STAT 4224 HW #2

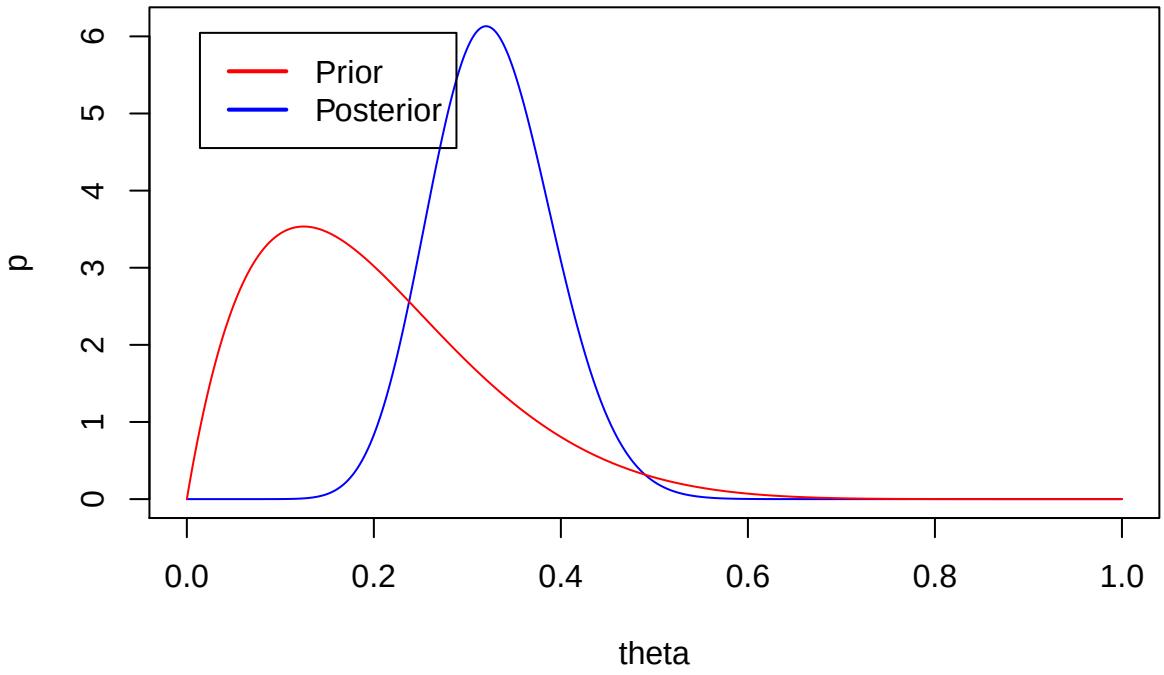
Carlyle Morgan

10/07/2021

1.

a.

```
theta<-seq(0,1,length.out = 1001)
q1aprior<-dbeta(theta,2,8)
q1aposterior<-dbeta(theta,17,35)
plot(theta,q1aposterior,type="l", col = "blue",ylab="p")
lines(theta,q1aprior,lty=1, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("red","blue"),legend=c("Prior", "Posterior"))
```



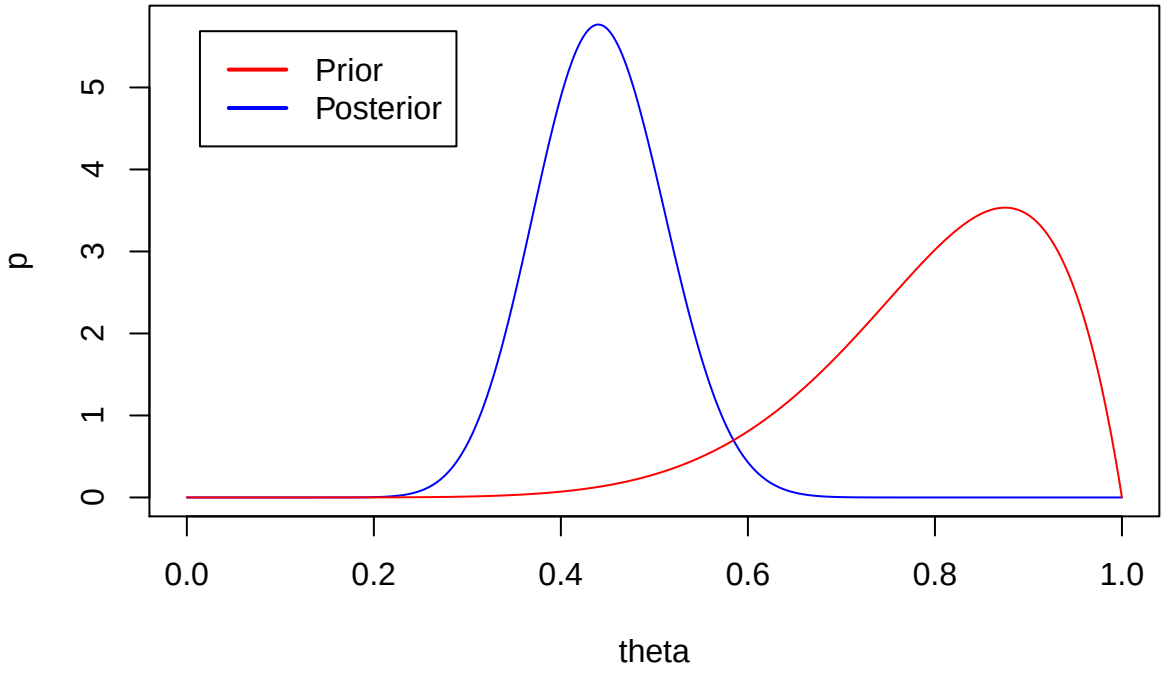

```
qbeta(c(0.025,0.25,0.5,0.75,0.975),17,35)
```

```
## [1] 0.2075833 0.2818156 0.3246886 0.3696256 0.4588729
```

The median value of the posterior is 0.325. A 50 percent confidence interval for θ is $(0.282, 0.370)$. A 95 percent confidence interval for θ is $(0.208, 0.459)$

b.

```
theta<-seq(0,1,length.out = 1001)
q1bprior<-dbeta(theta,8,2)
q1bposterior<-dbeta(theta,23,29)
plot(theta,q1bposterior,type="l", col = "blue",ylab="p")
lines(theta,q1bprior,lty=1, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("red","blue"),legend=c("Prior", "Posterior"))
```




```
qbeta(c(0.025,0.25,0.5,0.75,0.975),23,29)
```

```
## [1] 0.3112750 0.3953396 0.4415624 0.4884735 0.5775468
```

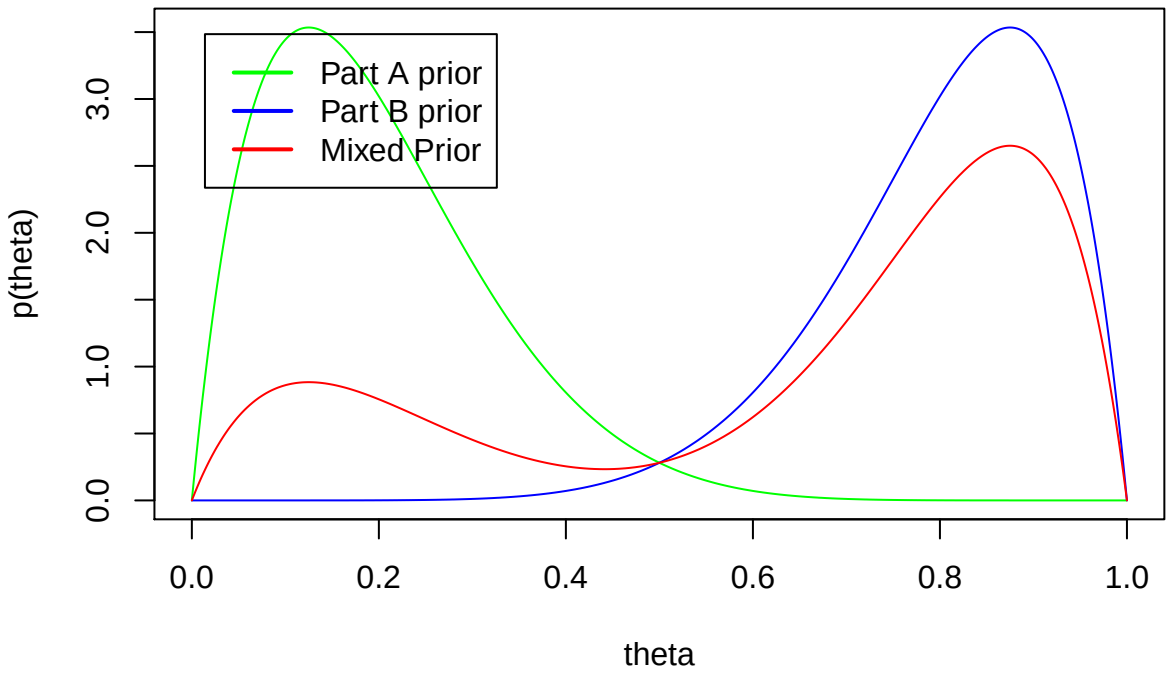
The median value of the posterior is 0.4416. A 50 percent confidence interval for θ is $(0.3953, 0.4885)$. A 95 percent confidence interval for θ is $(0.3113, 0.5775)$

c.

This mixed distribution represents the idea that we are 75% confident that the distribution of θ has a $\beta(8, 2)$ distribution and 25% confident θ has a beta distribution.

```
prior3<-function(theta){
  0.25*gamma(10)/(gamma(2)*gamma(8)) * ((theta * (1-theta)^7)+3*theta^7*(1-theta))
}
```

```r
plot(theta,q1aprior,type = "l", col = "green", ylab = "p(theta)")
lines(theta,q1bprior,lty=1, col = "blue")
lines(theta,prior3(theta), lty=1, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("green","blue","red"),legend=c("Part A prior", "Part B prior"
```



d.

i.

$$p(\theta) \times p(y|\theta) = (\frac{\Gamma(10)}{4\Gamma(2)\Gamma(8)}\theta(1-\theta)^7 + 3\theta^7(1-\theta)) \times \binom{42}{15}\theta^{15}(1-\theta)^{27}$$

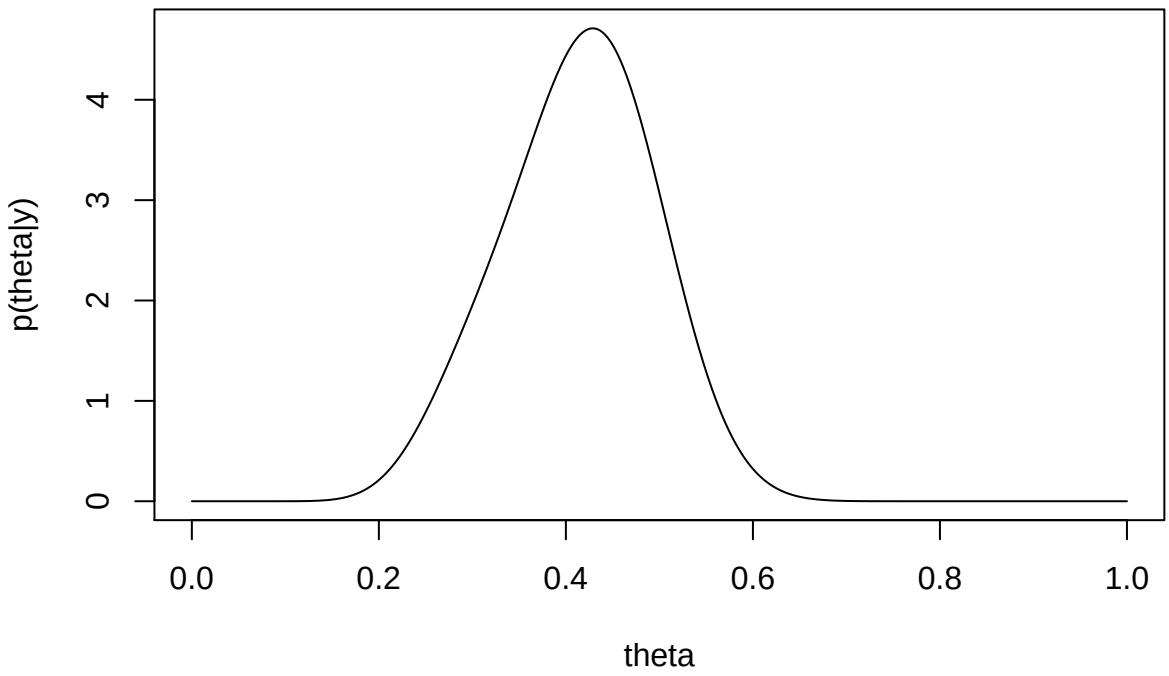
$$\Rightarrow p(\theta) \times p(y|\theta) \propto c(\theta^{16}(1-\theta)^{34} + 3\theta^{22}(1-\theta)^{28})$$

ii.

This is a mixture of a $\beta(17, 35)$ distribution with a $\beta(23, 29)$ distribution,

iii.

```r
mixpost<-function(theta){
  0.25*dbeta(theta,17,35) + 0.75*dbeta(theta,23,29)
}
plot(theta,mixpost(theta), type="l", xlim  = c(0,1), ylab = "p(theta|y)")
```



iv.

To calculate the confidence interval properly, the weights on the distributions that compose the mixture must be properly normalized to sum to 1. Using the calculations shown in part e:

```
0.1427 * qbeta(c(0.025),23,29) + 0.8573 * qbeta(c(0.025),17,35)
```

```
## [1] 0.2223801
```

```
0.1427 * qbeta(c(0.975),23,29) + 0.8573 * qbeta(c(0.975),17,35)
```

```
## [1] 0.4758077
```

Thus, a 95% CI for θ would be $[0.2224, 0.4758]$. Both the upper and lower bounds for this confidence interval are closer to that of the $\beta(2, 8)$ prior than that of the $\beta(8, 2)$ prior. This makes sense as the posterior derivative from $\beta(2, 8)$ is more heavily weighted.

e.

By di,

$$p(\theta) \times p(y|\theta) = (\frac{\Gamma(10)}{4\Gamma(2)\Gamma(8)}\theta(1-\theta)^7 + 3\theta^7(1-\theta)) \times \binom{42}{15}\theta^{15}(1-\theta)^{27}$$
$$\propto \Gamma(17)\Gamma(35)\beta(17, 35) + 3\Gamma(23)\Gamma(29)\beta(23, 29)$$

Normalizing the constant, we get:

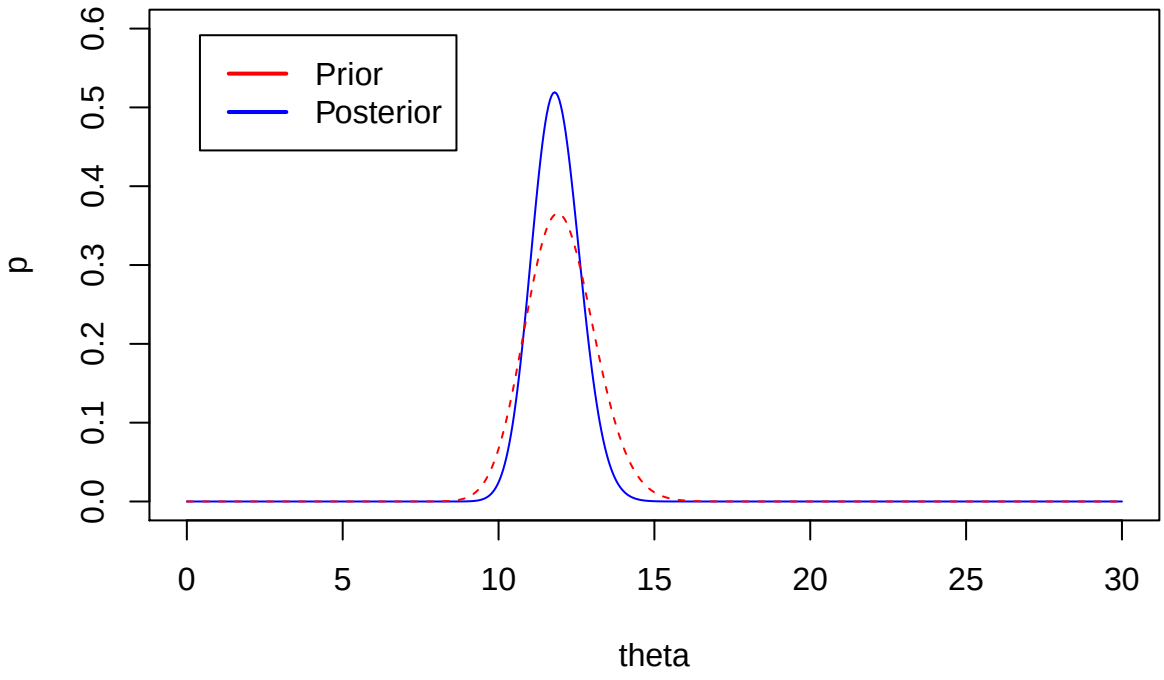
$$p(\theta|y) \approx 0.8573 \times \beta(17, 35) + 0.1427 \times \beta(23, 29)$$

From this, we can conclude that the observed data reinforced the idea that the true distribution of theta is more likely $\beta(17, 35)$ than $\beta(23, 29)$, as we had previously assumed that there was about a 75% chance it was the former and 25% chance the latter, but now the data is telling us that its more likely 85.7% vs 14.3%. These weights are also completely influenced by initial certainty about one distribution or the other, so if we had been completely off the mark to begin with(saying $\beta(23, 29)$ was more likely than $\beta(17, 35)$ for example) we would get much lower weights for $\beta(17, 35)$ and vice versa.

2.

a.

```
theta <- seq(0, 30, .05)
a_data<-c(12,9,12,14,13,13,15,8,15,6)
b_data<-c(11,11,10,9,9,8,7,10,6,8,8,9,7)

q2aprior<-dgamma(theta,120,10)
q2aposterior<-dgamma(theta,120+sum(a_data),10+length(a_data))

plot(theta,q2aposterior,type="l", col = "blue",ylab="p", ylim = c(0,0.6))
lines(theta,q2aprior,lty=2, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("red","blue"),legend=c("Prior", "Posterior"))
```



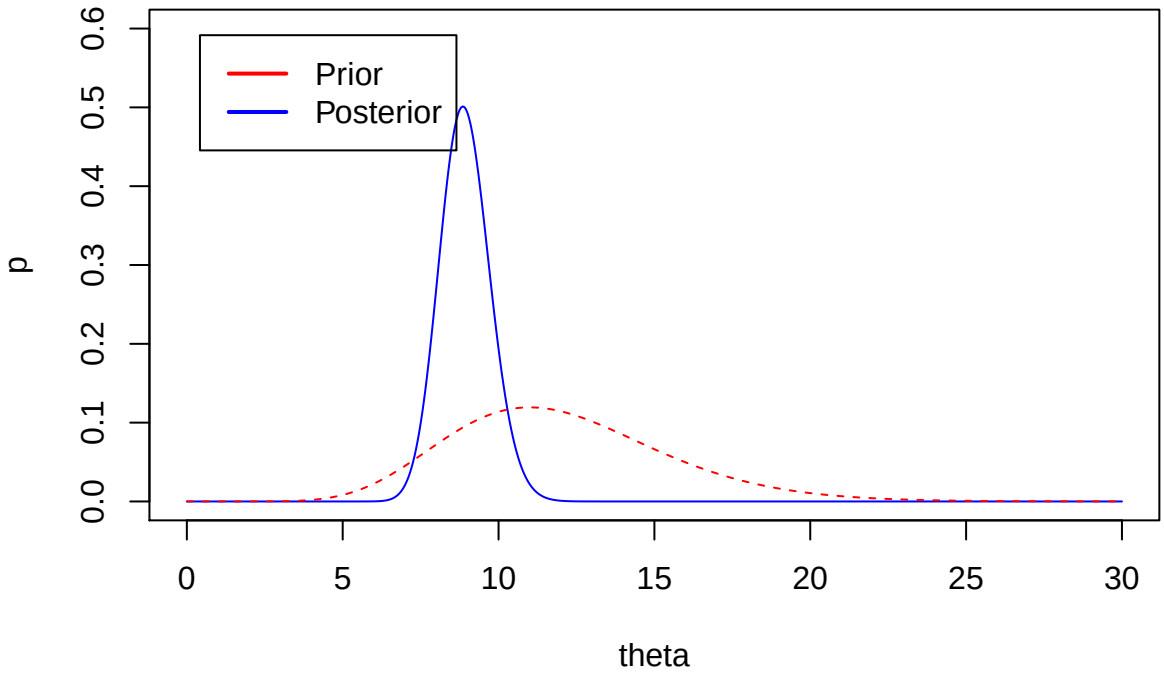

```
qgamma(c(0.025,0.25,0.5,0.75,0.975),120+sum(a_data),10+length(a_data))
```

```
## [1] 10.38924 11.32214 11.83334 12.35970 13.40545
```

The median value of the posterior is 11.833. A 50 percent confidence interval for θ is $(11.322, 12.360)$. A 95 percent confidence interval for θ is $(10.389, 13.405)$

b.

```
q2bprior<-dgamma(theta,12,1)
q2bposterior<-dgamma(theta,12+sum(b_data),1+length(b_data))

plot(theta,q2bposterior,type="l", col = "blue",ylab="p", ylim = c(0,0.6))
lines(theta,q2bprior,lty=2, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("red","blue"),legend=c("Prior", "Posterior"))
```



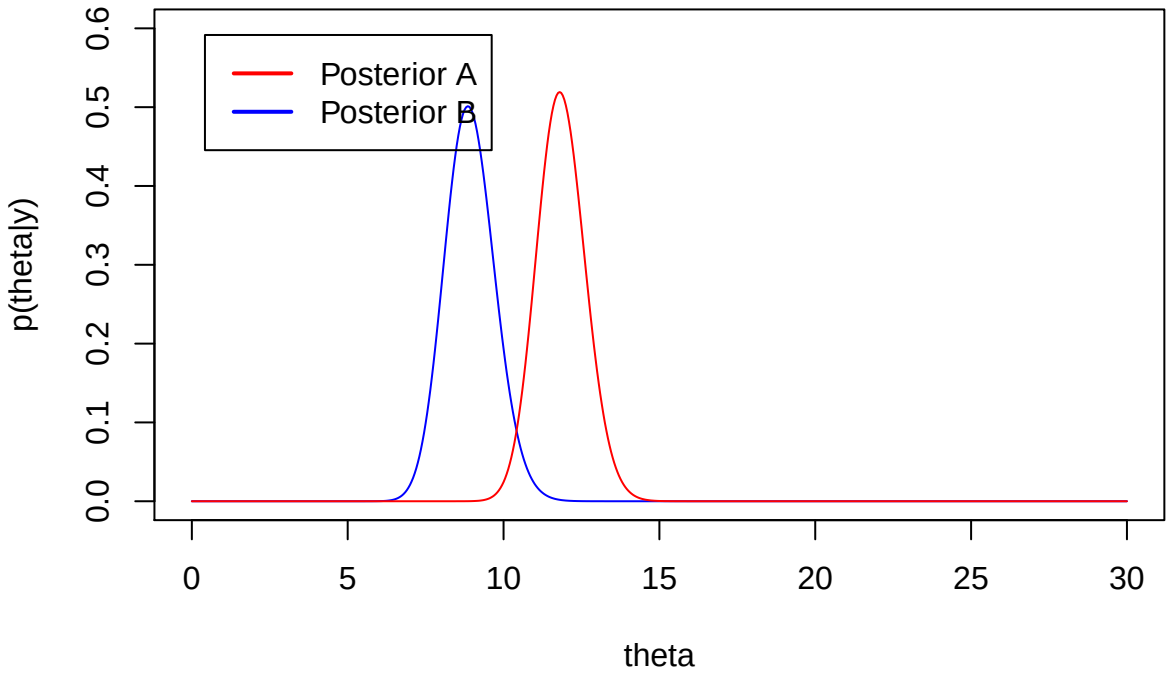

```
qgamma(c(0.025,0.25,0.5,0.75,0.975),12+sum(b_data),1+length(b_data))
```

```
## [1]  7.432064  8.377741  8.904773  9.453463 10.560308
```

The median value of the posterior is 8.905. A 50 percent confidence interval for θ is $(8.378, 9.453)$. A 95 percent confidence interval for θ is $(7.432, 10.560)$

c.

```
plot(theta,q2bposterior,type="l", col = "blue",ylab="p(theta|y)", ylim = c(0,0.6))
lines(theta,q2aposterior,lty=1, col = "red")
legend("topleft", inset=.05, lwd=2, col=c("red","blue"),legend=c("Posterior A", "Posterior B"))
```




```
1-pgamma(12,120+sum(a_data),10+length(a_data))
```

```
## [1] 0.4146161
```

```
1-pgamma(12,12+sum(b_data),1+length(b_data))
```

```
## [1] 0.0002276658
```

According to their respective posterior distributions, there should be only a 41% percent chance that $\theta_A > 12$ and a 0.02%(essentially zero) chance that $\theta_B > 12$.

d.

```
size_a<-120+sum(a_data)
mu_a<- size_a/(10+length(a_data))

size_b<-12+sum(b_data)
mu_b<- size_b/(1+length(b_data))

1-pnbinom(12, size=size_a, mu=mu_a)
```

```
## [1] 0.407275
```

```
1-pnbinom(12, size=size_b, mu=mu_b)
```

```
## [1] 0.1265673
```

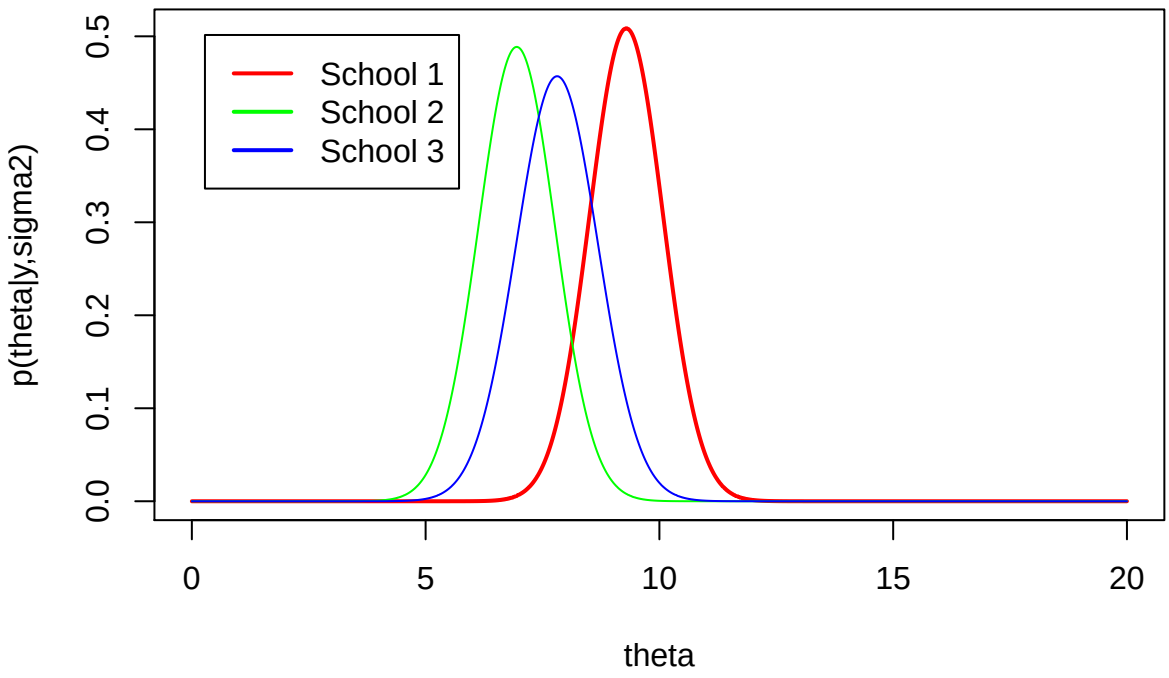
According to their respective posterior predictive distributions, there is probability 0.407 that the next observation of type A mice will have 12 or more tumors and a 0.127 probability that the next observation of

type B mice will have 12 or more tumors.

3.

a.

```
y1 <- scan("http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school1.dat")
y2 <- scan("http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school2.dat")
y3 <- scan("http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school3.dat")
theta<-seq(0,20,.001)
primean<-5
prisd<-4

y1bar <- mean(y1)
y2bar <- mean(y2)
y3bar <- mean(y3)

n1<-length(y1)
n2<-length(y2)
n3<-length(y3)

mu.n1 <- (primean/prisd^2 + n1*y1bar/16) / (1/prisd^2 + n1/16)
mu.n2 <- (primean/prisd^2 + n2*y2bar/16) / (1/prisd^2 + n2/16)
mu.n3 <- (primean/prisd^2 + n3*y3bar/16) / (1/prisd^2 + n3/16)

tau2.n1 <- 1 / (1/prisd^2 + n1/16)
tau2.n2 <- 1 / (1/prisd^2 + n2/16)
tau2.n3 <- 1 / (1/prisd^2 + n3/16)

plot(theta, dnorm(theta, mean=mu.n1, sd=sqrt(tau2.n1)),type="l", lwd=2, ylab="p(theta|y,sigma2)", col =
lines(theta, dnorm(theta, mean=mu.n2, sd=sqrt(tau2.n2)), col = "green")
lines(theta, dnorm(theta, mean=mu.n3, sd=sqrt(tau2.n3)), col = "blue")
legend("topleft", inset=.05, lwd=2, col=c("red","green","blue"),legend=c("School 1", "School 2", "Schoo
```



b.

```
qnorm(c(0.025,0.25,0.5,0.75,0.975),mean=mu.n1, sd=sqrt(tau2.n1))
```

```
## [1]  7.754785  8.763194  9.292308  9.821421 10.829830
```

```
qnorm(c(0.025,0.25,0.5,0.75,0.975),mean=mu.n2, sd=sqrt(tau2.n2))
```

```
## [1] 5.348446 6.398031 6.948750 7.499469 8.549054
```

```
qnorm(c(0.025,0.25,0.5,0.75,0.975),mean=mu.n3, sd=sqrt(tau2.n3))
```

```
## [1] 6.101584 7.223638 7.812381 8.401124 9.523178
```

The median value of θ_1 is 9.292. A 50 percent confidence interval for θ_1 is $(8.763, 9.821)$. A 95 percent confidence interval for θ_1 is $(7.755, 10.830)$. The median value of θ_2 is 6.949. A 50 percent confidence interval for θ_2 is $(6.398, 7.499)$. A 95 percent confidence interval for θ_2 is $(5.348, 8.549)$. The median value of θ_3 is 7.812. A 50 percent confidence interval for θ_3 is $(7.224, 8.401)$. A 95 percent confidence interval for θ_3 is $(6.102, 9.523)$.

c.

This is equivalent to finding $Pr(\theta_1 - \frac{\theta_2+\theta_3}{2} > 0)$, with $\theta_1 - \frac{\theta_2+\theta_3}{2}$ having a normal distribution.

```
1-pnorm(0, mean = (mu.n1-((mu.n2+mu.n3)/2)), sd = sqrt(tau2.n1+((tau2.n2+tau2.n3)/4)))
```

```
## [1] 0.9737223
```

There is probability 0.975 that the mean of school 1 is greater than the average of school 2 and school 3.

d.

The posterior predictive distributions of study time at each of the three schools is normal, and so a linear combonation of these PPDs is also normal.

```
1-pnorm(0, mean = (mu.n1-((mu.n2+mu.n3)/2)), sd = sqrt((tau2.n1+var(y1)) +(((tau2.n2+var(y2))+(tau2.n3+
```

```
## [1] 0.6498349
```

There is probability 0.65 that a randomly selected student of school 1 is greater than the average of randomly selected students from school 2 and school 3.

4.

a.

By the information given, $L(y) = 0.8y - \Phi^{-1}(0.75)\sqrt{0.8}, U(y) = 0.8y + \Phi^{-1}(0.75)\sqrt{0.8}$

$$\Rightarrow Pr(L(y) < \theta_0 < U(y)|\theta = \theta_0) = Pr(0.8y - \Phi^{-1}(0.75)\sqrt{0.8} < \theta_0 < 0.8y + \Phi^{-1}(0.75)\sqrt{0.8})$$

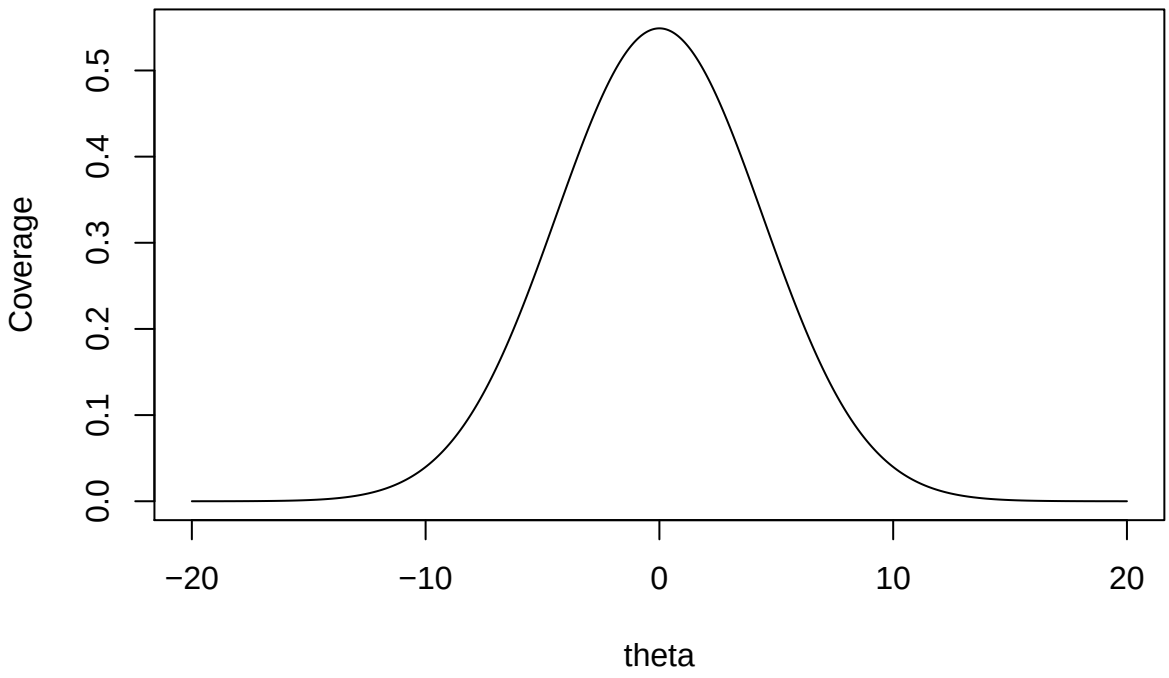
$$= Pr(\frac{\theta_0 - \Phi^{-1}(0.75)\sqrt{0.8}}{0.8} < y < \frac{\theta_0 + \Phi^{-1}(0.75)\sqrt{0.8}}{0.8})$$

$$= \Phi(\frac{\theta_0 + \Phi^{-1}(0.75)\sqrt{0.8}}{0.8} - \theta_0) - \Phi(\frac{\theta_0 - \Phi^{-1}(0.75)\sqrt{0.8}}{0.8} - \theta_0)$$

For $\theta_0 = 1$, $Pr(L(y) < 1 < U(y)|\theta = 1) = 0.535$

b.

```
postinterval<-function(theta0){
  pnorm((theta0+(0.674*sqrt(0.8)))/0.8 - theta0)-pnorm((theta0-(0.674*sqrt(0.8)))/0.8 - theta0)
}

theta<-seq(-20,20,0.01)
plot(theta, postinterval(theta), type = "l", ylab = "Coverage", xlab = "theta")
```



c.

```
intpostinterval<-function(theta0){
  postinterval(theta0)*dnorm(theta0,0,2)
}
integrate(intpostinterval,lower=-Inf,upper = Inf)
```

```
## 0.4996887 with absolute error < 2.4e-06
```

Integration via R yields that $Pr(L(y) < \theta_0 < U(y) = 0.5$. This makes sense as:

$$Pr(L(y) < \theta_0 < U(y) = \int Pr(L(y) < \theta_0 < U(y) | \theta = y) p(y) dy$$

Definition of $U(y)$ and $L(y)$ $\Rightarrow = 0.5 \int p(y) dy$

Definition of p $\Rightarrow = 0.5$.