Assignment 1

1.(a) $p(\theta|y)=p(y|\theta)*p(\theta)/p(y)$

uniform prior->$p(\theta)=1$

For movie 1, $p(\theta|y)\propto p^{150}(1-p)^{50}$ ~ Beta(151,51)

For movie 2, $p(\theta|y)\propto p^{4}(1-p)$ ~ Beta(5,2)

(b)Posterior Mean:

Movie 1: 151/(151+51)=0.7475

Movie 2: 5/(5+2)=0.7143

Movie 1 ranks higher.

Posterior Median:

Movie 1: qbeta(0.5, 151, 51)=0.748

Movie 2: qbeta(0.5, 5, 2)=0.736

Movie 1 ranks higher.

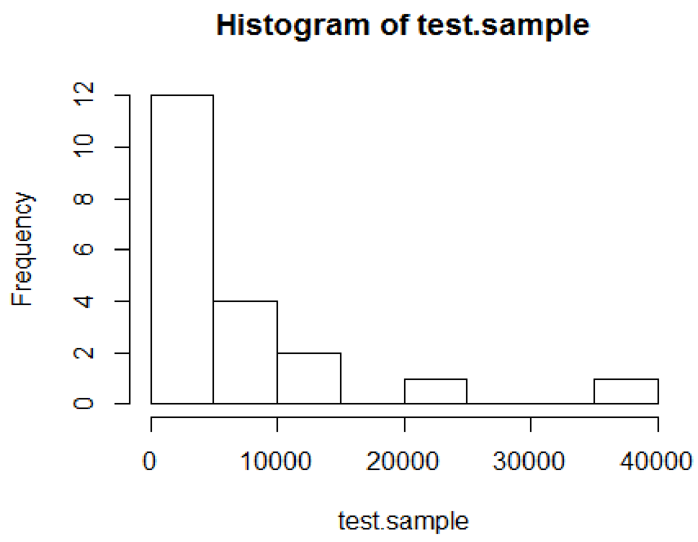
Posterior Mode:

Movie 1: (151-1)/(151+51-2)=0.75

Movie 2: (5-1)/(5+2-2)=0.80

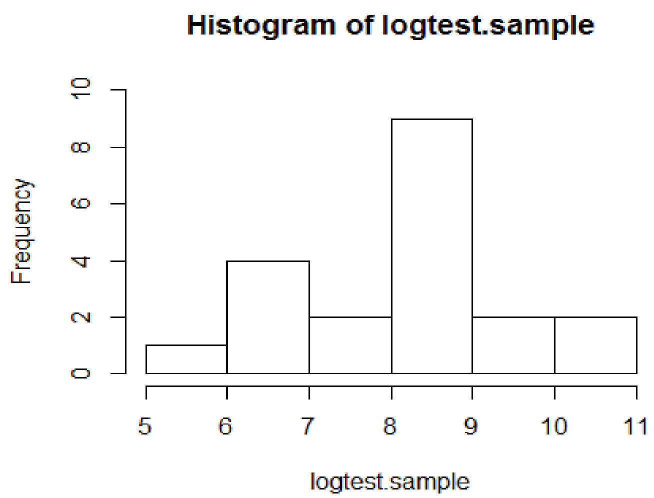
Movie 2 ranks higher.

2.(a)(i)



Description: The y-axis represents the counts of each data within value range of every 5000 interval in x-axis. 12 observations of article length are in 0-5000 bytes, 4 in 5000-10000 bytes, 2 in 10000-15000 bytes, 1 in 20000-25000 bytes, 1 in 35000-40000 bytes. In total there are 20 observations.

(ii)



Description: The y-axis represents the counts of each data's log value within the corresponding interval in x-axis. 1 observation of article length are in (5,6), which means original data is between e^5 and e^6. 4 in (6,7), 2 in (7,8), 9 in (8,9), 2 in (9,10), 2 in (10,11). In total there are 20 observations.

(iii) Coding Process:

```
hist(test.sample, plot=TRUE, freq=TRUE)
logtest.sample <- log(test.sample, base = exp(1))
hist(logtest.sample, plot=TRUE, freq=TRUE, ylim=c(0,10))
```

The second graph is better because it makes the x-axis in terval value more readable. The original value is in too la rge scale which is not good for analysis. Specifically, samp le variance, sample standard deviation and sample maxim um would be too large.

(b) mean(logtest.sample)

Sample Mean Result->8.1604

sd(logtest.sample)

Sample Standard Deviation Result->1.242427

(c) median of y_i=8.270

μ_{prior}=8.270

var_{prior}=var(logtest.sample)=1.543624

(i) posterior mean: 8.165612

```
> mun <- (median(logtest.sample)/var(logtest.sample)+
20*mean(logtest.sample)/var(logtest.sample))/(1/var(logt
est.sample)+20/var(logtest.sample))
> mun
[1] 8.165612
```

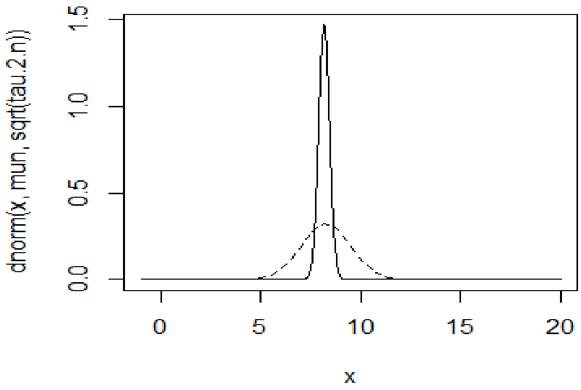
posterior variance: 0.07350592

```
> tau.2.n <-1/(1/var(logtest.sample)+20/var(logtest.sam
ple))
> tau.2.n
[1] 0.07350592
```

posterior precision: 13.60435

```
> 1/tau.2.n
[1] 13.60435
```

(ii)



Full line: posterior density

Imaginary line: prior density

Code:

```
> curve(dnorm(x,mun,sqrt(tau.2.n)), -1, 20, n=1000)
> curve(dnorm(x,median(logtest.sample),sd(logtest.sampl
e)), -1, 20,add=TRUE, lty=2)
```

(iii) Code:

```
> mun+c(-1,1)*1.645*sqrt(tau.2.n)
[1] 7.719620 8.611604
```

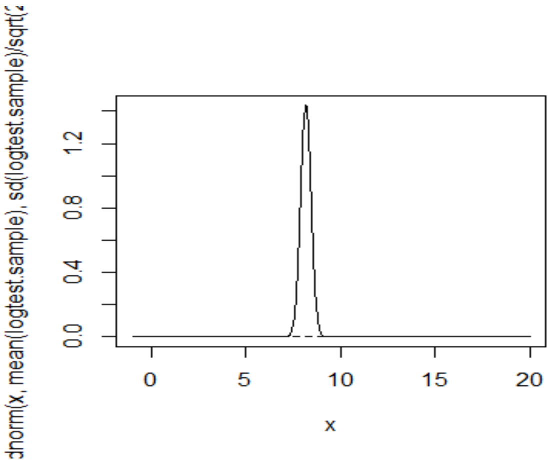
90% central posterior interval (7.719620, 8.611604)

(d)(i) posterior mean=sample mean=8.1604

posterior variance=sample variance/n=1.543624/20=0.07718121 posterior precision=sample precision=0.647826

```
> n/var(logtest.sample)
[1] 12.95652
```

(ii)



Full line: posterior density

Imaginary line: prior density

Code:

```
> curve(dnorm(x,mean(logtest.sample),sd(logtest.sample)/s
qrt(20)), -1, 20, n=1000)
> curve(dnorm(x,1,0), -1, 20,add=TRUE, lty=2)
```

(iii) 90% central posterior interval: (7.703394, 8.617406)

Code:

```
> mean(logtest.sample)+c(-1,1)*1.645*sd(logtest.sample)/s
qrt(20)
[1]  7.703394  8.617406
```

(e)(i) mean: 6754.0 variance: 3880315

```
> post.mu.sim <- rnorm(1000, mean(test.sample), sd(test.sample)/sqrt(20))
> summary(post.mu.sim)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  924.7   5426.0   6685.0   6754.0   7975.0  14590.0
> var(post.mu.sim)
[1] 3880315
```

(ii) Assume each article is i.i.d, total number of bytes equal to the sum of bytes of each article. Thus, total bytes estimation = 5.7million*6754.0 = 38497.8 million