ADVANCED BAYESIAN MODELING - ASSIGNMENT 1

Shyam Shah

09/02/2021

1 (a)

Posterior distribution of p_1:

$\theta \sim U(0,1)$ so $p(\theta) = 1$

$n = 10$

$y = 9$

The posterior likelihood can be defined as:

$p(\theta|y = 9) \propto p(\theta) \cdot p(y = 9|\theta)$

$\propto 1 * \binom{10}{9} \cdot \theta^9 \cdot (1-\theta)^{10-9}$

$\propto \theta^9 \cdot (1-\theta)^1$

$\theta^9 \cdot (1-\theta)^1$ is in the form of a $Beta$ distribution with $\alpha = 10$ and $\beta = 2$

Posterior distribution of p_2:

$\theta \sim U(0,1)$ so $p(\theta) = 1$

$n = 500$

$y = 425$

The posterior likelihood can be defined as:

$p(\theta|y = 425) \propto p(\theta) \cdot p(y = 425|\theta)$

$\propto 1 * \binom{500}{425} \cdot \theta^{425} \cdot (1-\theta)^{500-425}$

$\propto \theta^{425} \cdot (1-\theta)^{75}$

$\theta^{425} \cdot (1-\theta)^{75}$ is in the form of a $Beta$ distribution with $\alpha = 426$ and $\beta = 76$

1 (b)

Using posterior mean:

Formula for posterior mean is $\frac{\alpha}{\alpha+\beta}$. So for p_1, it is $10/12 \approx 0.83$ and for p_2 it is $426/502 \approx 0.85$ meaning movie 2 ranks higher.

Using posterior median:

```
median1 <- qbeta(0.5, 10, 2)
median2 <- qbeta(0.5, 426, 76)
median1>median2
```

```
## [1] TRUE
```

So movie 1 ranks higher.

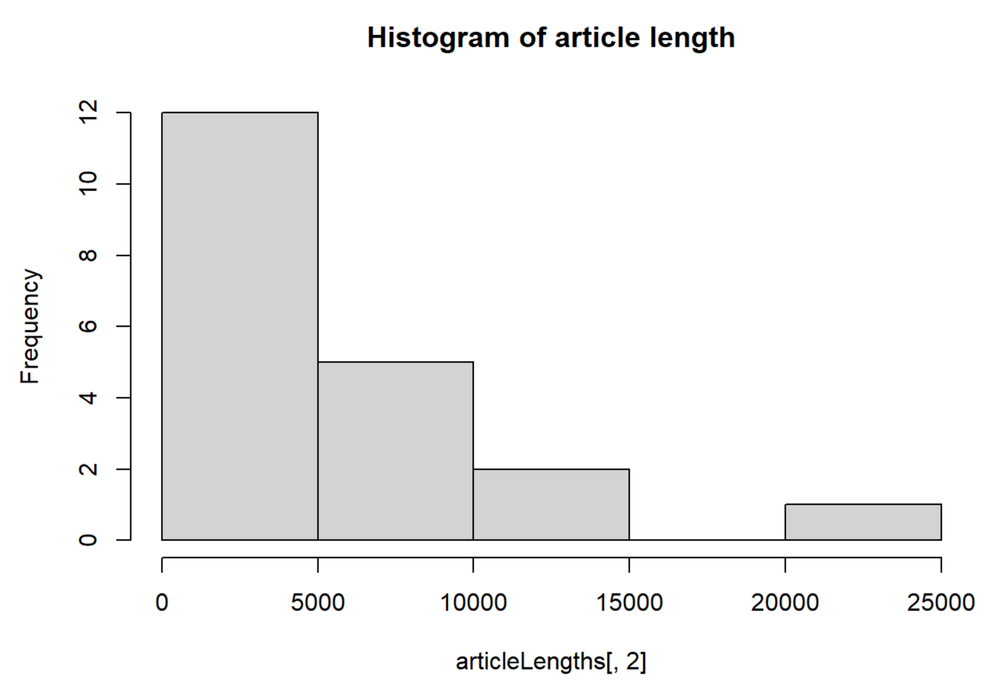
Using posterior mode:

Formula for posterior mean is $\frac{\alpha-1}{\alpha+\beta-2}$. So for p_1, it is $9/10 \approx 0.9$ and for p_2 it is $425/500 \approx 0.85$ meaning movie 1 ranks higher.

2 (a) (i)

The histogram looks like a beta distribution graph with a right skew(ie. high β and low α).

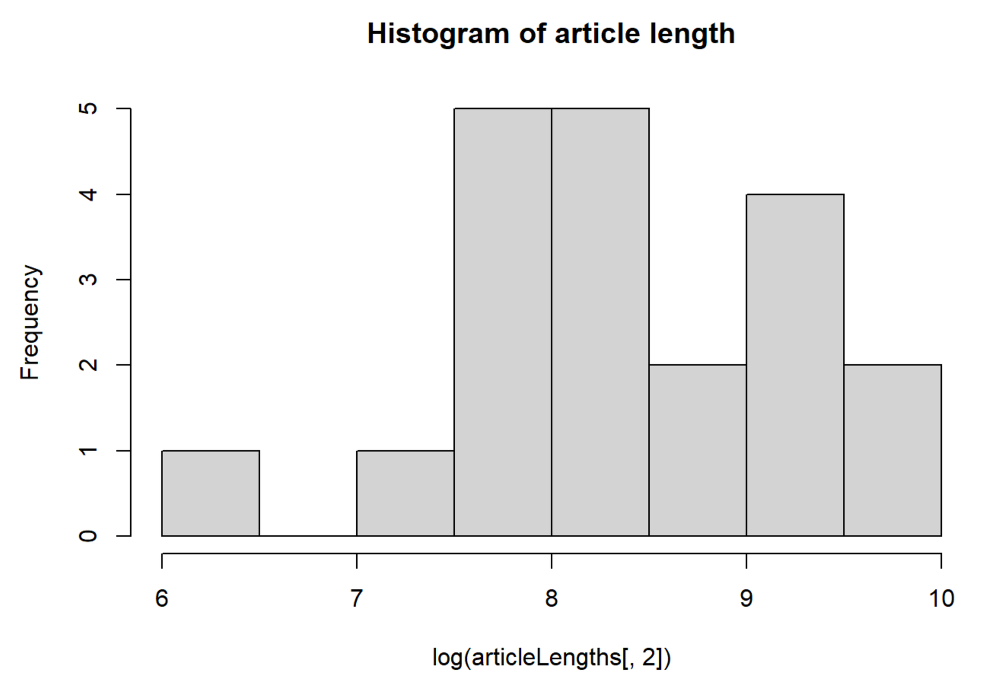
```
articleLengths <-read.table("randomwikipedia.txt")
hist(articleLengths[,2], main="Histogram of article length" )
```



2 (a) (ii)

This graph looks more like a normal distribution with a bit of a left skew.

```
hist(log(articleLengths[,2]), main="Histogram of article length" )
```



2 (a) (iii)

The log scale would be better because we want to use a normal sampling distribution for the later questions, and it resembles a normal distribution more than the original histogram.

2 (b)

```
sampleMean <- mean(log(articleLengths[,2]))
sampleMean
```

```
## [1] 8.381909
```

```
sampleVariance <- var(log(articleLengths[,2]))
sampleVariance
```

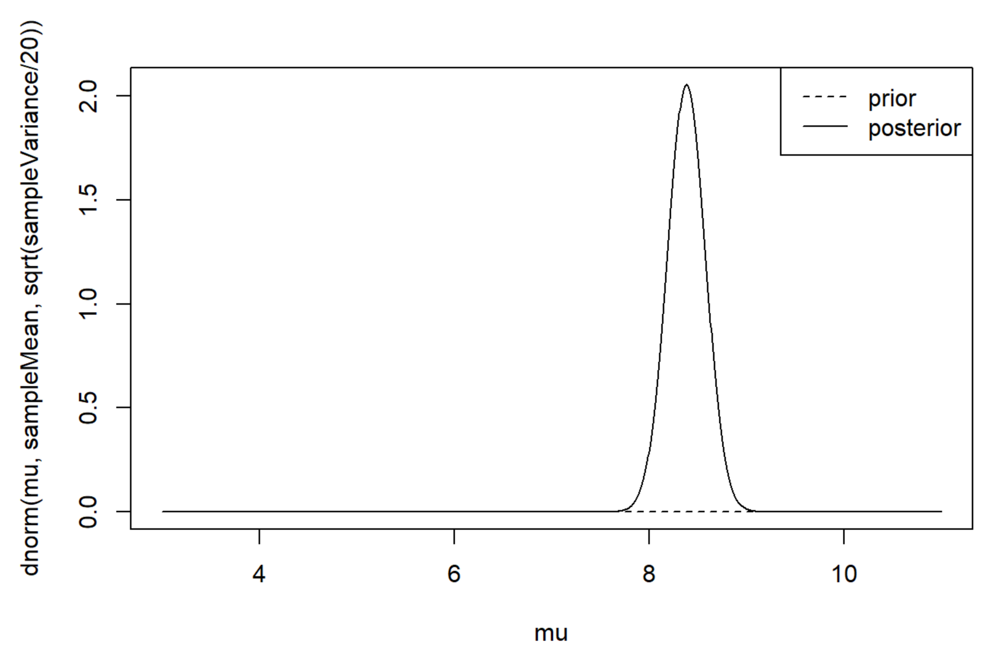
```
## [1] 0.7540378
```

2 (c) (i)

Since we are using a flat prior for μ, the posterior mean is the same as $\bar{y}$ which is 8.381909, the posterior variance is σ^2/n which is about 0.037701891, and the posterior precision is the reciprocal of the variance so it is about 26.5238685243.

2 (c) (ii)

```
curve(dnorm(mu,sampleMean,sqrt(sampleVariance/20)), 3, 11, xname="mu", n=1000) #posterior
curve(dunif(mu,0,1), 3, 11, xname="mu", n=1000, lty=2, add=TRUE) #prior
legend ("topright", c("prior","posterior"), lty=2:1)
```



2 (c) (iii)

Posterior interval can be calculated as $\bar{y} \pm 1.96 \cdot \sqrt{\sigma^2/n}$ which gives an interval of $8.381909 \pm 1.96 \cdot \sqrt{0.037701891}$ which is about $(8.1877392521, 8.5760787479)$

2 (d) (i)

We will be treating μ and σ^2 as independent. Given the prior density, applying Bayes Rule implies $\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$ and $\sigma^2|y \sim Inv_{\chi^2}(n-1, s^2)$. So first we will simulate values for σ^2 and then use them to simulate values for μ to calculate the posterior mean, variance, and precision.

```
sigma.2.sim <- (20-1) * sampleVariance / rchisq(1000, 20-1)
mu.sim <- rnorm(1000, sampleMean, sqrt(sigma.2.sim/20))
mean(mu.sim)
```

```
## [1] 8.366347
```

```
var(mu.sim)
```

```
## [1] 0.03866679
```

```
1/var(mu.sim)
```

```
## [1] 25.86198
```

So posterior mean is approximately 8.383812, posterior variance is approximately 0.04192662, and posterior precision is approximately 23.8512.

2 (d) (ii)

```
quantile(mu.sim, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 7.951993 8.756819
```

So a 95% central posterior interval for μ is $(7.972170, 8.778984)$

2 (d) (iii)

```
quantile(sigma.2.sim, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4467251 1.6526301
```

So a 95% central posterior interval for σ^2 is $(0.447000, 1.633389)$

2 (e) (i)

```
pred.sim <- rnorm(1000000, mu.sim, sqrt(sigma.2.sim))
exp(quantile(pred.sim,c(0.025, 0.975)))
```

```
##       2.5%      97.5%
##   659.1755 27995.2956
```

This is the 95% central posterior predictive interval on the original scale.

2 (e) (ii)

```
mean (pred.sim < log(min(articleLengths[,2])))
```

```
## [1] 0.022685
```

This is the posterior predictive probability that the length of a new row is less than the minimum article length in the data.

2 (e) (iii)

I am approaching this as calculating the probability that an event occurs exactly 0 times, where the event is that the length of a new randomly selected article is greater than the minimum length of the articles in the data. From the previous question, we have the probability that one event occurs. So we can use this to get the probability that it does not occur, and take it to the power of 20 to represent 20 different articles.

```
eventOccurs <- mean (pred.sim < log(min(articleLengths[,2])))
(1-eventOccurs)^20
```

```
## [1] 0.6319625
```

This is the probability that at least one out of 20 new randomly selected articles has a length less than the length seen in the data. This is the same as the probability that the minimum length out of 20 new randomly selected articles is less than the minimum length of the articles in the data.