Aman Arora (amana4@illinois.edu)

Answer 1:

a) **Movie 1**: Let p1 be the probability of positive reviews for movie 1.
y1 be observed positive reviews for movie 1.

p(p1 I y1) = p(y1|p1) * p(p1)/ p(y1)

Since p(p1) ~U(0,1) = > p(p1) = 1

=> p(p1|y) ~ p(y1|p1) = C(500,425) $p1^{425} \quad (1-p1)^{75}$

=> p(p1|y) = beta(426,76)

Normalizing constant = Γ(426+ 76) / Γ(426)*Γ(76)

Movie 2: Let p2 be the probability of positive reviews for movie 2.
y2 be observed positive reviews for movie 2.

p(p2 I y2) = p(y2|p2) * p(p2)/ p(y2)

Since p(p2) ~U(0,1) = > p(p2) = 1 and

=> p(p2|y) ~ p(y2|p2) = C(9,1) $(p2)^{9} \quad (1-p2)$

=> p(p2|y) = beta(10,2)

Normalizing constant = Γ(10+ 2) / Γ(10)*Γ(2) = 110

b) **Mean**:
p(p1) ~ Beta(426,76) => mean(p1) = 426/(426+76) = 0.849
p(p2) ~ Beta(10,2) => mean(p2) = 10/(10+2) = 0.833
Movie 1 ranks higher in posterior mean.

Median:
p(p1) ~ Beta(426,76) => median(p1) = qbeta(.5, 426,76) = .849
p(p2) ~ Beta(10,2) => median (p2) = qbeta(.5, 10,2) = .852
Movie 2 ranks higher in posterior median.

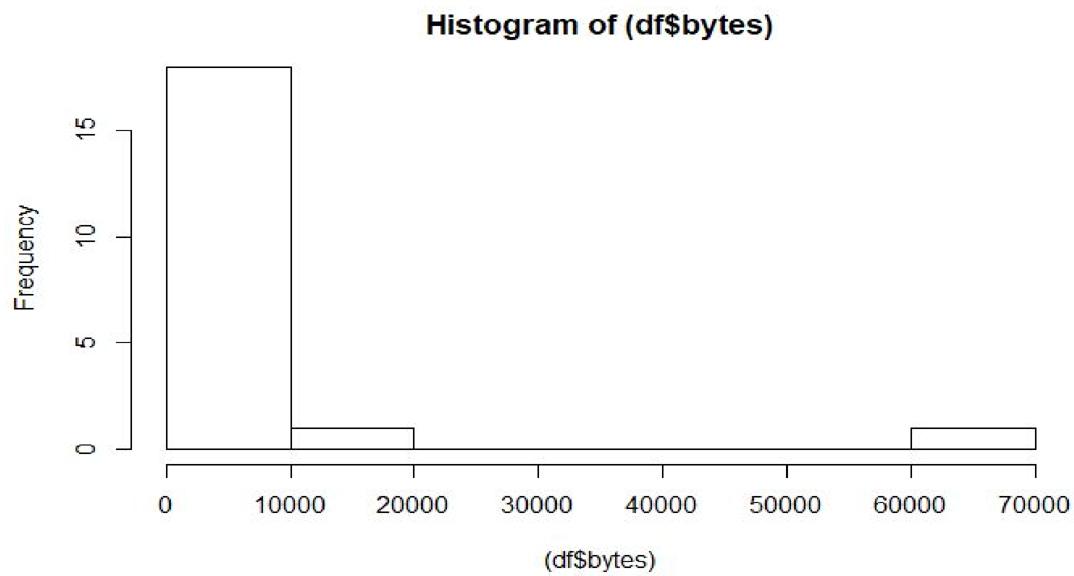
Mode:
p(p1) ~ Beta(426,76) => mode(p1) = (426-1)/(426+76-2) = .85
p(p2) ~ Beta(10,2) => mode(p2) = (10-1)/(10+2-2) = .9
Movie 2 ranks higher in posterior mode.

Answer 2

a) (i) **R-code**

```
df<- read.table("randomwikipedia.txt")
hist((df$bytes))
```

The distribution is highly skewed to the right with most articles in bin with length 0 and 10000, and few articles between 10000-20000 and 60000-70000 bins. Rest of the bins are empty.

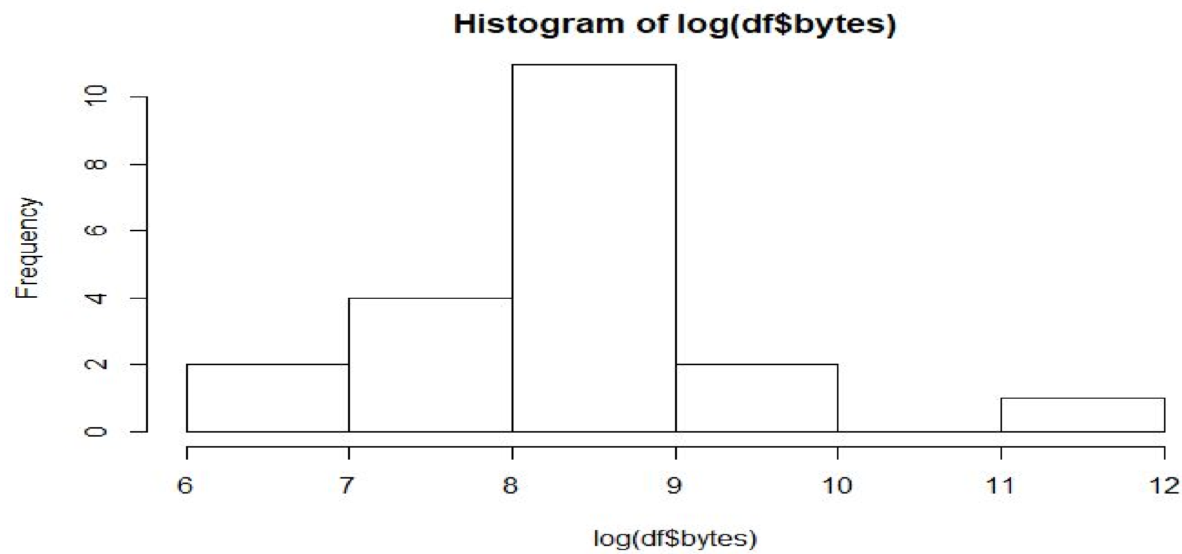


(ii) **R-code**

```
df<- read.table("randomwikipedia.txt")
hist(log(df$bytes))
```

The data after scaling with log is much less skewed and almost normally distributed.

iii) It will be better to use log scaling of data since the histogram of the log scaled data will likely be a good fit with a normal distribution and thus more amenable to further analysis.



b)

Mean

```
df<- read.table("randomwikipedia.txt")
mean(log(df$bytes))
8.331264
```

Variance

```
df<- read.table("randomwikipedia.txt")
var(log(df$bytes))
0.9600394
```

c)

For flat prior the posterior distribution of μ converge converges in distribution to a normal with a mean equal to sample mean and variance equal to sample variance.

Thus $\mu \rightarrow N\left(\bar{y}, \frac{\sigma^2}{n}\right)$

$$\bar{y} = 8.331 \; and \; \sigma^2 = 0.96, \; n = 20 \; => \frac{\sigma^2}{n} = .048$$

Thus μ tends to be normal distributed with mean = 8.331 and variance = .0048 as the variance of prior approaches infinity (for flat prior)

i)Thus Posterior mean of μ = 8.331, posterior variance of μ = .048, posterior precision of μ = 20.833

ii) **R code**

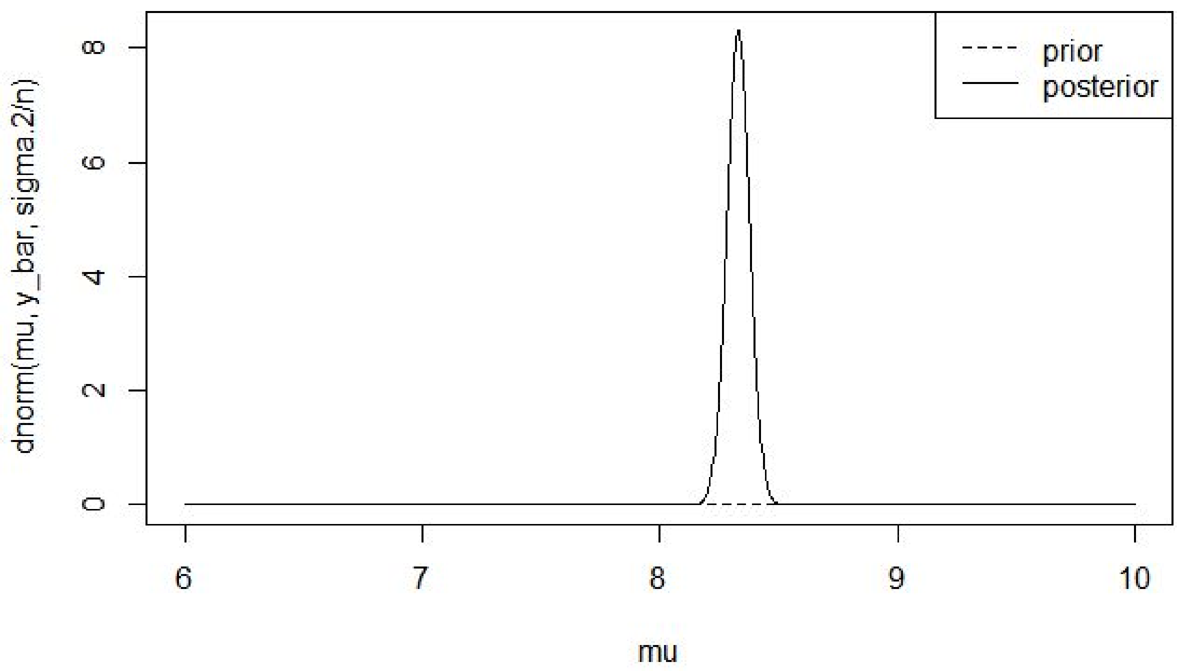
y_bar = 8.331264
sigma.2 = 0.9600394
n = 20

#Posterior
curve(dnorm(mu,y_bar,sigma.2/n), 6,10,xname ="mu",n=1000)

#Approximate Flat Prior drawn using very high variance normal dist (.Machine$double.xmax - largest floating point number)
curve(dnorm(mu,0,.Machine$double.xmax), 6,10,xname ="mu",n=1000,add = TRUE,lty=2)

#add legend
legend("topright",c("prior","posterior"), lty = 2:1)

Plot



III) 95% Posterior interval is (for normal distribution)

$$\bar{y} \pm 1.96\sqrt{\frac{\sigma^2}{n}}$$

(7.901841, 8.760687)

d)

i) Posterior mean/variance and precision of mu

```
n = 20
df = read.csv("randomwikipedia.txt",header = TRUE,sep = "")
s.2 = var(log(df$bytes))
y_bar = mean(log(df$bytes))
```

#Simulate distribution of variance/mean of mu

```
sigma.2.sim <- (n-1)*s.2/rchisq(10000,n-1)
mu.sim <-rnorm(10000,y_bar,sqrt(sigma.2.sim/n))
```

#posterior mean, variance and precision of mu

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

Mean 8.330382
Variance: 0.0545595
Precision 18.32862

ii) **95% posterior interval for mu**

```
quantile(mu.sim,c(.025,.975))
```

2.5%   97.5%
7.873134 8.786715

Iii **95% posterior interval for sigma^2**

```
quantile(sigma.2.sim,c(.025,.975))
```

  2.5%   97.5%
0.5556223 2.0489778

e)

i)

R-code

```
n = 20
df = read.csv("randomwikipedia.txt",header = TRUE,sep = "")
s.2 = var(log(df$bytes))
y_bar = mean(log(df$bytes))
```

#Simulate distribution of variance/mean of mu

Aman Arora (amana4@illinois.edu)

```
sigma.2.sim <- (n-1)*s.2/rchisq(10^6,n-1)
mu.sim <-rnorm(10^6,y_bar,sqrt(sigma.2.sim/n))
```

#Generate 100,000 post predictive samples

```
post.pred.sim.log <-rnorm(10^6, mu.sim, sqrt(sigma.2.sim) )
```

#Convert to linear scale

```
post.pred.sim = exp(post.pred.sim.log)
```

#95% central posterior interval of linear samples

```
quantile(post.pred.sim,c(.025,.975))
```

```
     2.5%      97.5%
 511.6614 33998.0668
```

ii) We know that a single posterior predicted observation $\bar{y}$ ~ N(μ, σ^2). We will use forward simulation to generate the samples and probability

R-code

```
n = 20
df = read.csv("randomwikipedia.txt",header = TRUE,sep = "")
s.2 = var(log(df$bytes))
y_bar = mean(log(df$bytes))
```

#Simulate distribution of variance/mean of mu

```
sigma.2.sim <- (n-1)*s.2/rchisq(10^6,n-1)
mu.sim <-rnorm(10^6,y_bar,sqrt(sigma.2.sim/n))
```

#Generate post predictive samples

```
post.pred.sim.log <-rnorm(10^6, mu.sim, sqrt(sigma.2.sim) )
```

#convert to linear scale

```
post.pred.sim = exp(post.pred.sim.log)
```

#Find the Posterior predictive probability that a single prediction is above the max value

```
mean(post.pred.sim> max(df$bytes))
```

Answer: 0.006856

III) For 20 observations problem, lets first calculate the probability of complementary event first i.e. probability that none of the 20 new random observation exceeds the maximum of prior data

We know that a single posterior predicted observation $\bar{y}$ ~ N(μ, σ^2) and from solution c (ii) above

P($\bar{y}$ >max(data)) = 0.006856 (for single observation)
= > P($\bar{y}$ <max(data)) = 1- 0.006856 = 0.993144

For 20 observation probability that all 20 observations <max (since they are iid) =
P($\bar{y}$ <max)^20 = 0.993144^20 = 0.871454

=> P(at least one of 20 observation >prior max) = P(max(20 observations) > max(prior data))
= 1 - P(all 20 observations are less than prior max)
= 1- 0.871454 = 0.128546