Midterm 03 - STAT440

Joseph Sepich (jps6444)

11/20/2020

```
set.seed(42)
```

Problem 1 Bernoulli Bayes

Part a: Likelihood Parameters

The Bernoulli likelihood takes one parameter often referred to as p. This parameter can take a value between 0 and 1, which is often referred to as the probability of a successful event. Note the likelihood function:

$$f(X_i|\theta) = \Pi_{i=1}^{n} \theta^{X_i}(1-\theta)^{(1-X_i)}$$

Part b: Conjugate Prior

The conjugate prior to a Bernoulli likelihood is the Beta distribution. The domain of this distribution lies between 0 and 1. This distribution takes two parameters α and β. These two parameters help to shape the distribution. Since these parameters shape the distribution that lie between 0 and 1 (which is the values our parameter can take on), the parameters dictate the distribution of the parameter θ or p. Higher values of alpha creates a decaying tail towards zero and lower values (less than 1) create a growing tail toward zero. The opposite occurs near 1 for beta. In this way a high alpha would mean our parameter is more likely to be higher (towards 1), while the same value of alpha and beta would create a symmetric distribution for our parameter.

Part c: Posterior Parameters

The posterior parameters based off the data and prior are as follows:

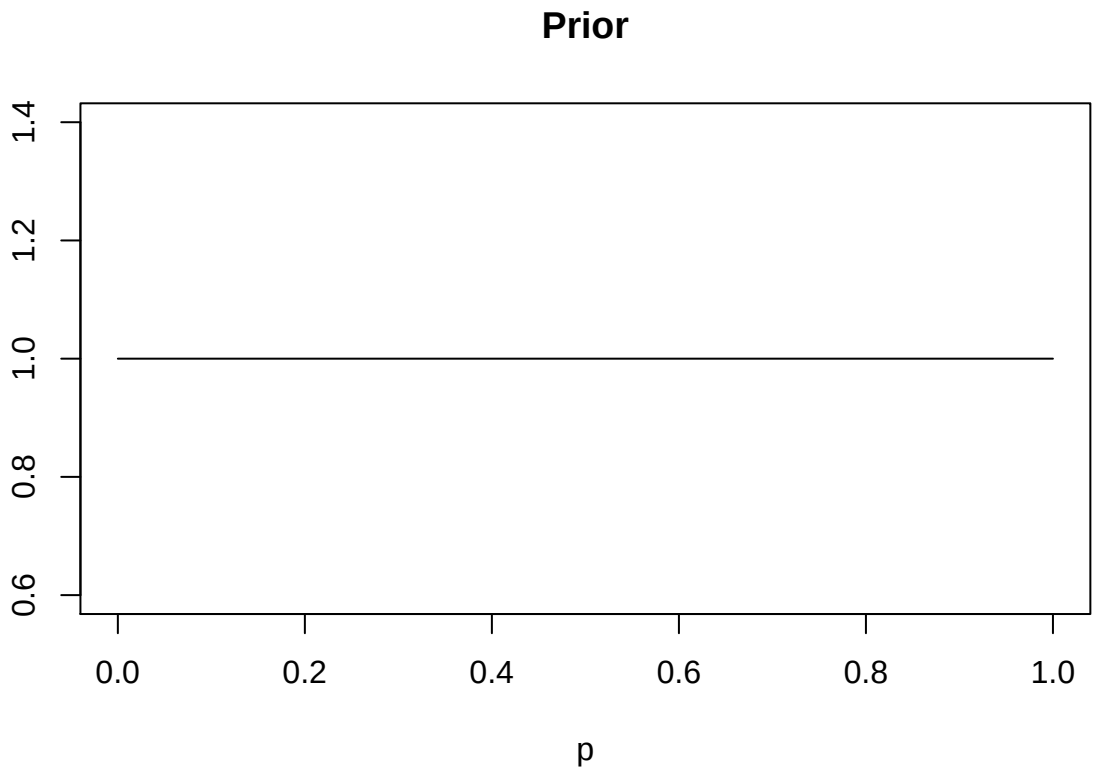
$$\alpha \to \alpha + n\overline{X}$$
$$\beta \to \beta + n - n\overline{X}$$

Both parameters depend on the term that is the summation of the events. α will become larger if there are many successes, which makes the beta distribution show that it is more likely for the success rate to be higher. β will become larger if there are many failures, which makes the beta distribution show that it is more likely for the success rate to be lower.

Part d: Plots

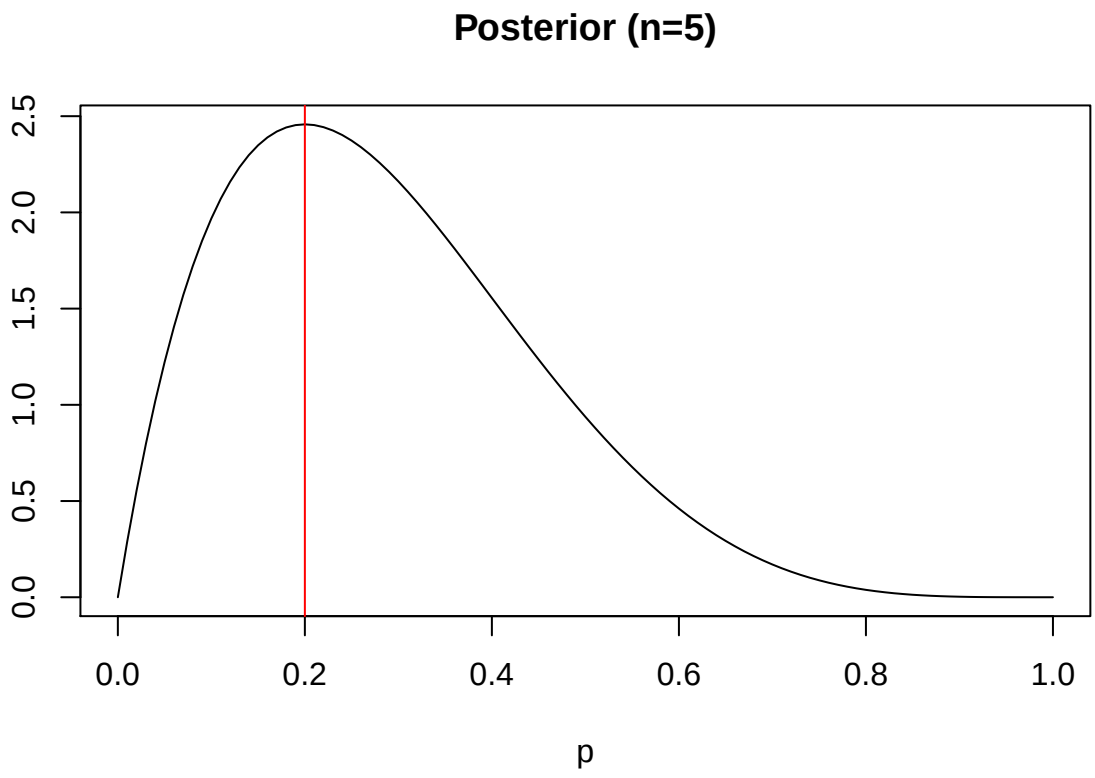
Below is our prior distribution with initial values of 1 each to show a initial belief that any parameter values from 0 to 1 is equally likely.

```
alpha_prior <- 1
beta_prior <- 1

curve(dbeta(x,alpha_prior,beta_prior),from=0,to=1,xlab="p",ylab="",main='Prior')
```



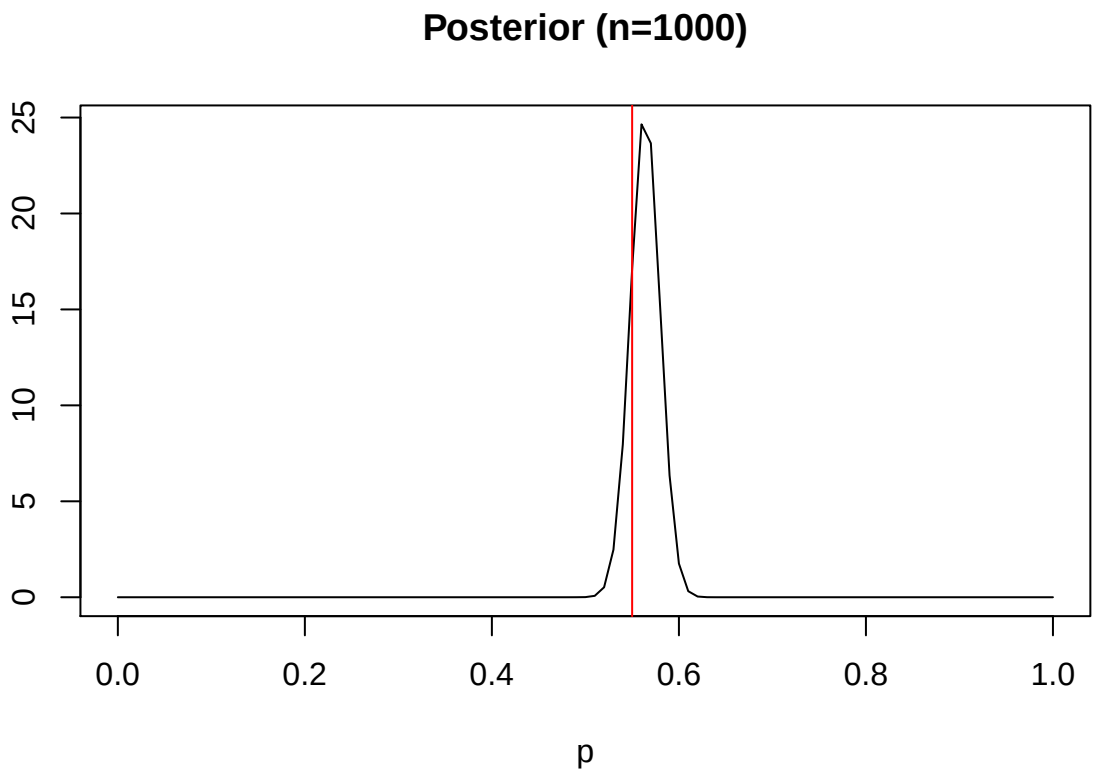
For this first example we have a small data sample size with very few successes.

```
X <- c(0, 0, 0, 0, 1)
n <- length(X)
alpha_post <- alpha_prior + n * mean(X)
beta_post <- beta_prior + n - n*mean(X)

curve(dbeta(x,alpha_post,beta_post),from=0,to=1,xlab="p",ylab="",main='Posterior (n=5)')
abline(v=0.2,col='red')
```



The second example we have many samples, but roughly half (p=0.55) are successes.

```
X <- rbinom(1, 1000, 0.55)
n <- 1000
alpha_post <- alpha_prior + X[1]
beta_post <- beta_prior + n - X[1]

curve(dbeta(x,alpha_post,beta_post),from=0,to=1,xlab="p",ylab="",main='Posterior (n=1000)')
abline(v=0.55,col='red')
```



We can see that the posterior distribution updates our beliefs about the parameter p based off the data we saw. For the first comparison we got a sample parameter value of 0.2, but we are not confident about that for the whole population with a small sample size, so the distribution is fat. The opposite in terms of fatness happens in the second case where we use a large number of observations, but the updated distribution is still centered around the sample parameter.

Problem 1 Exponential Likelihood

Part a: Likelihood Parameters

The exponential likelihood take one parameter λ known as its rate of decay. This parameter must be greater than. This rate dictates the rate of decay in the exponential decay, so a larger parameter will show a faster rate of decay. This can be seen in the likelihood function:

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

Part b: Conjugate Prior

The conjugate prior to an exponential likelihood is the Gamma distribution. The domain of this distribution is real numbers greater than 0, which matches our likelhood parameter possible values. The gamma distribution has two parameters α and β which pertain to the shape and rate of the pdf. A larger *beta* will create a prior pdf with a sharper rate of decay, meaning the parameter is more likely to be closer to zero. α will affect the shape of the distribution where a lower value will look like an exponential decay and a highervalue will create a modal distribution centered around a value greater than 0.

Part c: Posterior Parameters

The posterior parameters based off the data and prior are as follows:

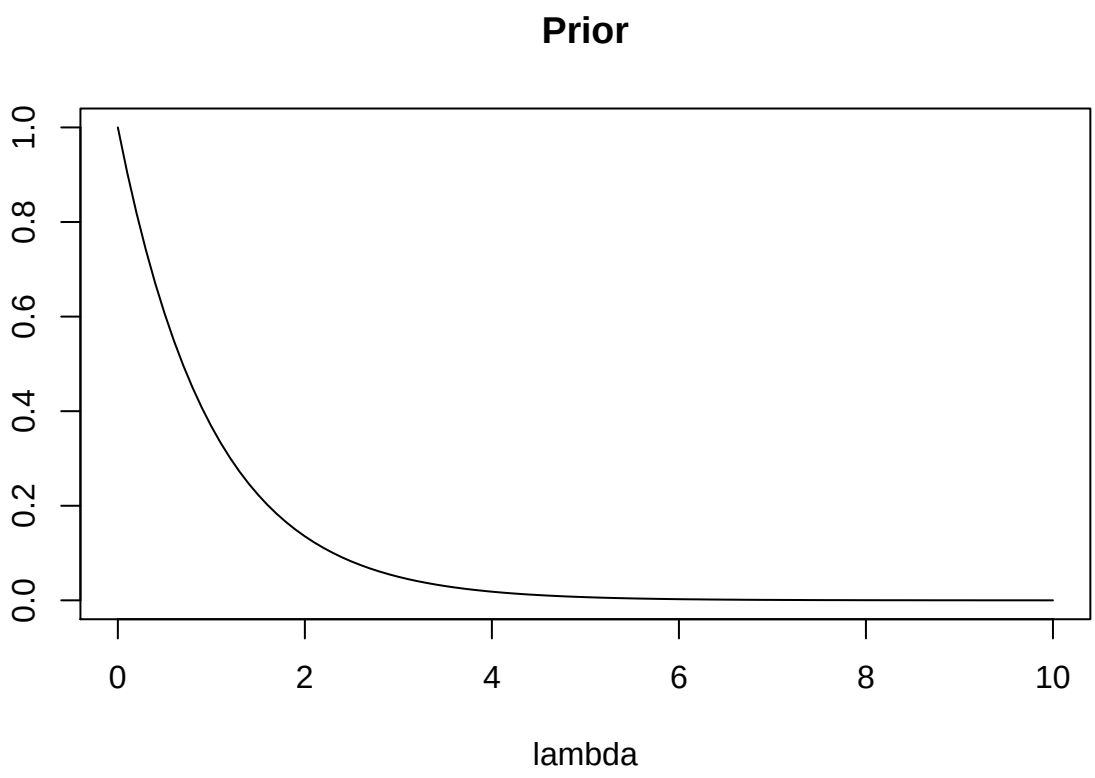
$$\alpha \to \alpha + n$$
$$\beta \to \beta + \Sigma_{i=1}^{n} X_i$$

Here we have two different effects captured with each parameter. The parameter α is acting as a count of the sample size of observed samples. With more samples we have a greater α, which will help to center the peak around an observed parameter. The parameter β is steadily increasing by taking on a value of the summation of the values found in the data. A larger value of the parameter will create a tighter distribution, so essentially we are more confident with large values in our data, or a large value from many observations.

Part d: Plots

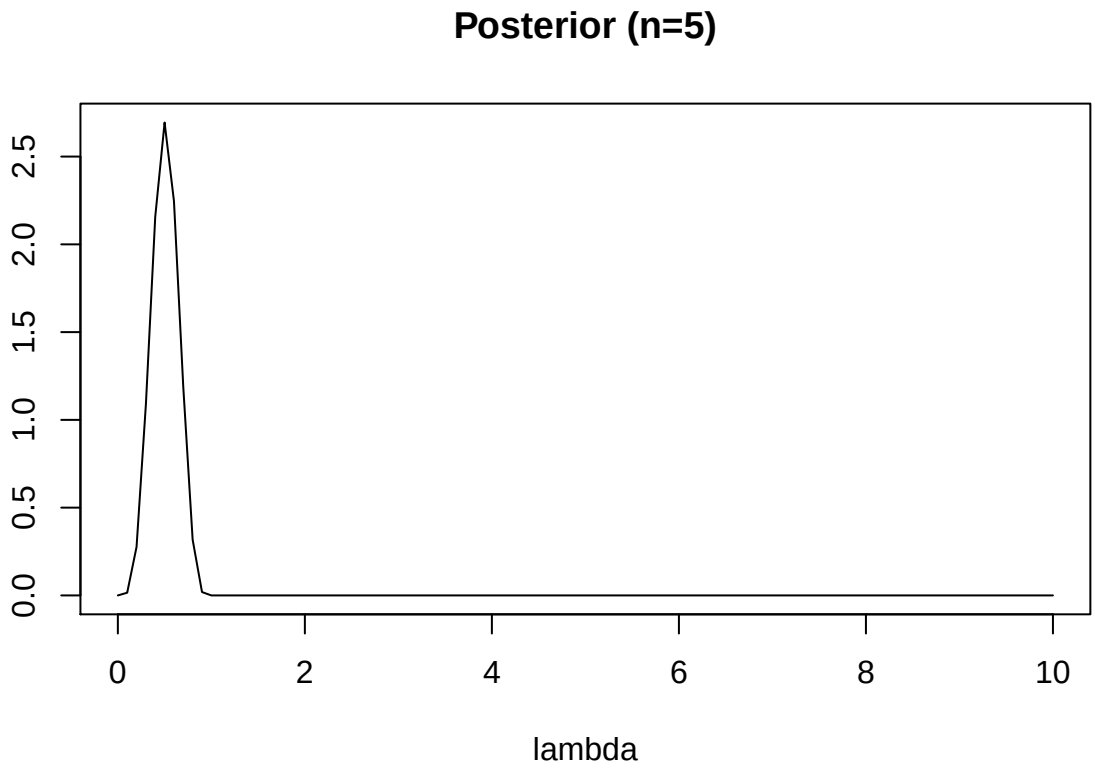
Below is our prior distribution with initial values of 1.

```
alpha_prior <- 1
beta_prior <- 1

curve(dgamma(x,alpha_prior,beta_prior),from=0,to=10,xlab="lambda",ylab="",main='Prior')
```



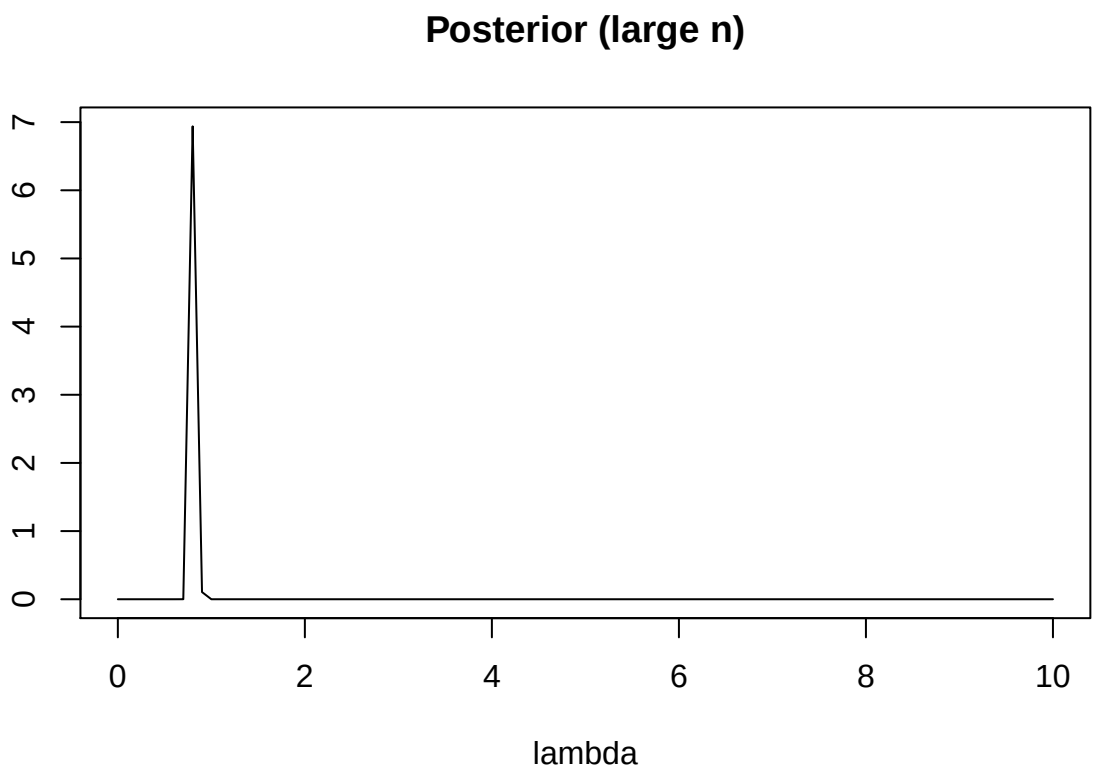
For this first example we have a small data sample size with a few large values.

```r
X <- c(0.1, 0.5, 0.1, 0.2, 4)
n <- length(X)
alpha_post <- alpha_prior + n
beta_post <- beta_prior + sum(X)

curve(dbeta(x,alpha_post,beta_post),from=0,to=10,xlab="lambda",ylab="",main='Posterior (n=5)')
```



The second example we have many samples, but all very small values.

```r
X <- seq(0.1, 0.3, 0.001)
n <- length(X)
alpha_post <- alpha_prior + n
beta_post <- beta_prior + sum(X)

curve(dbeta(x,alpha_post,beta_post),from=0,to=10,xlab="lambda",ylab="",main='Posterior (large n)')
```



Problem 3: Gaussian Likelihood (Known Sigma)

Part a: Likelihood Parameters

A Gaussian likelihood has two parameters, the mean and variance, also called μ and σ. The mean can take on any real value while $\sigma > 0$. The mean or μ dictates where the symmetrica gaussian distribution is centered. The variance or σ^2 dictates how high the peak is compared to how wide/large the tails are. A lower variance corresponds to a tigher/higher peak. In this problem we will hold σ fixed and "known".

Part b: Conjugate Prior

The conjugate prior for a Gaussian likelihood with a deterministic σ is also a gaussian distribution. The domain of this distribution is the entire real line. The prior parameters for this Gaussian are a μ_0 and σ_0. Where μ_0 would be our initial belief of the actual value of μ and σ would be larger the less confidence we have in our initial belief.

Part c: Posterior Parameters

The posterior parameters based off the data and prior are as follows:

$$\sigma_0^2 \rightarrow (\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1}$$

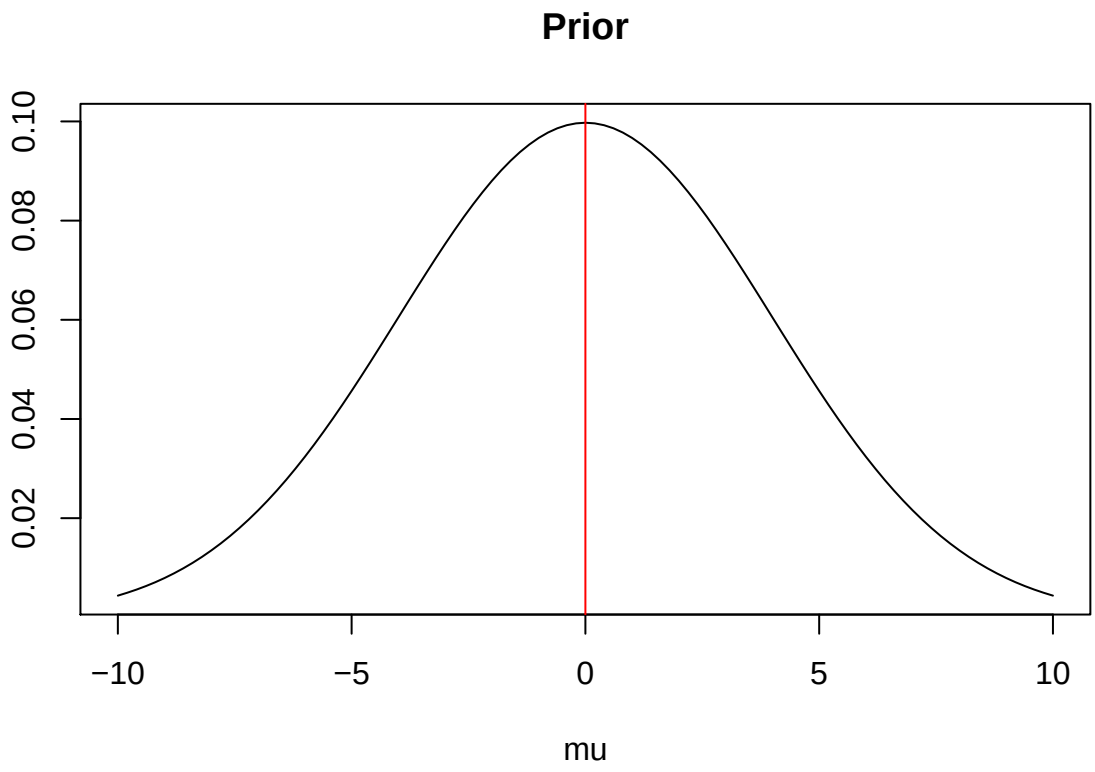
$$\mu_0 \to (\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1}(\frac{n\overline{X}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})$$

Here we have a steadily decreasing variance that decreases as our sample size increases, which makes sense, since we are essentially more confident about the true value of the parameter with more evidence. The estimated mean has two components in the posterior. The updated variance multipled by a second term. This second term captures adding the new sample mean to our existing knowledge in μ_0

Part d: Plots

Below is our prior distribution with initial values of $\mu_0 = 0$ and $\sigma = 4$. I am choosing a larger varaince the expected variance to communicate uncertainty, since I will be choosing values close to zero in my example (actual variance would be much smaller)

```
mu_prior <- 0
sig_prior <- 4

curve(dnorm(x,mu_prior,sig_prior),from=-10,to=10,xlab="mu",ylab="",main='Prior')
abline(v=0,col='red')
```

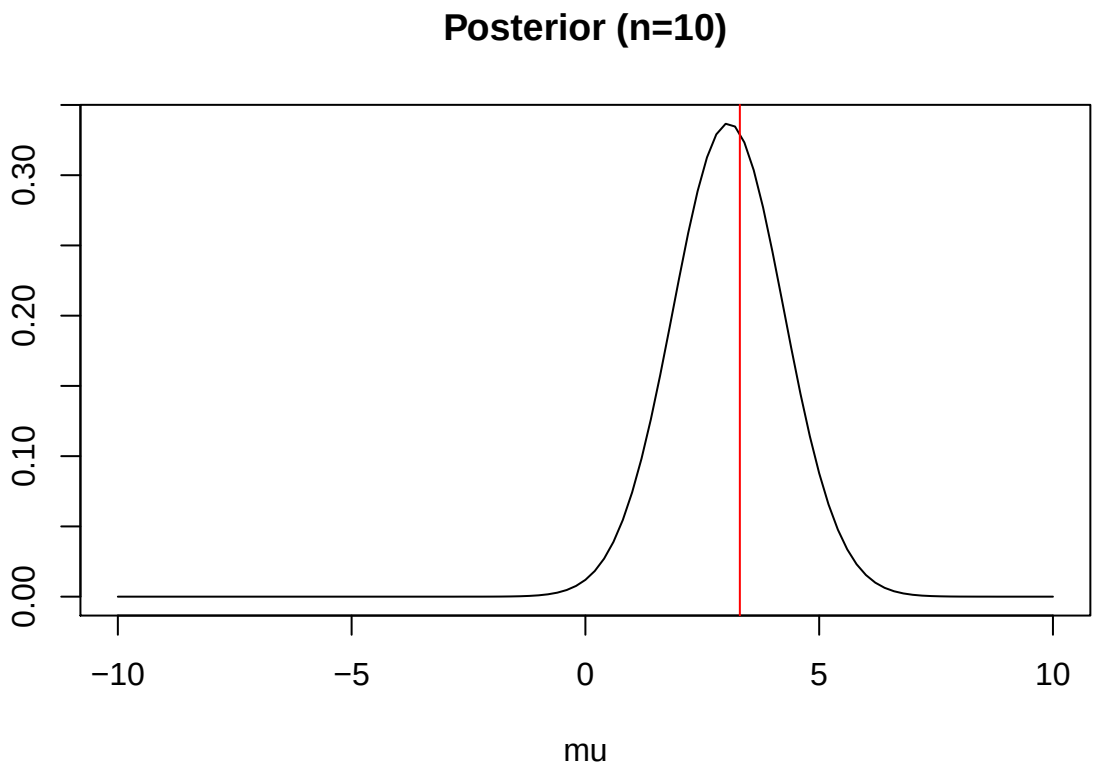


For this first example we have a small data sample size.

```
X <- rnorm(10, 0, 4)
print(mean(X))
```

```
## [1] 3.303289
```

```
n <- length(X)
sigma_post <- (n / sd(X)^2 + 1 / sig_prior^2) ^ -1
mu_post <-  sigma_post * (n * mean(X) / sd(X)^2 + mu_prior / sig_prior^2)

curve(dnorm(x,mu_post,sigma_post),from=-10,to=10,xlab="mu",ylab="",main='Posterior (n=10)')
abline(v=mean(X),col='red')
```

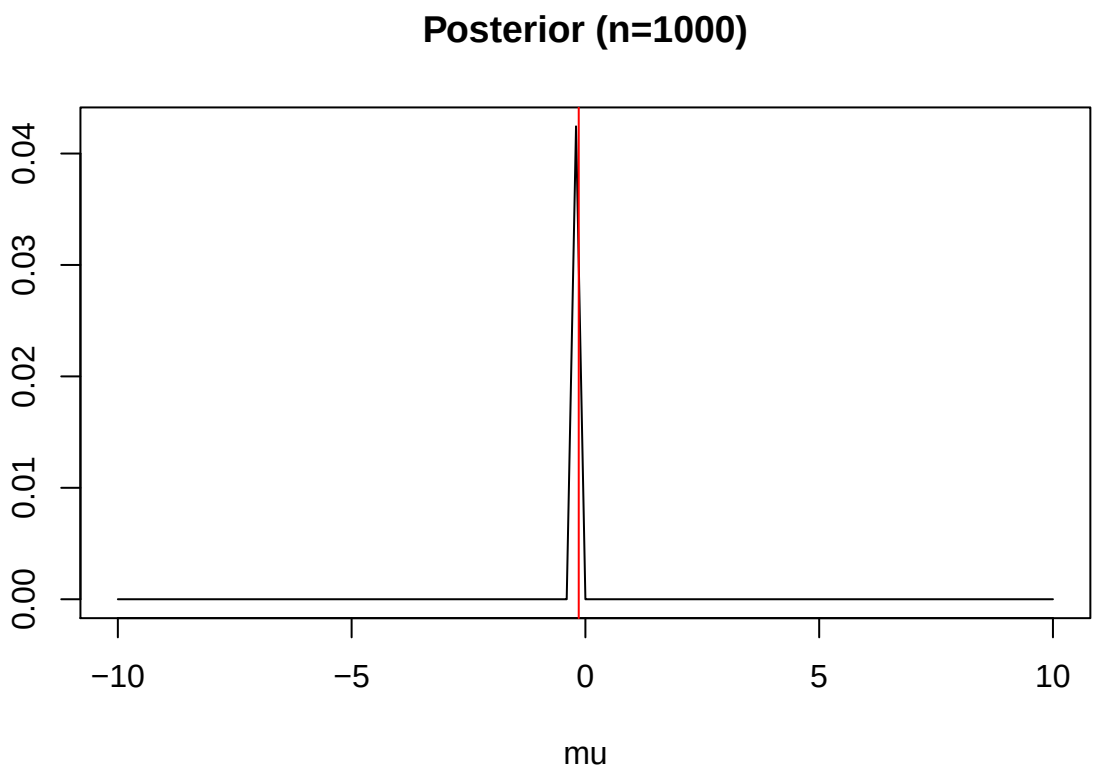


The second example we have many samples.

```
X <- rnorm(1000, 0, 4)
print(mean(X))
```

```
## [1] -0.1427202
```

```
n <- length(X)
sigma_post <- (n / sd(X)^2 + 1 / sig_prior^2) ^ -1
mu_post <-  sigma_post * (n * mean(X) / sd(X)^2 + mu_prior / sig_prior^2)

curve(dnorm(x,mu_post,sigma_post),from=-10,to=10,xlab="mu",ylab="",main='Posterior (n=1000)')
abline(v=mean(X),col='red')
```



We can see that in both cases the posterior parameter for μ adjusts to the sample mean, but the variance is very different based on the sample size of the evidence.

Problem 4: Gaussian Likelihood (Known Mean)

Part a: Likelihood Parameters

A Gaussian likelihood has two parameters, the mean and variance, also called μ and σ. The mean can take on any real value while $\sigma > 0$. The mean or μ dictates where the symmetrica gaussian distribution is centered. The variance or σ^2 dictates how high the peak is compared to how wide/large the tails are. A lower variance corresponds to a tigher/higher peak. In this problem we will hold μ fixed and "known".

Part b: Conjugate Prior

The conjugate prior for a Gaussian likelihoood with known mean is an inverse gamma distribution. The domain of this distribution is $(0, \infty)$. The prior parameters for this distribution are α and β. α is known as the shape parameter and β is known as the scale parameter. A smaller β will create a larger rate of decay on the right tail and a smaller *alpha* will result in a flatter distribution. Note that the mean of this distribution is $\frac{\beta}{\alpha - 1}$.

Part c: Posterior Parameters

The posterior parameters based off the data and prior are as follows:

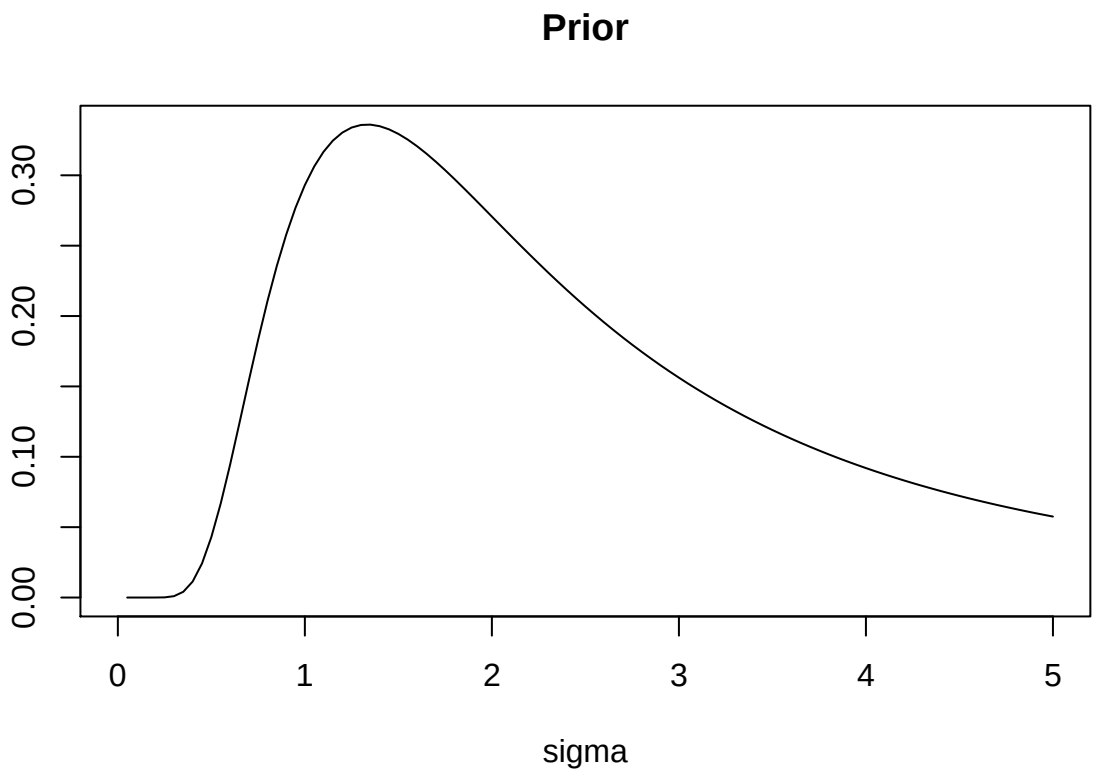
$$\alpha \to \alpha + n/2$$
$$\beta \to \beta + \frac{1}{2}\Sigma(X_i - \mu)^2$$

The α parameter is a measure of the number of observations in our evidence while the β parameter is updated by adding the sample variance from the set of incoming observtions.

Part d: Plots

We will base our plots below around a know mean of 0. We will start our initial guess of variance around 4 where $\beta = 4$ and $\alpha = 2$.

```
library(invgamma)
alpha_prior <- 2
beta_prior <- 4

curve(dinvgamma(x,alpha_prior,beta_prior),from=0,to=5,xlab="sigma",ylab="",main='Prior')
```

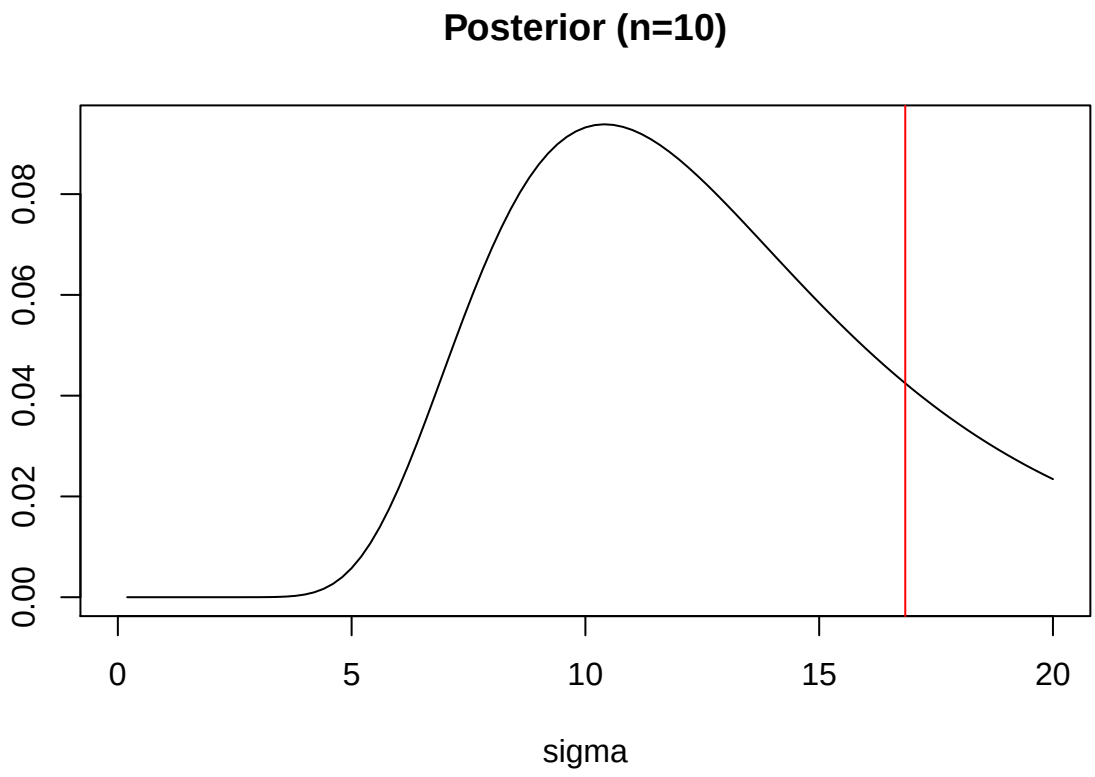


For this first example we have a small data sample size.

```
X <- rnorm(10, 0, 4)
print(sd(X)^2)
```

```
## [1] 16.84106
```

```r
n <- length(X)
alpha_post <- alpha_prior + n/2
beta_post <-  beta_prior + 1/2*sum(X^2)

curve(dinvgamma(x,alpha_post,beta_post),from=0,to=20,xlab="sigma",ylab="",main='Posterior (n=10)')
abline(v=sd(X)^2,col="red")
```

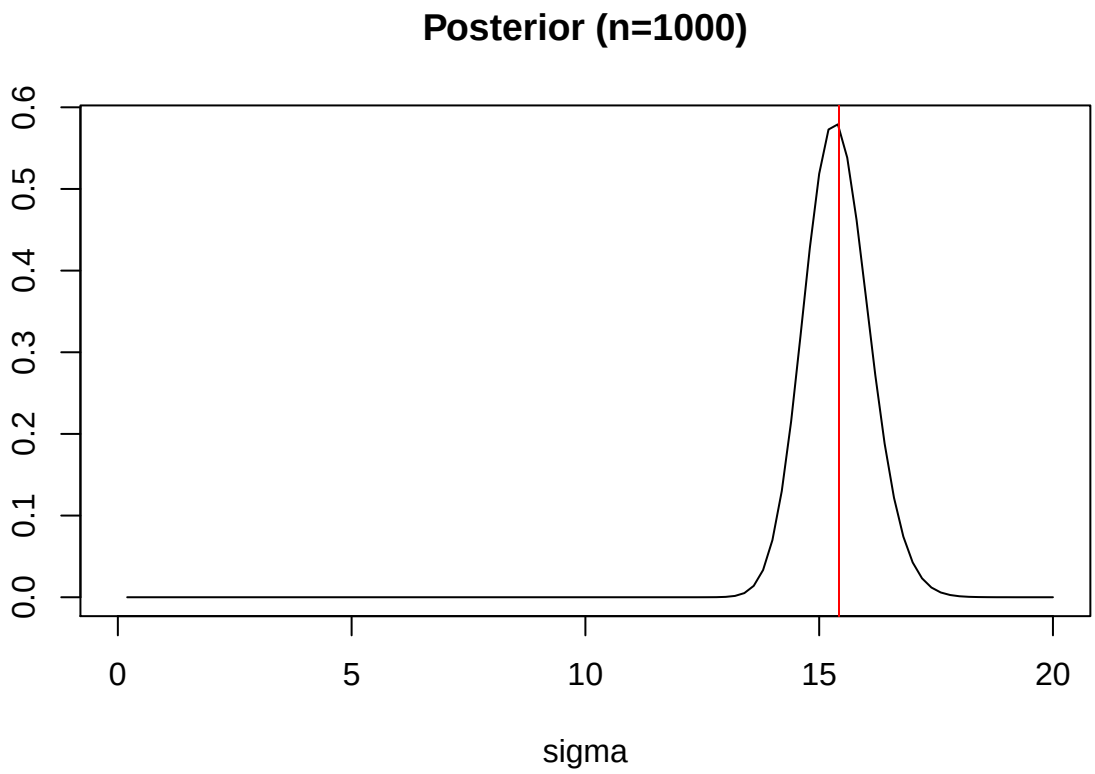


The second example we have many samples.

```r
X <- rnorm(1000, 0, 4)
print(sd(X)^2)
```

```
## [1] 15.42391
```

```r
n <- length(X)
alpha_post <- alpha_prior + n/2
beta_post <-  beta_prior + 1/2*sum(X^2)

curve(dinvgamma(x,alpha_post,beta_post),from=0,to=20,xlab="sigma",ylab="",main='Posterior (n=1000)')
abline(v=sd(X)^2,col='red')
```

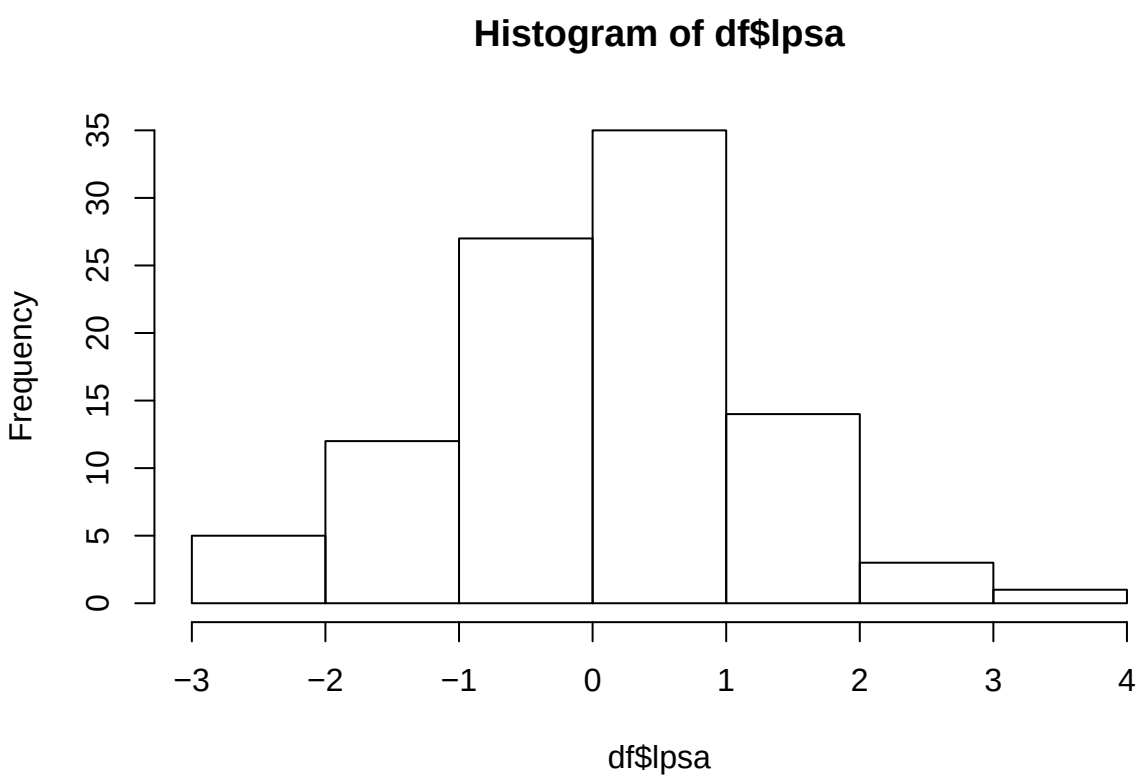


We can see in both cases that the posterior distribution mean is the value of the sample variance exhibited in our data or evidence, but when we have a much larger sample size (n) then the inverse gamma distribution has much smaller tails, which would shrink our credible interval.

Problem 5

First let's read the data and print the basic statistics.

```
df <- read.csv('./data/prostate.csv')
hist(df$lpsa)
```




```
print(mean(df$lpsa))
```

```
## [1] -1.969206e-15
```

```
print(sd(df$lpsa)^2)
```

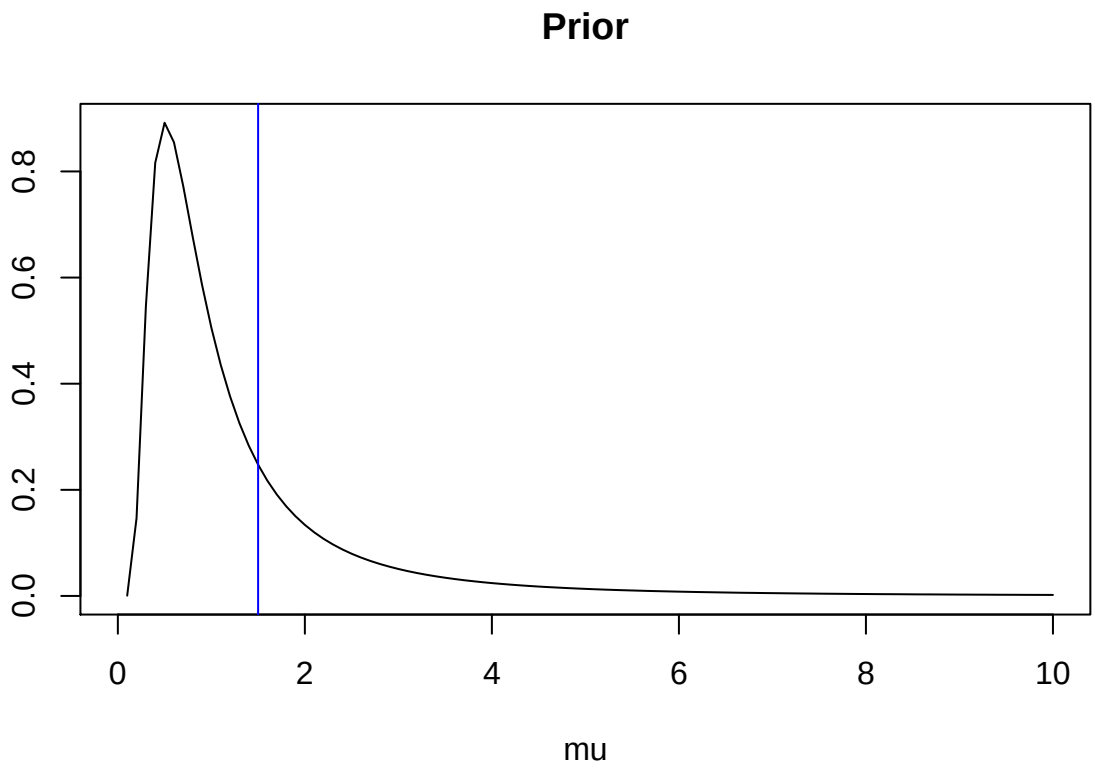
```
## [1] 1.332476
```

Part a

As stated in problem 4, in a normal Bayesian setting with a known μ we use the Inverse Gamma distribution as the conjugate prior to model the variance parameter for the normal likelihood function. For our prior we need to set two parameters α and β as described above. We will set the parameters to make the expected value of variance around 1.5 and have a high initial variance of the inverse gamma distribution to compensate for our initial lack of knowledge.

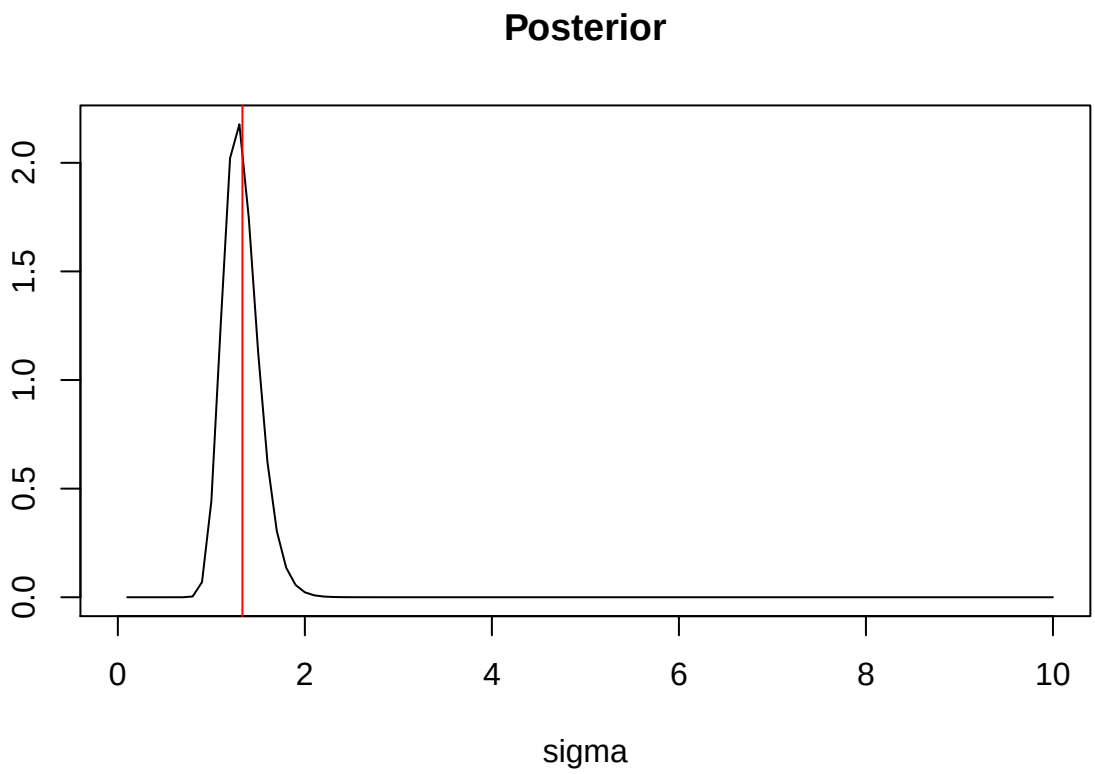
$$\alpha_0 = 2 + \frac{1.5}{100}$$
$$\beta_0 = (\alpha_0 - 1) * 1.5$$

```r
expected <- 1.5
var <- 100

alpha_prior <- 2 + (expected / var)
beta_prior <- (alpha_prior - 1)*expected

curve(dinvgamma(x,alpha_prior,beta_prior),from=0,to=10,xlab="mu",ylab="",main='Prior')
abline(v=expected,col='blue')
```



Part b

```r
X <- df$lpsa
n <- length(X)
alpha_post <- alpha_prior + n/2
beta_post <-  beta_prior + 1/2*sum(X^2)

curve(dinvgamma(x,alpha_post,beta_post),from=0,to=10,xlab="sigma",ylab="",main='Posterior')
abline(v=sd(X)^2,col='red')
```



We can see that our posterior distribution maintains a similar mean, but the variance is largely decreased. This is expected as we initialized the prior parameters to have a mean near our sample mean (with our sneak peek!), so the posterior distribution takes this evidence and centers our belief distribution around the sample mean. The much smaller variance takes into account the amount of evidence we have gained through using the data. The more samples, the smaller the variance will be.

Part c

We will compute samples from the posterior as the product of our likelihood and prior. This function will help determine the acceptance probability.

```
X <- df$lpsa

posterior <- function(sig) {
    likelihood <- prod(dnorm(X,mean(X),sig))
    prior <- dinvgamma(sig, alpha_prior, beta_prior)
    likelihood * prior
}
```

Next we will create our proposal function. This proposal function will be a uniform distribution with width 2τ where τ is a chosen width. Standard deviations produced lower than or equal to 0 will be rounded back up to just above 0.

```
tau <- 1
```

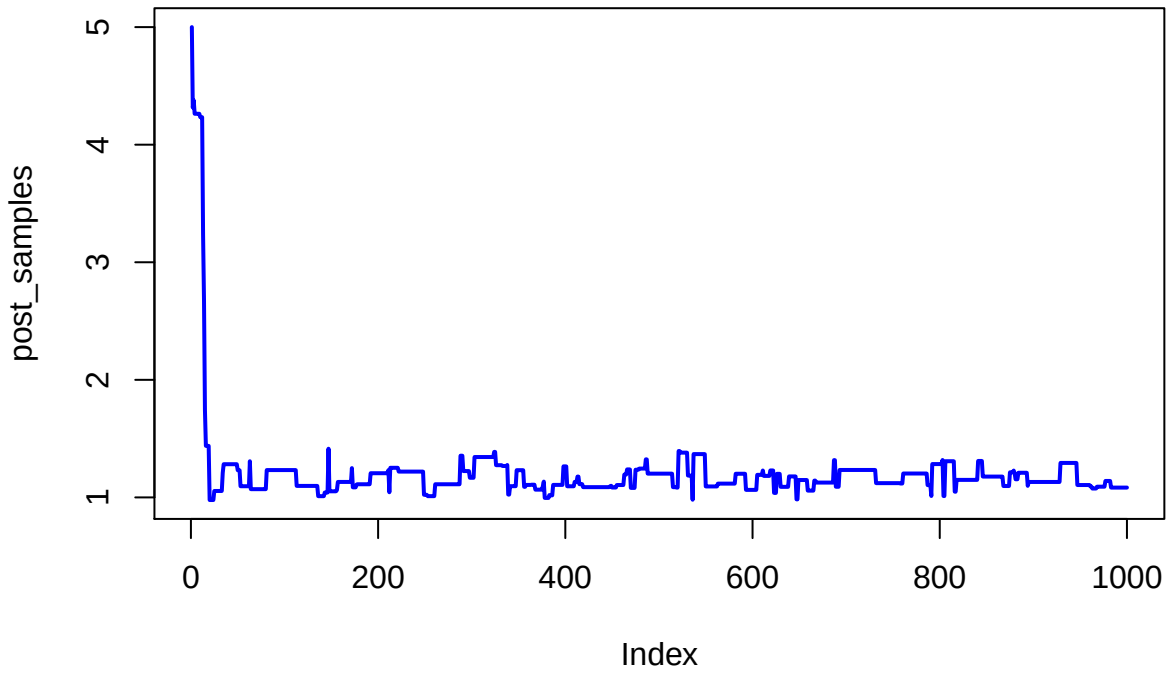
```
prop <- function(xprop,x) {
    dunif(xprop, x-tau, x+tau)
}

gen_prop <- function(x) {
    prop <- runif(1, x-tau, x+tau)
    if (prop <= 0) {
        prop <- 0.001
    }
    prop
}
```

```
accept <- function(xp, x, prop_func) {
    ratio <- prop_func(x, xp)*posterior(xp)/(prop_func(xp,x)*posterior(x))
    if (is.na(ratio)) {
        print(x)
        print(xp)
    }
    if (ratio < 1) {
        ratio
    } else {
        1
    }
}
```

```
mcmc <- function(x0, M, prop_func, gen_prop_func) {
    # init vector
    result <- numeric(M)
    result[1] <- x0
    for (i in 2:M) {
        # get current point
        x <- result[i-1]
        # propose a new point
        xprop <- gen_prop_func(x)
        prob_accept <- accept(xprop, x, prop_func)
        # choose whether to accept
        accepted <- as.logical(rbinom(1,1,prob_accept))
        if (accepted) {
            result[i] <- xprop
        } else {
            result[i] <- x
        }
    }
    result
}
```

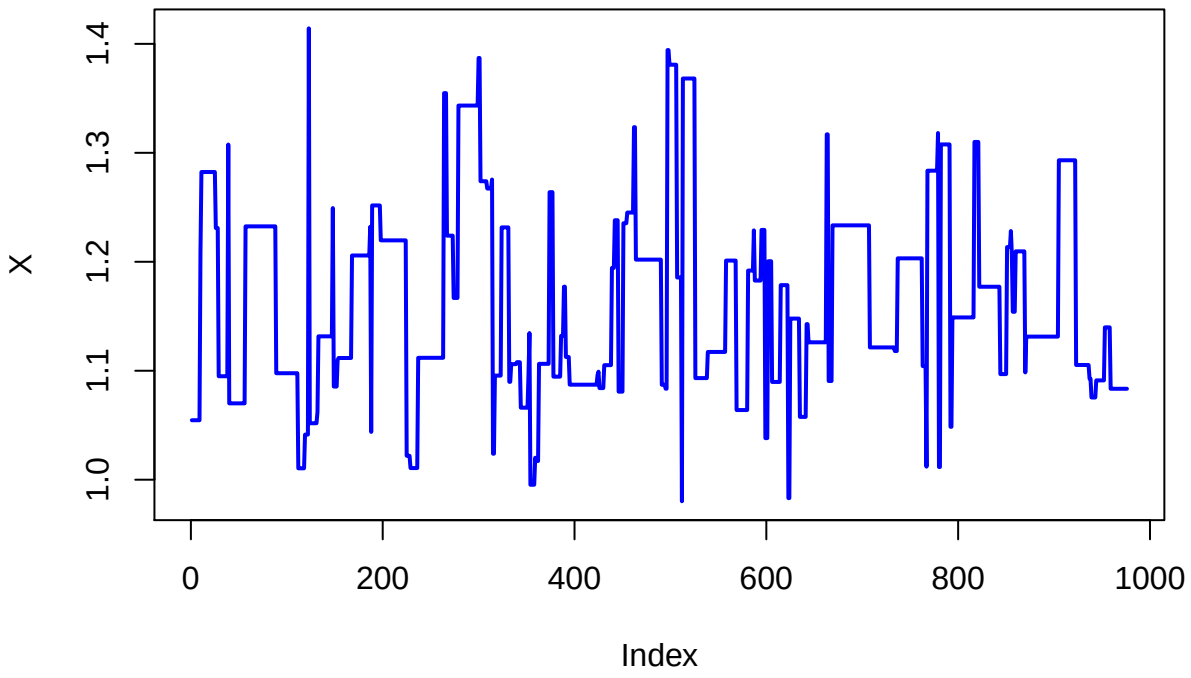
```
# sample size
M <- 1000
# inital point
x0 <- 5

post_samples <- mcmc(x0, M, prop, gen_prop)
```

```
plot(post_samples,type="l",lwd=2,col="blue")
```



We can see that the samples converge rather quickly, so we will just throw out the first 25:

```
X <- post_samples[25:length(post_samples)]
plot(X,type="l",lwd=2,col="blue")
```



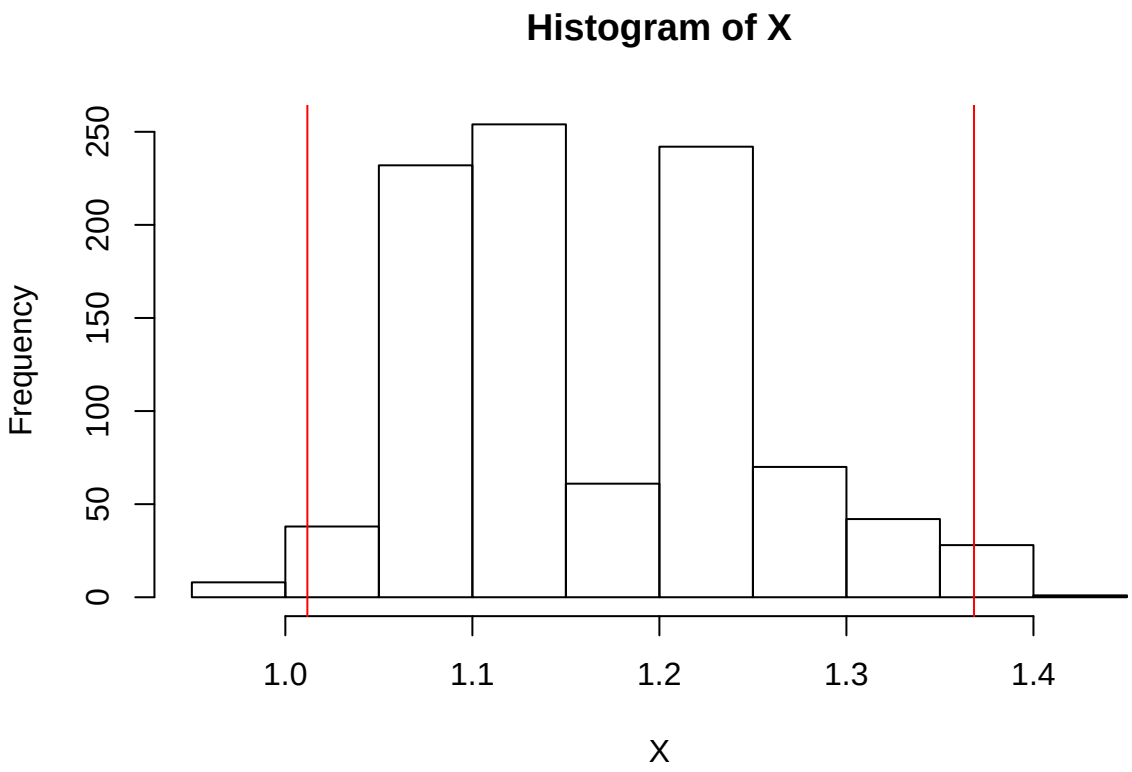
Part d

```
cred <- 0.05
int <- quantile(X, probs=c(cred/2, 1-(cred/2)))
int
```

```
##     2.5%    97.5%
## 1.011772 1.368227
```

Part e

```
hist(X)
abline(v=int,col='red')
```


Histogram of X
Frequency
0
50
100
150
200
250
1.0
1.1
1.2
1.3
1.4
X