Homework 10 - STAT440

Joseph Sepich (jps6444)

11/15/2020

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

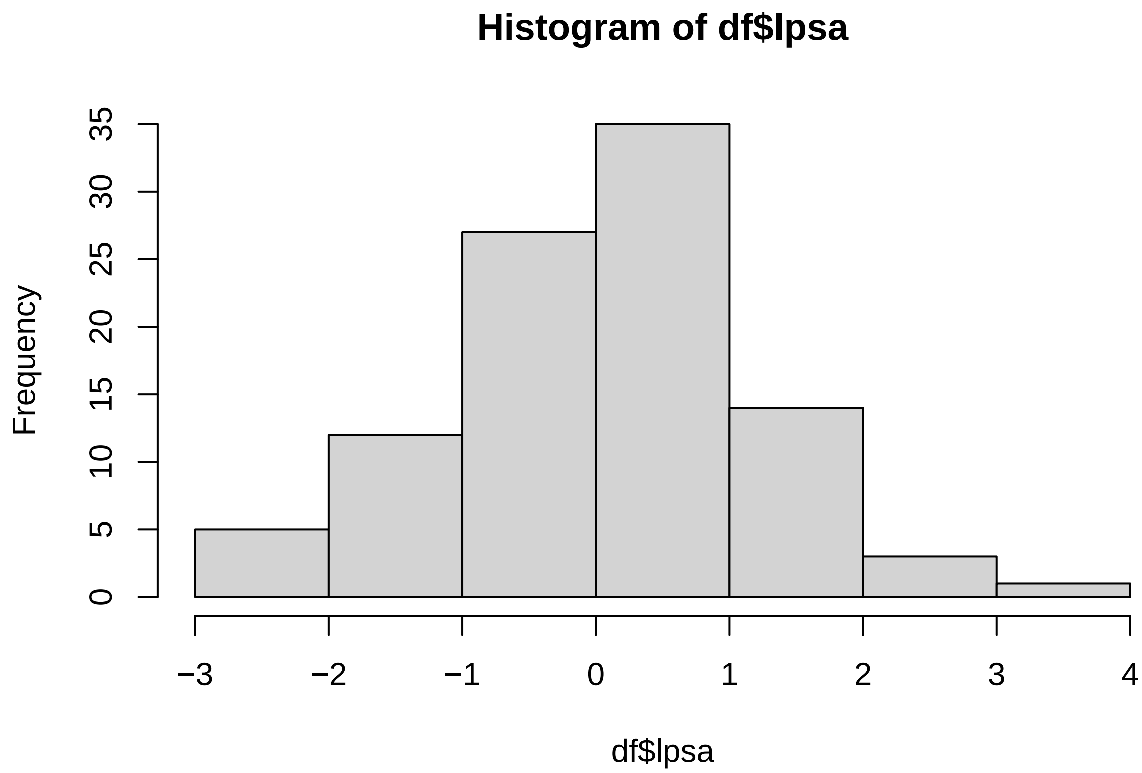
```r
df <- read.csv('./data/prostate.csv')
head(df)
```

```
##   X     lcavol    lweight        age      lbph        lcp    gleason      lpsa
## 1 1 -1.6373556 -2.0062118 -1.8624260 -1.024706 -0.8631712 -1.0421573 -2.909170
## 2 2 -1.9889805 -0.7220088 -0.7878962 -1.024706 -0.8631712 -1.0421573 -2.640906
## 3 3 -1.5788189 -2.1887840  1.3611634 -1.024706 -0.8631712  0.3426271 -2.640906
## 4 4 -2.1669171 -0.8079939 -0.7878962 -1.024706 -0.8631712 -1.0421573 -2.640906
## 5 5 -0.5078745 -0.4588340 -0.2506313 -1.024706 -0.8631712 -1.0421573 -2.106823
## 6 6 -2.0361285 -0.9339546 -1.8624260 -1.024706 -0.8631712 -1.0421573 -1.712919
```

Problem 1

Part a

```r
hist(df$lpsa)
```



I am going to use the normal distribution as the form of our liklihood function. The data is symmetric and has a single mode in the center. This mimics the form of a normal distribution.

Part b

Note that we will use the sample variance as our "known", fixed variance value.

We will use a normal distribution as our prior distribution for our mean. We will use this as it is the conjugate prior for the normal distribution likelihood, so the posterior will also be a normal distribution.

Part c

A large portion of the values are near 0, so we will use the prior parameter $\mu_0 = 0$. It is likely that the 0 level is considered the "normal" level for measurement purposes as (via very brief research) there is a certain level that doctors use as a cutoff for whether further tests are needed (in the non-log variable). The sample mean is also very close to this value.

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

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

Part d

We will use Metropolis-Hastings to produce a sample from the posterior distribution using our liklihood and prior (which are both normal distributions). Our proposal distribution will be a uniform distribution with parameters roughly equal to the fifth and ninety-fifth percentiles of our sample. We will set them to x-2 and x+2.

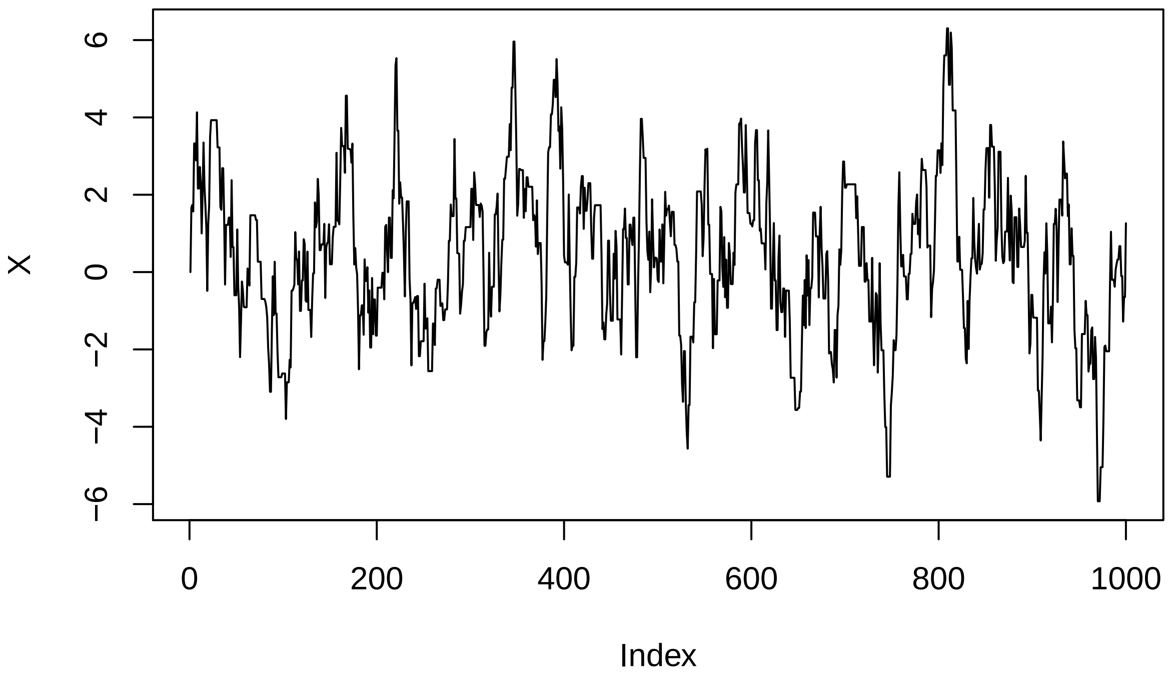
```
probs <- c(0.05, 0.95)
quantile(df$lpsa, probs)
```

```
##        5%       95%
## -1.791700  1.702283
```

```
# init params
n <- 1000
sig_true <- sd(df$lpsa)

mu_prior <- 0
sig_prior <- sig_true # deterministic

X <- numeric(n)
X[1] <- mu_prior

mu <- function(){rnorm(1,mean=mu_prior,sd=sig_true)} # prior
f <- function(x){dnorm(x,mean=mu(),sd=sig_true)} # likelihood
Q <- function(x1,x2){dunif(x1,min=x2-2,max=x2+2)}

accept_fun <- function(x_c,x_p) {
    accept <- (Q(x_c,x_p)/Q(x_p,x_c))*(f(x_p)/f(x_c))
    return(min(accept,1))
}
```

```
for(i in 2:n) {
    x_proposed <- X[i-1] + runif(1,min=-2,max=2)
    accept <- accept_fun(X[i-1],x_proposed)
    decision <- rbinom(1,1,accept)
    if(decision == 1){
        X[i] = x_proposed
    } else {
        X[i] = X[i-1]
    }
}
```

```
plot(X,type='l')
```



We will not throw away any observations as it appears that we initialized our value around the convergence point.

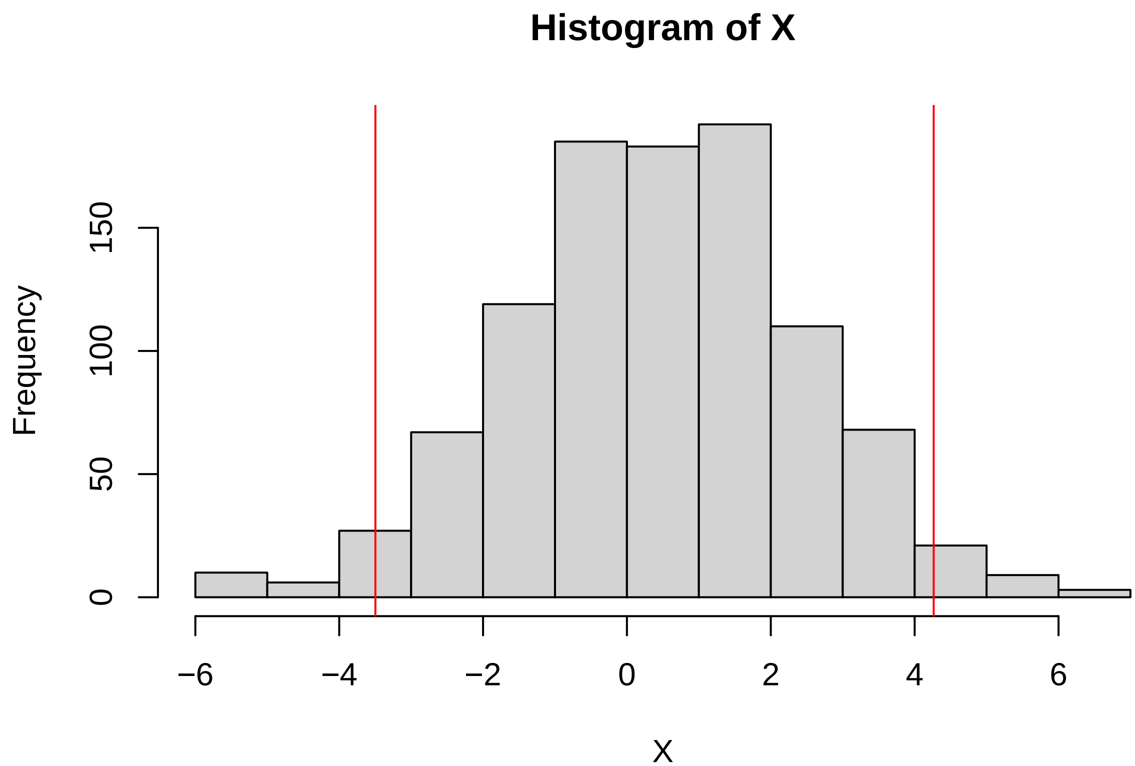
Part e

```
c_val <- 0.05
cred_int <- quantile(X, probs=c(c_val/2, 1-c_val/2))
cred_int
```

```
##      2.5%     97.5%
## -3.497190  4.264728
```

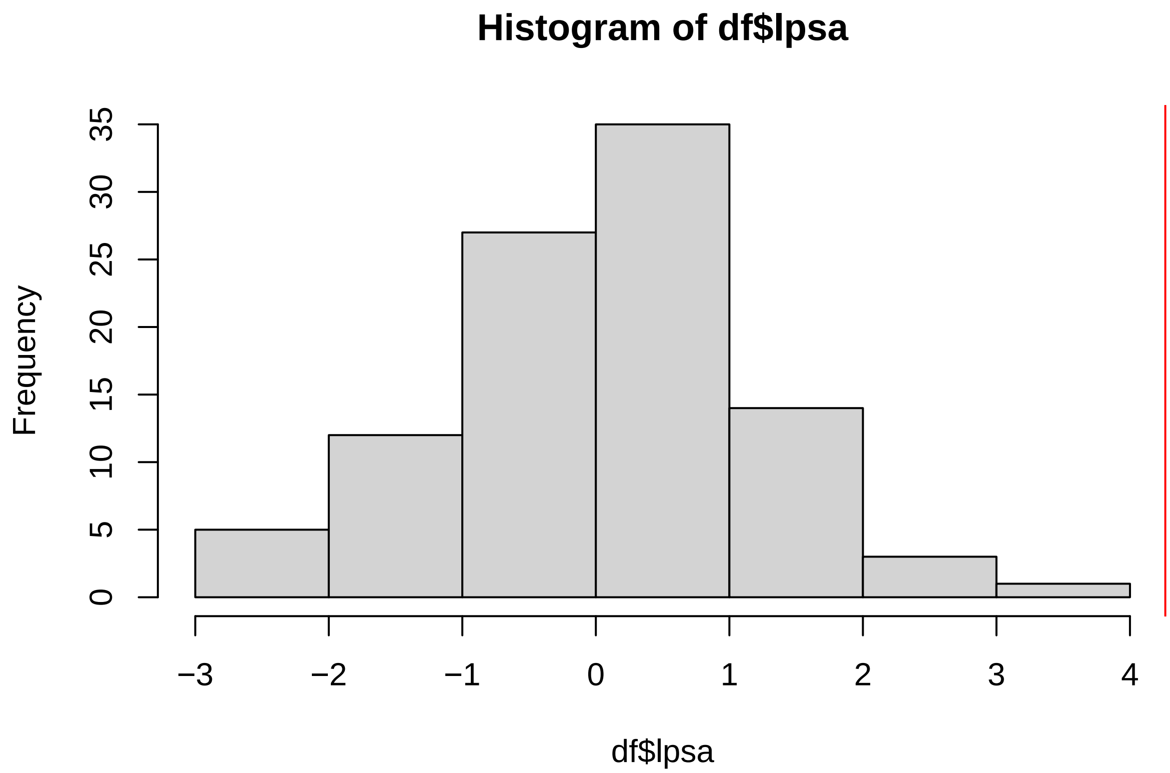
Part f

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



This credible interval tells me that my data has a mean between -4 and 4 and the data falls within that interval as well.

```
hist(df$lpsa)
abline(v=cred_int,col='red')
```



Problem 2

Part a

Part b

Part c

Part d

Part e