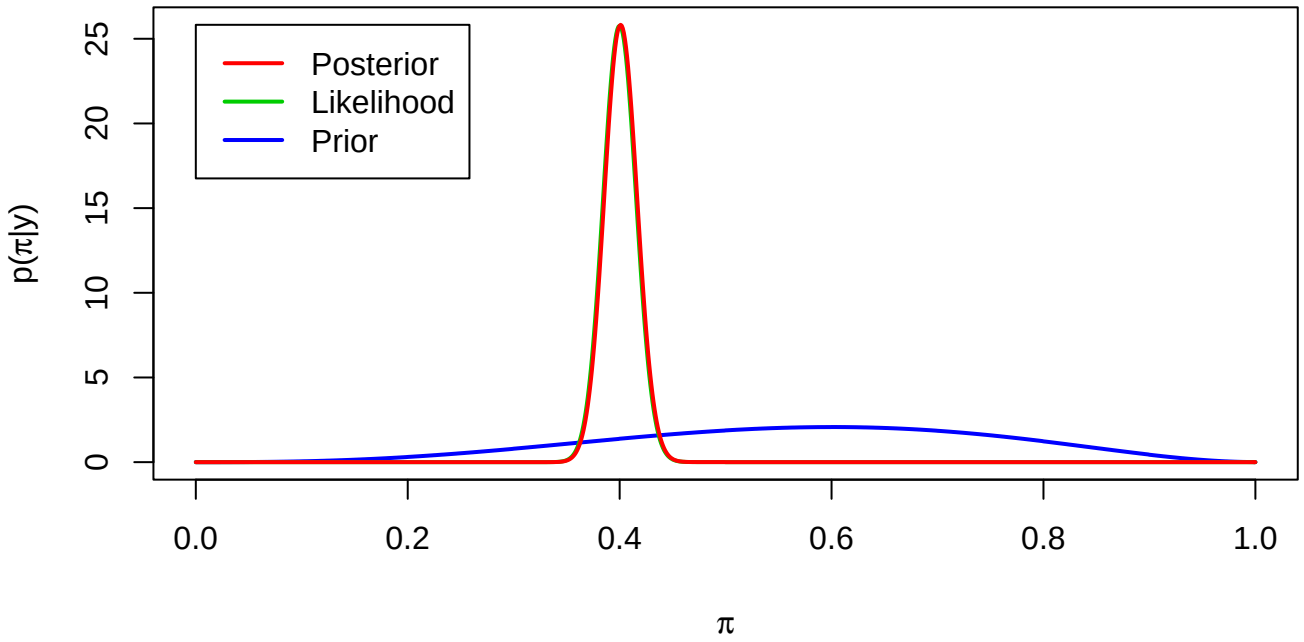


$$E[\pi] = \frac{4}{7} \qquad \hat{\pi} = \frac{2}{5} \qquad E[\pi|y] = \frac{44}{107} = 0.411$$

$$\text{Var}(\pi) = \frac{3}{98} = 0.0306 \qquad \text{Var}(\pi|y) = 0.0022$$

$$\text{SD}(\pi) = 0.175 \qquad \text{SD}(\pi|y) = 0.047$$

$n = 1000, y = 400, a = 4, b = 3$



$$E[\pi] = \frac{4}{7} \qquad \hat{\pi} = \frac{2}{5} \qquad E[\pi|y] = \frac{404}{1007} = 0.401$$

$$\text{Var}(\pi) = \frac{3}{98} = 0.0306 \qquad \text{Var}(\pi|y) = 0.0024$$

$$\text{SD}(\pi) = 0.175 \qquad \text{SD}(\pi|y) = 0.0154$$

Prediction

Another useful summary is the posterior predictive distribution of a future observation, $\tilde{y}$

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta$$

In many situations, $\tilde{y}$ will be conditionally independent of y given θ. Thus the distribution in this case reduces to

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

In many situations this can be difficult to calculate, though it is often easy with a conjugate prior.

For example, with Binomial-Beta model, the posterior distribution of the success probability is $Beta(a_1, b_1)$ (for some a_1, b_1). Then the distribution of the number of successes in m new trials is

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\pi)p(\pi|y)d\pi \\ &= \int_0^1 \binom{m}{\tilde{y}} \pi^{\tilde{y}}(1-\pi)^{m-\tilde{y}} \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} \pi^{a_1-1}(1-\pi)^{b_1-1} d\pi \\ &= \binom{m}{\tilde{y}} \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} \frac{\Gamma(a_1+\tilde{y})\Gamma(b_1+m-\tilde{y})}{\Gamma(a_1+b_1+m)} \end{aligned}$$

Which is an example of the Beta-Binomial distribution.

The mean this distribution is

$$E[\tilde{y}|y] = m\frac{a_1}{a_1+b_1} = m\tilde{\pi}$$

This can be gotten by applying

$$E[\tilde{y}|y] = E[E[\tilde{y}|\pi]|y] = E[m\pi|y]$$

The variance of this can be gotten by

$$\begin{aligned} \mathrm{Var}(\tilde{y}|y) &= \mathrm{Var}(E[\tilde{y}|\pi]|y) + E[\mathrm{Var}(\tilde{y}|\pi)|y] \\ &= \mathrm{Var}(m\pi|y) + E[m\pi(1-\pi)|y] \end{aligned}$$

This is of the form

$$m\tilde{\pi}(1-\tilde{\pi})\{1+(m-1)\tau^2\}$$

One way of thinking about this is that there is two pieces of uncertainty in predicting a new observation.

1. Uncertainty about the true success probability

2. Deviation of an observation from its expected value

This is more clear with the Normal-Normal model with fixed variance. As we saw earlier, the posterior distribution is of the form

$$\theta|y \sim N(\mu_n, \tau_n^2)$$

Then

$$p(\tilde{y}|y) = \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right) \frac{1}{\tau_n\sqrt{2\pi}} \exp\left(-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right) d\theta$$

A little bit of calculus will show that this reduces to a normal density with mean

$$E[\tilde{y}|y] = \mu_n = E[E[\tilde{y}|\mu_n]|y] = E[\theta|y]$$

and variance

$$\begin{aligned} \mathrm{Var}(\tilde{y}|y) &= \tau_n^2 + \sigma^2 \\ &= \mathrm{Var}(E[\tilde{y}|\theta]|y) + E[\mathrm{Var}(\tilde{y}|\theta)|y] \\ &= \mathrm{Var}(\mu_n|y) + E[\sigma^2|y] \end{aligned}$$

An analogue to this is the variance for prediction in linear regression. It is exactly of this form

$$\mathrm{Var}(\hat{y}|x) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}\right)$$

Simulating the posterior predictive distribution

This is easy to do, assuming that you can simulate from the posterior distribution of the parameter, which is usually feasible.

To do it involves two steps:

1. Simulate θ_i from $\theta|y; i = 1, \ldots, m$

2. Simulate $\tilde{y}_i$ from $\tilde{y}|\theta_i$ $(= \tilde{y}|\theta_i, y); i = 1, \ldots, m$

The pairs $(\theta_i, \tilde{y}_i)$ are draws from the joint distribution $\theta, \tilde{y}|y$. Therefore the $\tilde{y}_i$ are draws from $\tilde{y}|y$.

Why interest in the posterior predictive distribution?

- You might want to do predictions. For example, what will happen to a stock in 6 months.

- Model checking: Is your model reasonable?

  There are a number of ways of doing this. Future observations could be compared with the posterior predictive distribution.

  Another option might be something along the lines of cross validation. Fit the model with part of the data and compare the remaining observation to the posterior predictive distribution calculated from the sample used for fitting.

Other One Parameter Models

Poisson

Example: Prussian Cavalry Fatailities Due to Horse Kicks

10 Prussian cavalry corp were monitored for 20 years (200 Corp-Years) and the number of fatalities due to horse kicks were recorded

$x = \#$ Deaths	Number of Corp-Years with x Fatalities
0	109
1	65
2	22
3	3
4	1

Let $y_i, i = 1, \ldots, 200$ be the number of deaths in observation i.

Assume that $y_i \overset{iid}{\sim} Poisson(\theta)$. (This has been shown to be a good description for this data). Then the MLE for θ is

$$\hat{\theta} = \bar{y} = \frac{122}{200} = 0.61$$

This can be seen from

$$p(y|\theta) = \prod_{i=1}^{200} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \propto \theta^{\sum y_i} e^{-n\theta} = \theta^{n\bar{y}} e^{-n\theta}$$

Instead lets take a Bayesian approach. For a prior, lets use $\theta \sim Gamma(\alpha, \beta)$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

Note that this is a conjugate prior for θ.

The posterior density satisfies

$$p(\theta|y) \propto \theta^{n\bar{y}} e^{-n\theta} \theta^{\alpha-1} e^{-\beta\theta} = \theta^{n\bar{y}+\alpha-1} e^{-(n+\beta)\theta}$$

which is proportional to a $Gamma(\alpha + n\bar{y}, \beta + n)$ density

The mean and variance of a $Gamma(\alpha, \beta)$ are

$$E[\theta] = \frac{\alpha}{\beta} \qquad \text{Var}(\theta) = \frac{\alpha}{\beta^2}$$

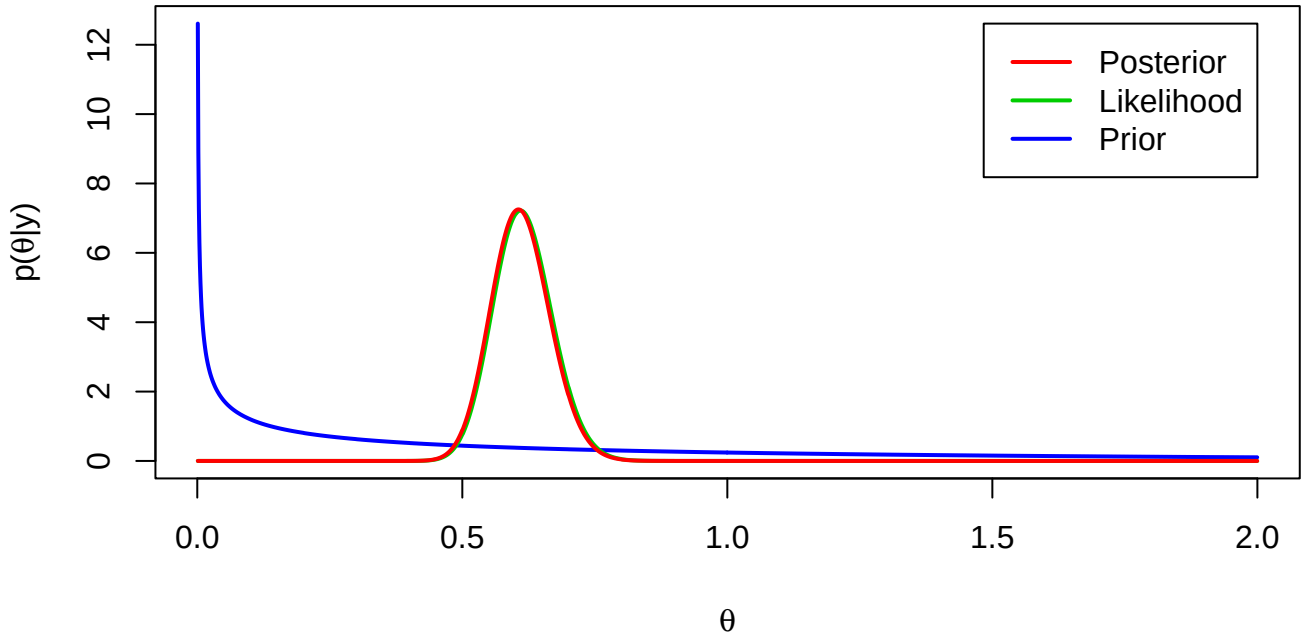
So the posterior mean and variance in this analysis are

$$E[\theta|y] = \frac{\alpha + n\bar{y}}{\beta + n} \qquad \text{Var}(\theta|y) = \frac{\alpha + n\bar{y}}{(\beta + n)^2}$$

Similarly to before, the posterior mean is a weighted average of the prior mean and the MLE (weights β and n).

Lets examine the posteriors under different prior choices

$n = 200, \bar{y} = 0.61, \alpha = \beta = 0.5$

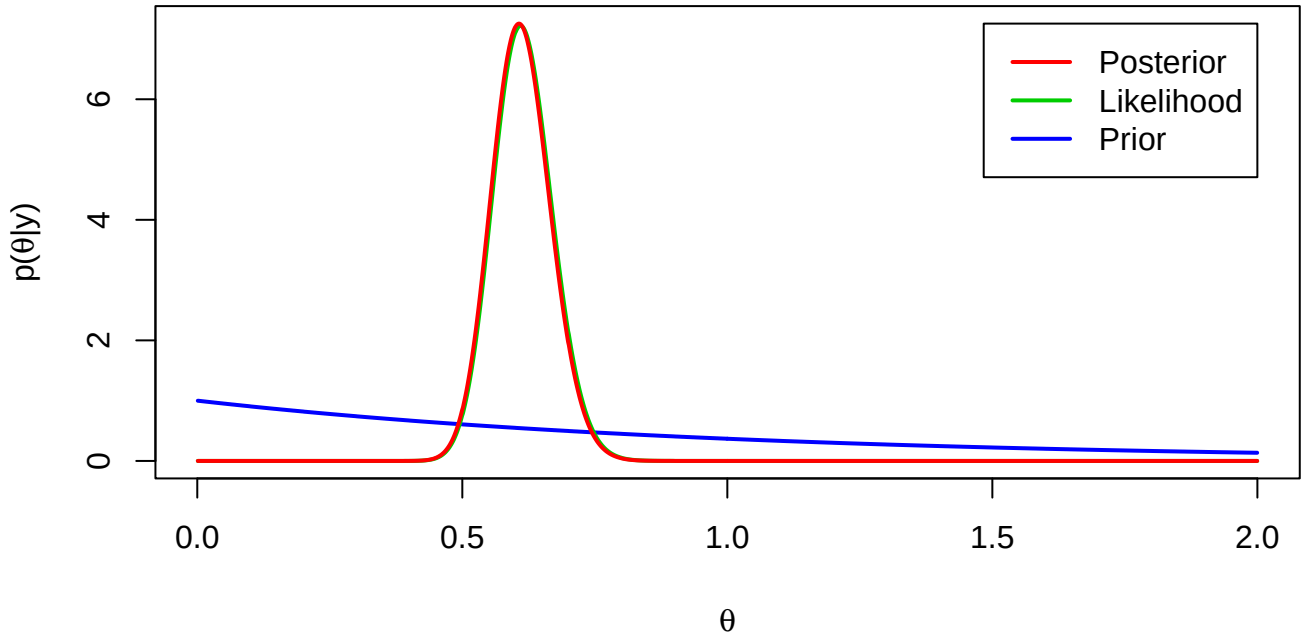


$$E[\theta] = 1 \qquad \hat{\theta} = 0.61 \qquad E[\theta|y] = 0.611$$

$$\text{Var}(\theta) = 2 \qquad \text{Var}(\theta|y) = 0.0030$$

$$\text{SD}(\theta) = 1.412 \qquad \text{SD}(\theta|y) = 0.055$$

$n = 200, \bar{y} = 0.61, \alpha = \beta = 1$

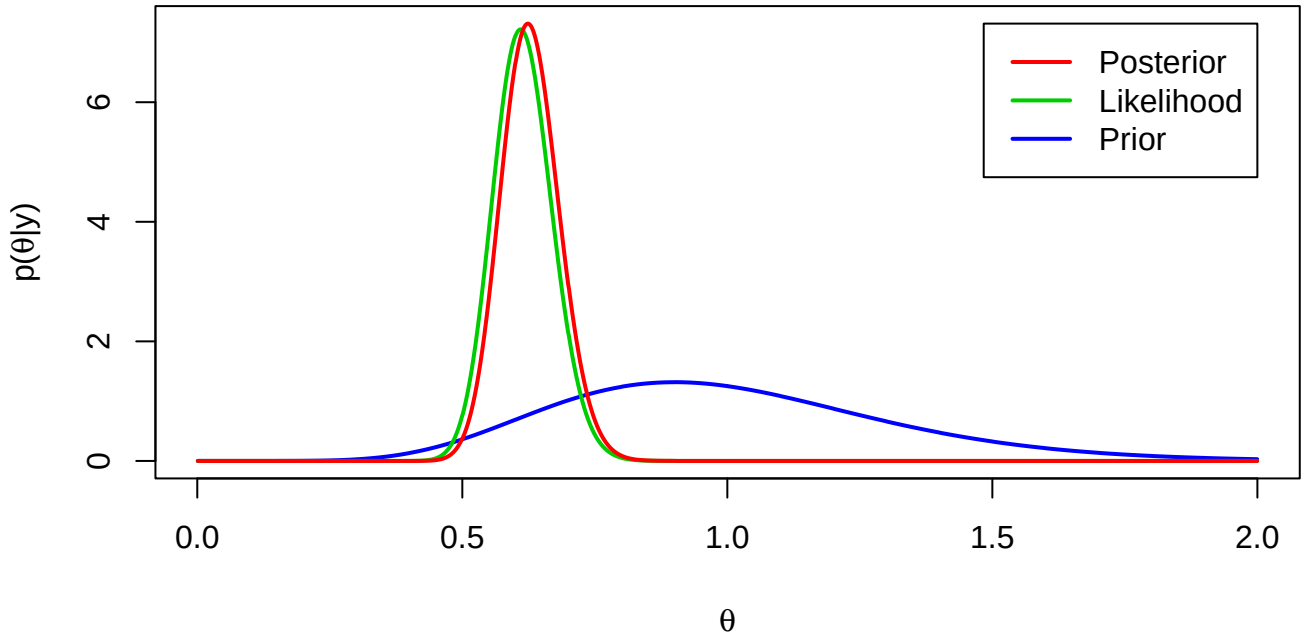


$$E[\theta] = 1 \qquad \hat{\theta} = 0.61 \qquad E[\theta|y] = 0.612$$

$$\text{Var}(\theta) = 1 \qquad \text{Var}(\theta|y) = 0.0030$$

$$\text{SD}(\theta) = 1 \qquad \text{SD}(\theta|y) = 0.055$$

$n = 200, \bar{y} = 0.61, \alpha = \beta = 10$

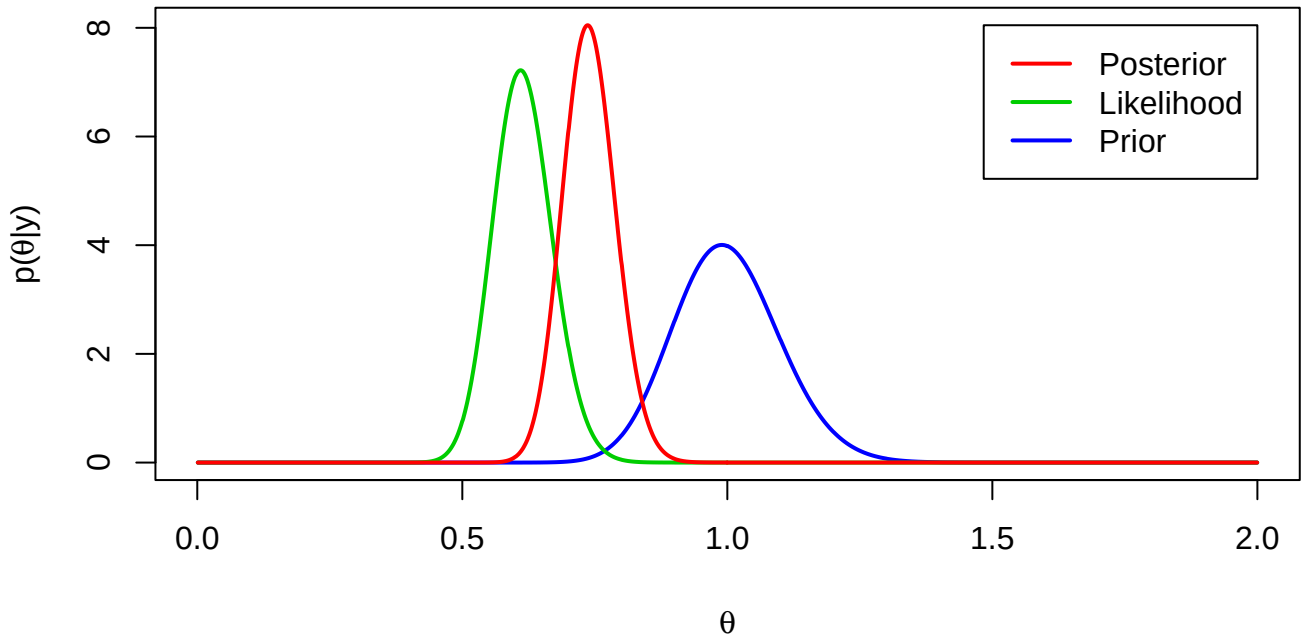


$$E[\theta] = 1 \qquad \hat{\theta} = 0.61 \qquad E[\theta|y] = 0.629$$

$$\text{Var}(\theta) = 0.1 \qquad \text{Var}(\theta|y) = 0.0030$$

$$\text{SD}(\theta) = 0.316 \qquad \text{SD}(\theta|y) = 0.055$$

$n = 200, \bar{y} = 0.61, \alpha = \beta = 100$



$$E[\theta] = 1 \qquad \hat{\theta} = 0.61 \qquad E[\theta|y] = 0.74$$

$$\text{Var}(\theta) = 0.01 \qquad \text{Var}(\theta|y) = 0.0025$$

$$\text{SD}(\theta) = 0.1 \qquad \text{SD}(\theta|y) = 0.050$$

One way to think of the gamma prior in this case is that you have a data set with β observations with and observed Poisson count of α.

Note that the Gamma distribution can be parameterized many ways.

Often the scale parameter form $\lambda = \frac{1}{\beta}$ is used.

Also it can be parameterized in terms of mean, variance, and coefficient of variation (only two are needed).

This gives some flexibility in thinking about the desired form of the prior for a particular model.

In the example, I fixed the mean at 1 and let the variance decrease.