STAT 578 - Fall 2019 - Assignment 1

Frederick (Eric) Ellwanger - fre2

September 14, 2019

Exercise 1

Consider the following hypothetical scenario:

```
Movie 1: 150 positive reviews out of 200 (75%)
Movie 2: 4 positive reviews out of 5 (80%)
```

Assume that reviews of Movie i are independent with a common probability p_i of being positive (depending on the movie).
Assume a $U(0,1)$ prior on each p_i.

(1)(a) Determine the posterior distribution of p_1 and of p_2 (separately).

The sampling distribution is a binomial distribution with n = 200 and $\theta = 150$

$y_1|\theta_1 \sim Bin(n = 200, \theta = 150)$

$p_1(\theta_1|y_1) \propto p(\theta_0) * p(y_1|\theta_1)$

Since the prior is assumed to be a uniform distribution between 0 and 1, it's density will be 1.

$p_1(\theta_1|y_1) \propto 1 * \binom{200}{150} * \theta_1^{150} * (1-\theta_1)^{200-150}$

the posterior distribution, p_1 can be written as:

$\theta_1|y_1 \sim Beta(\alpha = 151, \beta = 51)$

or we can write the density as

$\mathbf{p_1}(\boldsymbol{\theta_1}|\mathbf{y_1}) \propto \boldsymbol{\theta}_\mathbf{1}^{\mathbf{150}} * (\mathbf{1}-\boldsymbol{\theta}_\mathbf{1})^{\mathbf{50}}, \qquad 0 < \theta_1 < 1$

Following similar reasoning the posterior distribution, p_2, can be written:

$\theta_2|y_2 \sim Beta(\alpha = 5, \beta = 2)$

or we can write the density as

$\mathbf{p_2}(\boldsymbol{\theta_2}|\mathbf{y_2}) \propto \boldsymbol{\theta}_\mathbf{2}^{\mathbf{4}} * (\mathbf{1}-\boldsymbol{\theta}_\mathbf{2})^{\mathbf{1}}, \qquad 0 < \theta_1 < 1$

(1)(b) Which movie ranks higher according to posterior mean, posterior median, and posterior mode?

```
#P1 Posterior Mean
alpha1 = 151
beta1  = 51
(mean_p1 = alpha1 / (alpha1 + beta1))
```

```
## [1] 0.748
```

posterior mean (p1) $\approx$ 0.748

```
#P2 Posterior Mean
alpha2 = 5
beta2 = 2
(mean_p2 = alpha2 / (alpha2 + beta2))
```

```
## [1] 0.714
```

posterior mean (p2) $\approx$ 0.714

Movie 1 would rank higher according to posterior mean (0.748 $>$ 0.714).

Which movie ranks higher according to posterior median?

```r
#P1 Median
(median_p1 = qbeta(0.5, alpha1, beta1))
```

```
## [1] 0.748
```

posterior median (p1) $\approx$ 0.748

```r
#P2 Median
(median_p2 = qbeta(0.5, alpha2, beta2))
```

```
## [1] 0.736
```

posterior median (p2) $\approx$ 0.736

Movie 1 would rank higher according to posterior median (0.748 $>$ 0.736).

Which movie ranks higher according to posterior mode?

```r
#P1 Mode
(mode_p1 = (alpha1 - 1) / (alpha1 + beta1 - 2))
```

```
## [1] 0.75
```

posterior mode (p1) $\approx$ 0.75

```r
#P2 Mode
(mode_p2 = (alpha2 - 1) / (alpha2 + beta2 - 2))
```

```
## [1] 0.8
```

posterior mode (p2) $\approx$ 0.8

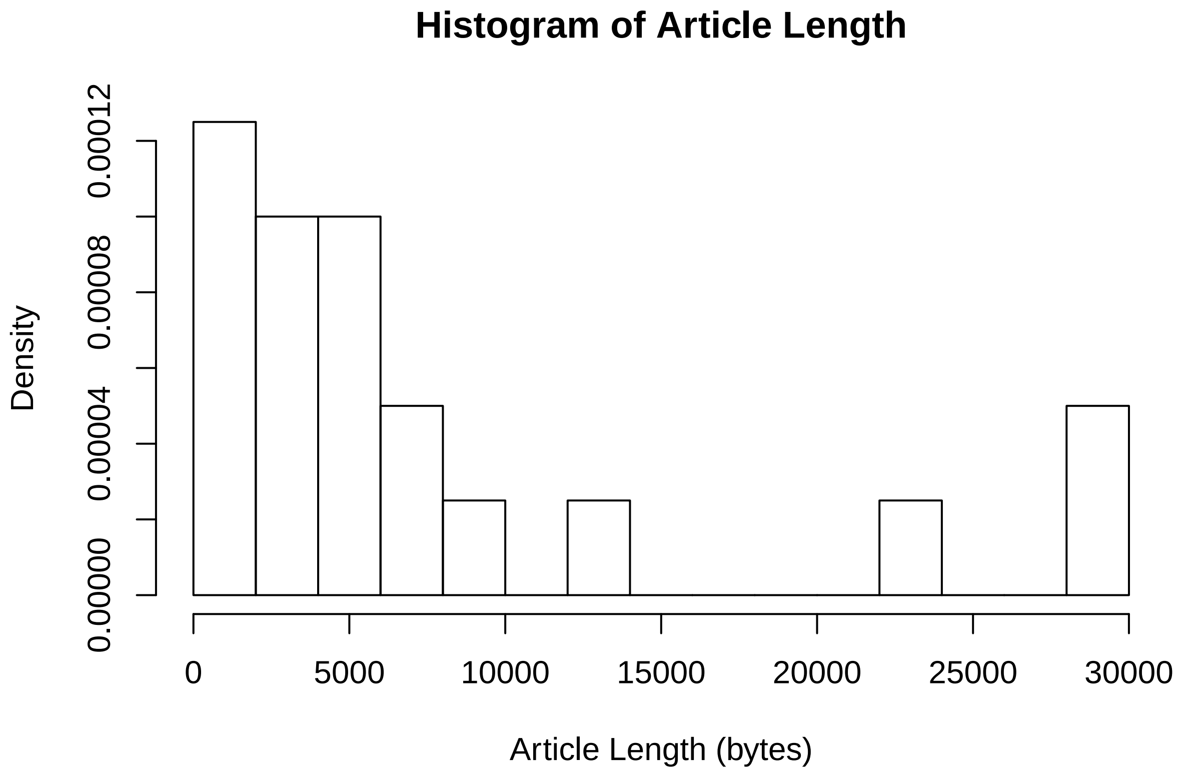
Movie 2 would rank higher according to posterior mode (0.8 $>$ 0.75).

Exercise 2

File randomwikipedia.txt contains the ID number and number of bytes in length for 20 randomly selected English Wikipedia articles.

(2)(a)(i) Draw a histogram of article length, and describe the distribution.

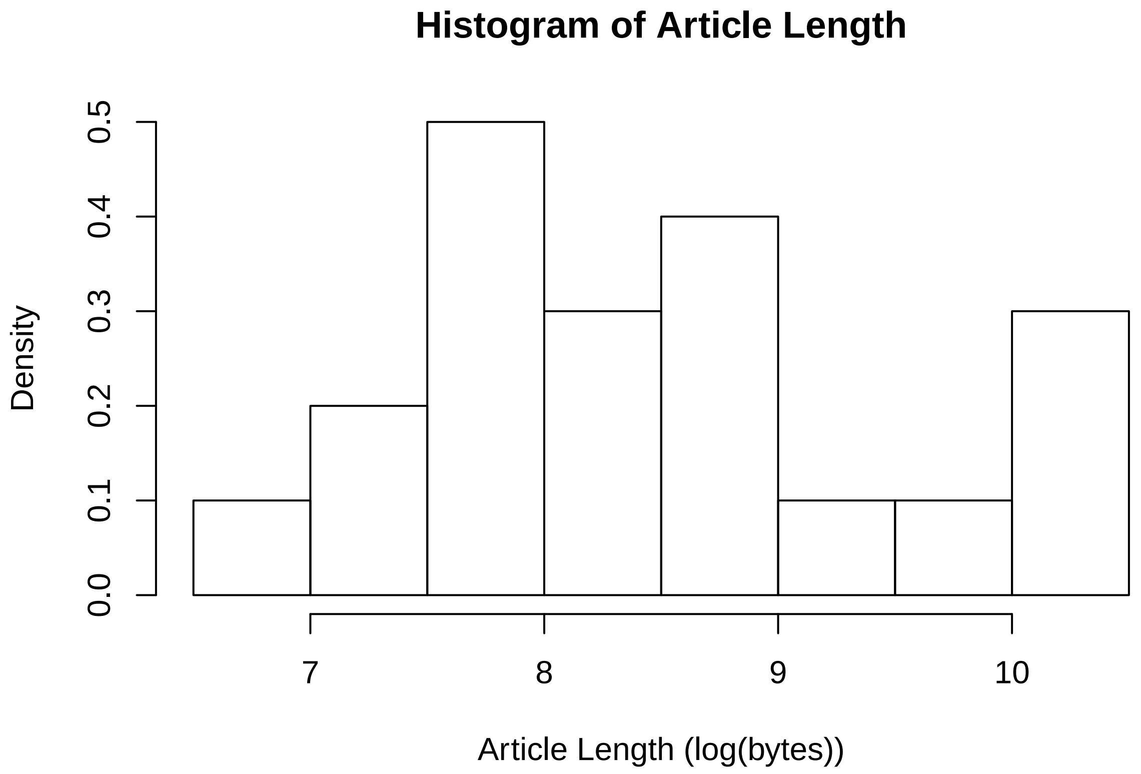
```r
hist(wiki_data$bytes, breaks = 10, main = "Histogram of Article Length",
     freq = FALSE, xlab = "Article Length (bytes)")
```



The distribution of the data is skewed right with most of the data falling below 15,000 bytes and some data above 20,000 bytes. The mean of the data is roughly 7,700 bytes, the median is roughly 4,400 bytes.

(2)(a)(ii) Transform article length to the (natural) log scale. Then re-draw the histogram and describe the distribution.

```
hist(log(wiki_data$bytes), breaks = 10, main = "Histogram of Article Length",
     freq = FALSE, xlab = "Article Length (log(bytes))")
```



The distribution of the data still has some structure to it, but looks more compact. The mean of the data is approximately 8.45, the median is approximately 8.4

(2)(a)(iii) Based on your histograms, explain why the log scale would be better to use for the remainder of the analysis. (Read below.)

The log scaled data has a better shape and less skew, making it a better candidate to be modeled by a normal distribution.

(2)(b) Let y_i be length of article i on the log scale (i.e., the natural logarithm of the number of bytes). Compute the sample mean and sample variance of $y_1, ..., y_{20}$.

```
#Sample Mean
(y_bar = mean(log(wiki_data$bytes)))
```

```
## [1] 8.45
```

Sample mean ($\bar{y}$) $\approx$ 8.454

```
#Sample Variance
(s.2 = var(log(wiki_data$bytes)))
```

```
## [1] 1.01
```

Sample variance (s^2) $\approx$ 1.015

In the remaining parts, assume the $y_i's$ have a normal sampling distribution with mean μ and variance σ^2.

(2)(c) Assume σ^2 is known to equal the sample variance. Consider a flat prior for μ. Use it to:

(2)(c)(i) Compute the posterior mean, posterior variance, and posterior precision of μ.

$p(\mu|y) \propto p(\mu)p(y|\mu) \propto 1 * p(y|\mu)$
$p(\mu|y) \propto exp^{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2}$

Posterior mean $= \bar{y} \approx$ 8.454

```
#Posterior variance of mu
n = nrow(wiki_data)
(sigma.n.2 = s.2 / n)
```

```
## [1] 0.0507
```

Posterir variance of $\mu = \sigma^2/n \approx$ 0.051

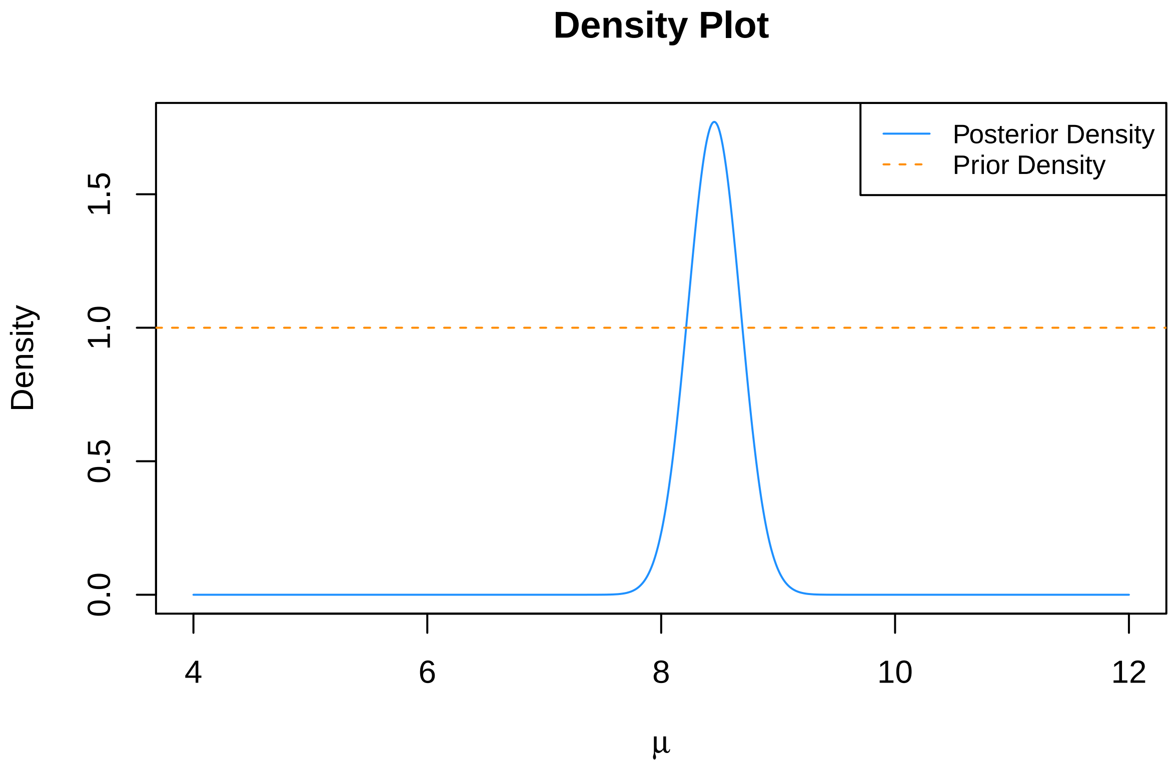
```
#Posterior precision of mu
(p.precision = n / s.2)
```

```
## [1] 19.7
```

Posterior precision of $\mu = \frac{n}{\sigma^2} \approx$ 19.71

(2)(c)(ii) Plot the prior density and the posterior density of μ together in a single plot. Label which is which.

```
curve(dnorm(x, y_bar, sqrt(sigma.n.2)), 4, 12, n = 1000, col = 'dodgerblue',
      ylab = "Density", xlab = expression(mu), main = "Density Plot")
abline(h=1, col = 'darkorange', lty = 2)
legend("topright", c("Posterior Density", "Prior Density"),
       col=c("dodgerblue", "darkorange"), lty=1:2, cex=0.8)
```



(2)(c)(iii) Compute a 95% central posterior interval for μ.

```
#95% central posterior interval
#(conf_int = y_bar + c(-1, 1) * 1.96 * sqrt(sigma.n.2))
(conf_int = qnorm(c(0.025, 0.975), y_bar, sqrt(sigma.n.2)))
```

```
## [1] 8.01 8.90
```

The 95% central posterior interval fo μ is $\approx$ (8.013, 8.896)

(2)(d) Now let μ and σ^2 have prior

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1} \qquad\qquad (\sigma^2) > 0$$

Use it to:

(2)(d)(i) Compute the posterior mean, posterior variance, and posterior precision of μ. (If you cannot compute explicitly, use a good computational approximation.)

This simplifies the distribution of μ to: $\mu|\sigma^2, y \sim \mathcal{N}(\bar{y}, \sigma^2/n)$
Posterior mean of $\mu = \bar{y} \approx$ 8.454

```
#Posterior variance of mu
set.seed(19690223)
post.sigma.2.sim = (n-1) * s.2 / rchisq(1000000, n-1)
post.mu.sim = rnorm(1000000, y_bar, sqrt(post.sigma.2.sim / n))
(mu_var = var(post.mu.sim))
```

```
## [1] 0.0567
```

Posterior variance of $\mu \approx$ 0.057

```r
#Posterior precision of mu
(p.precision = 1 / mu_var)
```

```
## [1] 17.6
```

Posterior precision of $\mu \approx$ 17.641

(2)(d)(ii) Approximate a 95% central posterior interval for μ.

```r
#95% central posterior interval for mu
(conf_int = quantile(post.mu.sim, c(0.025, 0.975)))
```

```
##  2.5% 97.5%
##  7.98  8.92
```

The 95% central posterior interval for μ is $\approx$ (7.982, 8.924)

(2)(d)(iii) Approximate a 95% central posterior interval for σ^2.

```r
#95% central posterior interval for sigma^2
(conf_int = quantile(post.sigma.2.sim, c(0.025,0.975)))
```

```
##  2.5% 97.5%
## 0.587 2.163
```

The 95% central posterior interval for σ^2 is $\approx$ (0.587, 2.163)

(2)(e) Assume the prior of the previous part. Use simulation in R to answer the following, based on 1,000,000 draws from the posterior.

(2)(e)(i) Approximate a 95% central posterior predictive interval for the length (in bytes) of a single (new) randomly drawn article. (Note that this is on the original scale, not the log scale.)

```r
#95% central posterior predictive interval for single new draw
set.seed(19690223)
post.sigma.2.sim = (n-1) * s.2 / rchisq(1000000, n-1)
post.mu.sim = rnorm(1000000, y_bar, sqrt(post.sigma.2.sim / n))
post.pred.sim = rnorm(1000000, post.mu.sim, sqrt(post.sigma.2.sim))
(conf_int = quantile(exp(post.pred.sim), c(0.025,0.975)))
```

```
##  2.5% 97.5%
##   540 40601
```

The 95% central posterior interval for length in bytes of a single (new) randomly drawn article is $\approx$ (540, 40601)

(2)(e)(ii) Approximate the posterior predictive probability that the length of a single (new) randomly drawn article will exceed the maximum article length in the data.

```r
#Posterior predictive probability length > max length
(prob = mean(post.pred.sim > max(log(wiki_data$bytes))))
```

```
## [1] 0.0489
```

The posterior predicictive proability Pr($new_{length} > max(log(y's))$) is $\approx$ 0.049

(2)(e)(iii) Approximate the posterior predictive probability that the maximum length of 20 (new) randomly drawn articles will exceed the maximum article length in the data. (Be careful! The 20 randomly drawn articles have the same value for μ and for σ^2.)

```r
set.seed(19690223)

#Get maximum of 20 random draws from normal dist. w/ same mean, variance
get_max_20 = function(n, y_bar, s.2){
  post.sigma.2.sim = (n-1) * s.2 / rchisq(1, n-1)
  post.mu.sim = rnorm(1, y_bar, sqrt(post.sigma.2.sim / n))
  max(rnorm(20, post.mu.sim, sqrt(post.sigma.2.sim)))
}

#Repeat finding max of 20 draws, 1,000,000 times
ppp = replicate(1000000, get_max_20(n, y_bar, s.2))
(prob = mean(ppp > max(log(wiki_data$bytes))))
```

```
## [1] 0.553
```

The posterior predicictive proability $Pr(max(20new_{lengths}) > max(log(y's))$ **is $\approx$ 0.553**