STAT 8700 Homework 2

Brian Detweiler

Friday, September 9, 2016

1. Suppose we observe y successes in n trials where the probability of success in each trial is θ and suppose we use a $Beta(1,1)$ prior for θ. Show that the posterior mean can be written as a weighted average of the prior mean and the observed proportion of successes, meaning that the posterior mean will always fall between those two values.

Recognizing that $Beta(1,1)$ is the uniform distribution, $U(0,1)$. Since a uniform distribution must integrate to 1, then $\int p(\theta)d\theta = 1$, we have

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$
$$= \binom{n}{y}\theta^y(1-\theta)^{n-y} \cdot 1$$
$$\propto \theta^y(1-\theta)^{n-y}, 0 \leq \theta \leq 1$$
$$= c \cdot \theta^y(1-\theta)^{n-y}$$

We need to find the missing constant, c. Recognizing that the functional form of $p(\theta|y)$ is the same as $Beta(\theta|\alpha+y, \beta+n-y)$, then we have conjugacy. We now know the expected value of $Beta(\theta|\alpha+y, \beta+n-y)$ which turns out to be

$$E[\theta|y] = \frac{\alpha+y}{\alpha+y+\beta+n-y}$$
$$= \frac{\alpha+y}{\alpha+\beta+n}$$

Since we have the uniform distribution, which is a special case of the Beta, $Beta(1,1)$, then $\alpha=1, \beta=1$, and thus we have

$$E[\theta|y] = \frac{y+1}{y+1+n-y+1}$$
$$= \frac{y+1}{n+2}$$
$$= \frac{n}{n}\frac{y}{n+2} + \frac{2}{2}\frac{1}{n+2}$$
$$= \frac{n}{n+2}\frac{y}{n} + \frac{2}{n+2}\frac{1}{2}$$

Because $\frac{n}{n+2} + \frac{2}{n+2}$ sum to 1, it should be clear that this is a weighted average of $\frac{y}{n} + \frac{1}{2}$, where $\frac{y}{n}$ is the proportion of successes, and $\frac{1}{2}$ is the prior (uniform) mean.

Thus, if we let y take on extreme values, $y = 0$ and $y = n$, then we have

$$
\begin{aligned}
E[\theta|y = 0] &= \frac{n}{n+2}\frac{0}{n} + \frac{2}{n+2}\frac{1}{2} \\
&= \frac{1}{n+2} \\
E[\theta|y = n] &= \frac{n}{n+2}\frac{n}{n} + \frac{2}{n+2}\frac{1}{2} \\
&= \frac{n}{n+2} + \frac{1}{n+2} \\
&= \frac{n+1}{n+2}
\end{aligned}
$$

■

2. Suppose we observe y successes in n trials where the probability of success in each trial is θ.

(a) Prove that if we choose a $Beta(1,1)$ (Uniform) prior then the posterior variance will be smaller than the prior variance.

Continuing from problem 1, we wish to find the posterior variance, $Var(\theta|y)$. We know the prior variance, which is the variance of the uniform, $Var(\theta) = \frac{1}{12}$.

We will show that $Var(\theta|y) < \frac{1}{12}$.

$$
\begin{aligned}
Var(\theta|y) &= \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \\
&= \frac{E[\theta|y]\big[1 - E[\theta|y]\big]}{\alpha + \beta + n + 1} \\
&= \frac{E[\theta|y]\big[1 - E[\theta|y]\big]}{n + 3}
\end{aligned}
$$

Evaluated at the extremes of $y = 0$ and $y = n$, we have

$$
\begin{aligned}
Var(\theta|y = 0) &= \frac{\frac{1}{n+2}\big[1 - \frac{1}{n+2}\big]}{n + 3} \\
Var(\theta|y = n) &= \frac{\frac{n+1}{n+2}\big[1 - \frac{n+1}{n+2}\big]}{n + 3}
\end{aligned}
$$

Evaluating n at the extremes of a single sample, $n = 1$ and extremely large samples, $n \to \infty$, we find

$$
\begin{aligned}
Var(\theta|y = 0; n = 1) &= \frac{\frac{1}{3}\big[1 - \frac{1}{3}\big]}{4} \\
&= \frac{1}{18} \\
Var(\theta|y = n; n = 1) &= \frac{\frac{2}{3}\big[1 - \frac{2}{3}\big]}{4} \\
&= \frac{1}{18} \\
\lim_{n \to \infty} Var(\theta|y = 0) &= \frac{\frac{1}{n+2}\big[1 - \frac{1}{n+2}\big]}{n + 3} = 0 \\
\lim_{n \to \infty} Var(\theta|y = n) &= \frac{\frac{n+1}{n+2}\big[1 - \frac{n+1}{n+2}\big]}{n + 3} = 0
\end{aligned}
$$

So $Var(\theta|y) < Var(\theta)$ for any value of y, under any sample size.

<div align="right">■</div>

(b) Show that the above isn't necessarily the case if we choose a general $Beta(\alpha, \beta)$ prior. That is, find set of values for α, β, n, y where the above is not true.

When trying to find a proof by example, it is helpful to start at the extremes. We can see this with $Beta(2, 1)$ by letting $\alpha = 2, \beta = 1, n = 1, y = 0$. Immediately, we find a point where this is the case.

$$E[\theta|y] = \frac{\alpha + y}{\alpha + y + \beta + n - y}$$
$$= \frac{2 + y}{3 + n}$$

$$Var(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$
$$= \frac{E[\theta|y]\left[1 - E[\theta|y]\right]}{\alpha + \beta + n + 1}$$
$$= \frac{\frac{2}{4}\left[1 - \frac{2}{4}\right]}{4}$$
$$= \frac{\frac{1}{2}\left[\frac{1}{2}\right]}{4}$$
$$= \frac{1}{16} < \frac{1}{12}$$

∎

3. Suppose we wish to estimate the proportion of a voting population that support a particular ballot initiative. We choose to use a Uniform prior for the proportion of voters who support the initiative. A random sample of 100 voters is polled and 55 are in favor of the ballot initiative.

(a) Find the posterior distribution of θ.

Under the uniform prior distribution, the posterior distribution for ballot initiative support is

$$
\begin{aligned}
\theta|y &\sim Beta(y+1, n-y+1) \\
&= Beta(55+1, 100-55+1) \\
&= Beta(56, 46)
\end{aligned}
$$

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
# Note: Borrowed from
# http://www.obscureanalytics.com/2012/07/04/to-the-basics-bayesian-inference-on-a-binomial-proportion/
betaplot <- function(a,b) {
  theta = seq(0,1,0.005)
  p_theta = dbeta(theta, a, b)
  p <- qplot(theta, p_theta, geom='line')
  p <- p + theme_bw()
  p <- p + ylab(expression(paste('p(',theta,')', sep = '')))
  p <- p + xlab(expression(theta))
  return(p)
}
betaplot(56, 46)
```

■
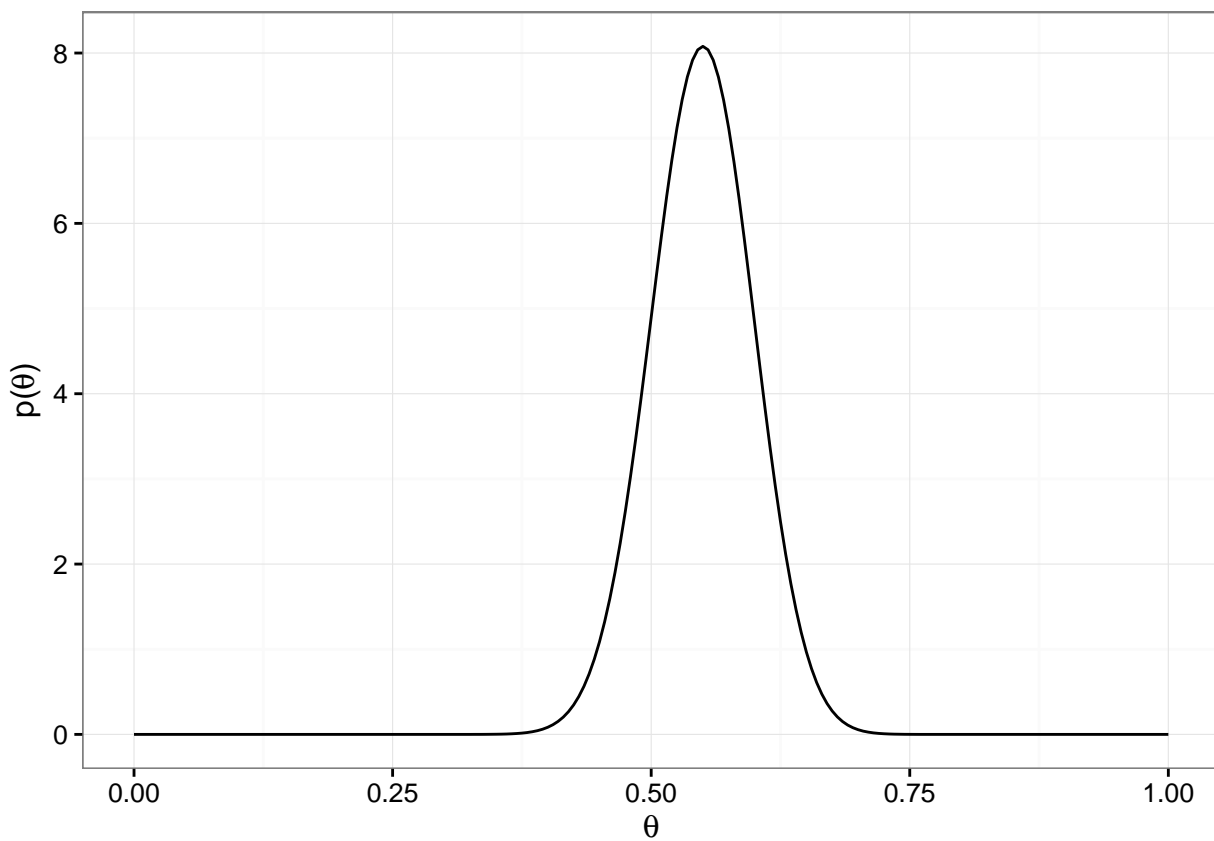
(b) What is the posterior mean and variance?

Using our findings from **1.)**, we have

$$
\begin{aligned}
E[\theta|y] &= \frac{n}{n+2}\frac{y}{n} + \frac{2}{n+2}\frac{1}{2} \\
&= \frac{55}{102} + \frac{1}{102} \\
&= \frac{56}{102} \\
&\approx 0.549
\end{aligned}
$$

and from **2.)**, we have

$$
\begin{aligned}
Var(\theta|y) &= \frac{E[\theta|y]\big[1 - E[\theta|y]\big]}{n+3} \\
&= \frac{\frac{56}{102}\big[1 - \frac{56}{102}\big]}{103} \\
&\approx 0.00240385512667
\end{aligned}
$$

$\blacksquare$

(c) The `binobp` command in the **Bolstad** package in **R** will calculate the posterior for binomial data and a beta prior. It requires 4 inputs (in order): y, n, α, β. The output includes a graph of the prior and posterior distributions. Include this graph in your assignment.

```
library(Bolstad)
```

```
## Warning: package 'Bolstad' was built under R version 3.2.4
```
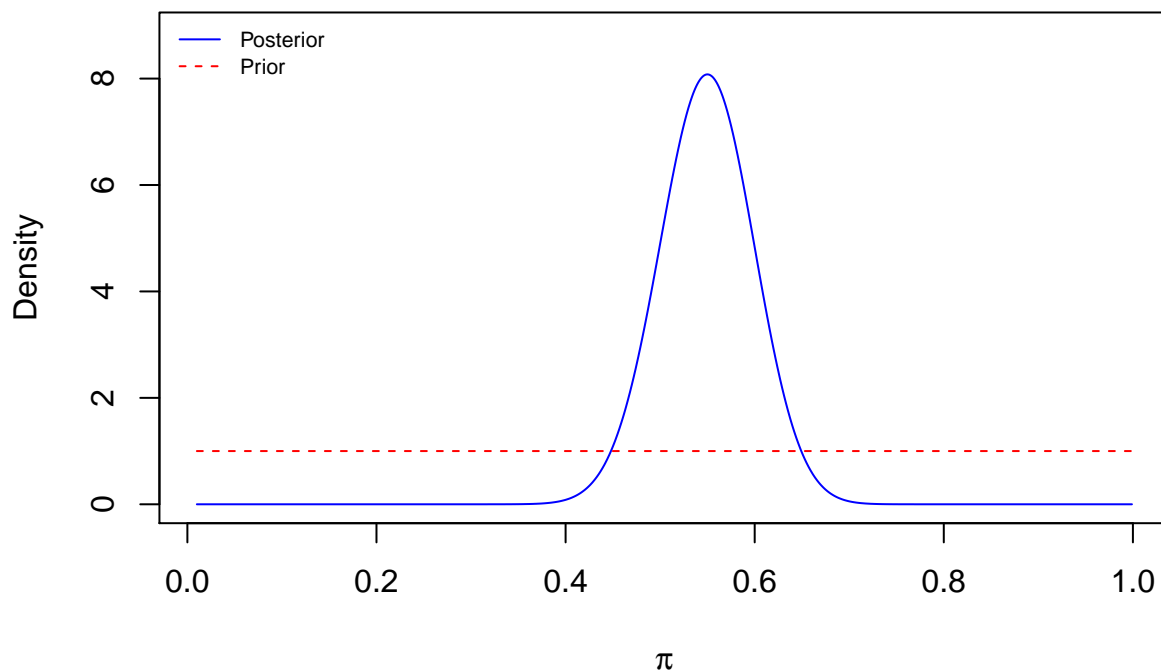
```
##
## Attaching package: 'Bolstad'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, sd, var
```

```
posterior <- binobp(x=55, n=100, a=1, b=1)
```



```
## Posterior Mean            :  0.5490196
## Posterior Variance        :  0.0024039
## Posterior Std. Deviation  :  0.0490291
##
```

```
## Prob.    Quantile
## ------   ---------
## 0.005    0.4220634
## 0.010    0.4342329
## 0.025    0.4522192
## 0.050    0.4677738
## 0.500    0.5493412
## 0.950    0.6291668
## 0.975    0.6439984
## 0.990    0.6609932
## 0.995    0.6723969
```

■

(d) The command `abline(v=location, col="colour")` adds a vertical line to a plot, where location should be replaced by the x co-ordinate of the vertical line, and colour should be replaced by the actual color. Add 3 vertical lines to your plot from the previous part: a black line representing the observed proportion of voters who support the initiative, a red line representing the prior mean, and a blue line representing the posterior mean.

```
posterior <- binobp(x=55, n=100, a=1, b=1)
```

```
## Posterior Mean            :  0.5490196
## Posterior Variance         :  0.0024039
## Posterior Std. Deviation :  0.0490291
##
## Prob.     Quantile
## ------    ---------
## 0.005     0.4220634
## 0.010     0.4342329
## 0.025     0.4522192
## 0.050     0.4677738
## 0.500     0.5493412
## 0.950     0.6291668
## 0.975     0.6439984
## 0.990     0.6609932
## 0.995     0.6723969
```

```
abline(v=55/100, col="black")
abline(v=1/2, col="red")
abline(v=posterior$mean, col="blue")
```
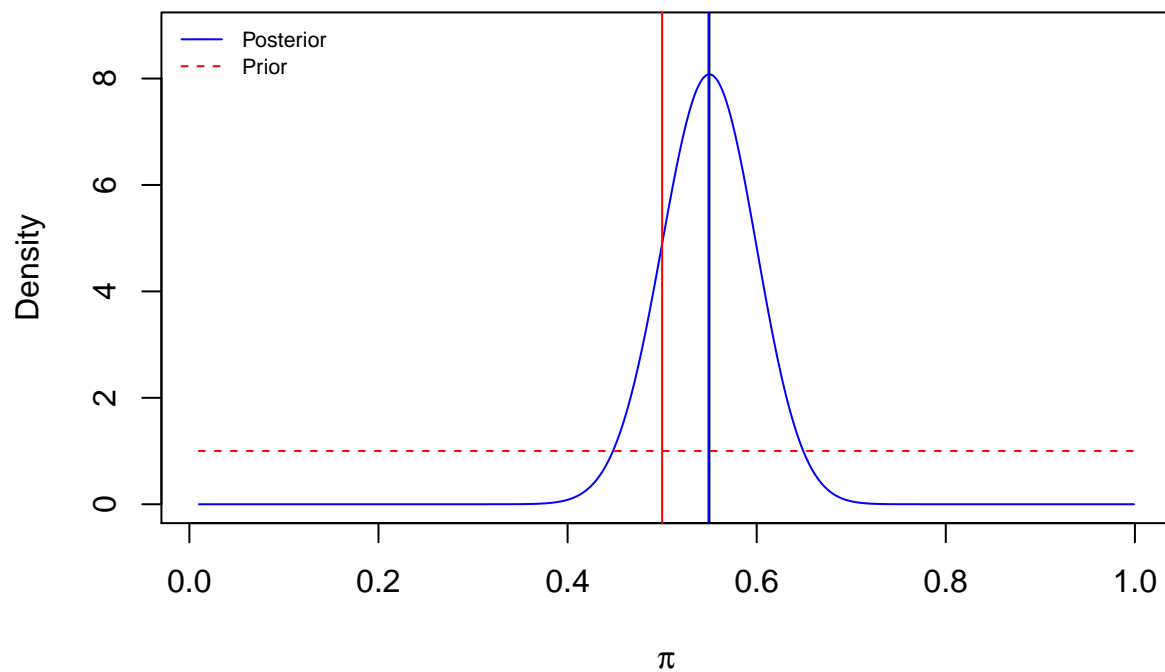
Notice, the observed proportion of successes is nearly the same as the posterior mean, so the blue and black lines are hard to distinguish. This is becaus the we have a weakly informed prior (uniform) which has a negligable impact on the observed data.

∎

(e) Also included in the output from `binobp` is a table of posterior quantiles. A 95% credible interval for the posterior distribution can be found by using the 0.025 and 0.975 quantiles. What is this 95% credible interval for your posterior distribution? (Note this interval is exactly what people wrongly assume the classical confidence interval is, that is there is a 95% chance that θ will take a value inside this interval).

The 95% CI for this distribution is obtained by taking the 0.025 and 0.975 probabilities, which results in

```
lower <- posterior$quantiles["0.025"]
upper <- posterior$quantiles["0.975"]
c(lower, upper)
```

```
##      0.025     0.975
## 0.4522192 0.6439984
```

∎

(f) What is of interest to us is whether or not the initiative will pass (that is, receive a majority of Yes votes). The R command `pbeta` computes the CDF of a beta distribution and requires 3 inputs (in order): The value where you wish to evaluate the CDF, α, β. Use this to calculate our posterior probability that the initiate will pass.

Since `pbeta` computes the CDF, we are interested in $f(X > 0.5) = 1 - f(X \leq 0.5)$.

```
1 - pbeta(q=0.5, 56, 46)
```

```
## [1] 0.8401363
```

There is an 88.6% chance the ballot will pass.

■

4. Consider the previous question. Suppose we wish to use an informative prior instead. We would like to use a Beta prior with a mean of 0.4 and a prior standard deviation of 0.1. What are the corresponding hyper-parameters of the prior distribution? Repeat all the steps of the previous question, using the new prior distribution.

The hyperparameters that most closely meet these requirements are $\alpha = 4, \beta = 6$.

$$\begin{aligned} E[\theta] &= \frac{\alpha}{\alpha + \beta} \\ &= \frac{4}{4 + 6} \\ &= 0.4 \end{aligned}$$
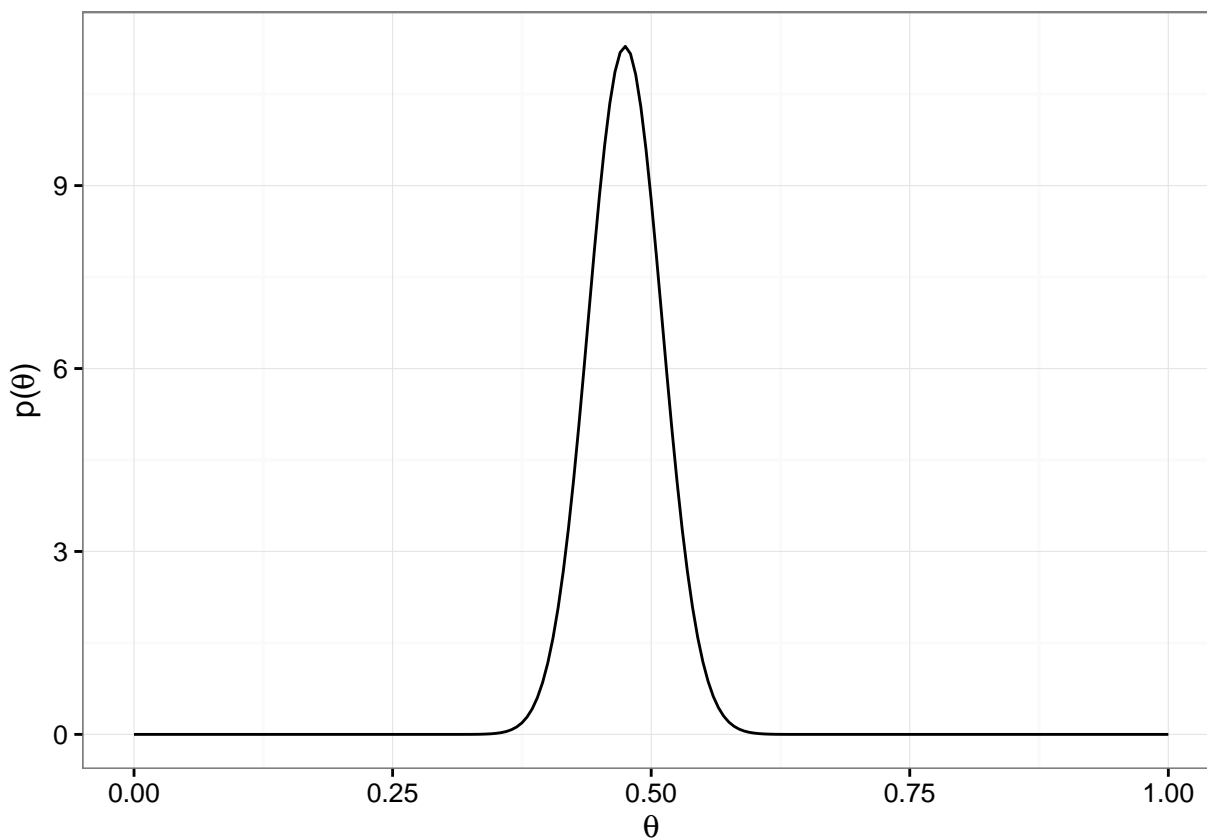
$$\begin{aligned} Var(\theta) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\ &= \frac{4 \cdot 6}{(4 + 6)^2(4 + 6 + 1)} \\ &= \frac{24}{1100} \\ &= \frac{6}{275} \\ &= 0.021818182 \end{aligned}$$

$$\begin{aligned} St.Dev &= \sqrt{0.021818182} \\ &= 0.14771 \\ &\approx 0.1 \end{aligned}$$

Under the informative Beta prior distribution, the posterior distribution for ballot initiative support is

$$\begin{aligned} \theta|y &\sim Beta(\alpha + y, \beta + n - y) \\ &= Beta(40 + 55, 60 + 100 - 55) \\ &= Beta(95, 105) \end{aligned}$$

```
betaplot(95, 105)
```

The posterior mean and variance are

$$
\begin{aligned}
E[\theta|y] &= \frac{\alpha + y}{\alpha + \beta + n} \\
&= \frac{40 + 55}{40 + 60 + 100} \\
&= \frac{95}{900} \\
&= 0.475
\end{aligned}
$$

$$
\begin{aligned}
Var(\theta|y) &= \frac{E[\theta|y]\big[1 - E[\theta|y]\big]}{\alpha + \beta + n + 1} \\
&= \frac{0.475(1 - 0.475)}{40 + 60 + 100 + 1} \\
&\approx 0.00124067
\end{aligned}
$$

Using the `Bolstad` package,

```
posterior <- binobp(x=55, n=100, a=40, b=60)
```

```
## Posterior Mean          :  0.475
## Posterior Variance       :  0.0012407
## Posterior Std. Deviation :  0.0352232
```

```
## 
## Prob.     Quantile
## ------    ---------
## 0.005     0.3851506
## 0.010     0.3936731
## 0.025     0.4062744
## 0.050     0.4171853
## 0.500     0.4749165
## 0.950     0.5330997
## 0.975     0.5441993
## 0.990     0.5570604
## 0.995     0.5657844
```

```r
abline(v=55/100, col="black")
abline(v=0.4, col="red")
abline(v=posterior$mean, col="blue")
```
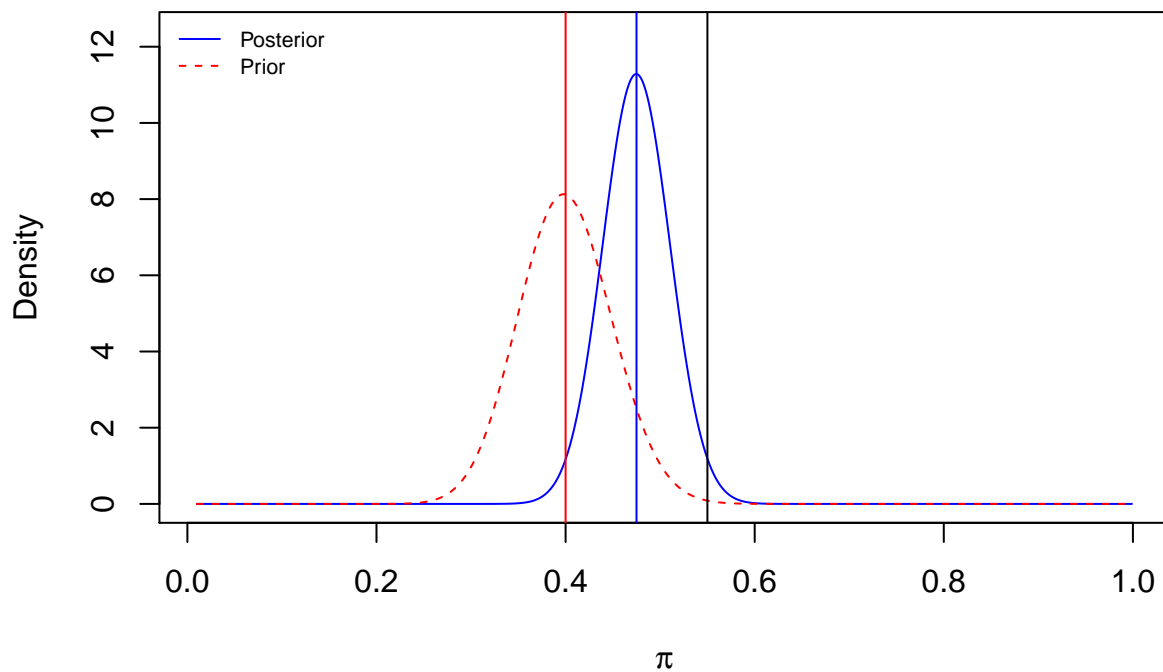

Notice how the new informative prior has shifted the posterior about half-way between its mean and the observed mean.

The 95% CI for this distribution is obtained by taking the 0.025 and 0.975 probabilities, which results in

```r
lower <- posterior$quantiles["0.025"]
upper <- posterior$quantiles["0.975"]
c(lower, upper)
```

```
##     0.025     0.975
## 0.4062744 0.5441993
```

Since `pbeta` computes the CDF, we are interested in $f(X > 0.5) = 1 - f(X \leq 0.5)$.

```
1 - pbeta(q=0.5, 95, 105)
```

```
## [1] 0.2392456
```

With our more informative prior, we can see the probability of passing is dragged down much further. We would want to make sure that we have good reason to use such a prior.

■

5. Each city bus in Omaha is numbered. Suppose that they are numbered sequentially $1, 2, \ldots, M$.

(a) If M were known, and Y represents the number of the next bus you see, find an expression for $P(Y = y|M)$. For what values of y is this valid?

If total number of busses, M is known, then we want the probability of seeing any given bus in the distribution next, any such value of which is equally likely (theoretically). Thus we have a uniform distribution of

$$P(Y = y|M) = \frac{1}{M}, \text{ for } 1 \leq y \leq M$$

∎

(b) Now suppose that M is unknown, as we assume a geometric prior distribution on M, that is that

$$p(M) = \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \text{ for } M = 1, 2, \ldots$$

Furthermore, suppose we observe a single bus, numbered 200. Find the posterior distribution of M (up to a constant of proportionality).

We know the likelihood is still discrete uniform, as we are just as likely to see bus 200 as we are any of the busses from 1 to M, so $p(M) \propto 1$. Therefore, we have

$$p(M|Y = 200) \propto P(Y = 200|M)P(M)$$
$$\propto \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \cdot 1$$
$$= \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \cdot c$$

for some constant, c.

$\blacksquare$

(c) Use software (for example Wolfram-Alpha or Maple) to find the constant of proportionality for the posterior, and thus find the posterior mean and variance.

Typing the following into WolframAlpha,

integral from 0 to 1 c(1/150)(149/150)^(M - 1) dM = 1

we solve for c, which is

$$c = 22350 \cdot log\left(\frac{150}{149}\right)$$

The posterior mean is then

$$
\begin{aligned}
E[M|Y = 200] &= \int_0^\infty M \cdot P(M|Y = 200)dM \\
&= \int_0^\infty M \cdot \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \cdot \left[22350 \cdot log\left(\frac{150}{149}\right)\right]dM \\
&= \frac{150}{log\left(\frac{150}{149}\right)} \\
&\approx 22425
\end{aligned}
$$

... which does not seem right.

The posterior variance is then

$$
\begin{aligned}
Var(M|Y = 200) &= \int_0^\infty (M - E[M|Y = 200])^2 \cdot \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \cdot \left[22350 \cdot log\left(\frac{150}{149}\right)\right]dM \\
&= \int_0^\infty (M - 22425)^2 \cdot \frac{1}{150}\left(\frac{149}{150}\right)^{M-1} \cdot \left[22350 \cdot log\left(\frac{150}{149}\right)\right]dM
\end{aligned}
$$

... which does not come out, but probably isn't right anyway.

■

(d) If we had decided to use the improper uniform prior $p(M) \propto 1$, would this have produced a proper or improper posterior distribution? Show your work.

∎