Michael Leibert

Math 640

Homework 2

1. Let $y_i$, $i = 1, \ldots, n$ be iid normally distributed with mean $\mu$ and known variance $\sigma^2$. Assuming a conjugate prior for $\mu$ with parameters $\mu_0$ and $\tau_0^2$, fully specify the posterior distribution.

$$\mathcal{L}(\mathbf{y}|\mu) = \prod_{i=1}^{n} P(y_i|\mu) \qquad\qquad \pi(\mu) = \left(2\pi\tau_0^2\right)^{-\frac{1}{2}} \exp\left[\frac{-(\mu - \mu_0)^2}{2\tau_0^2}\right]$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \qquad\qquad \propto \exp\left[\frac{-(\mu - \mu_0)^2}{2\tau_0^2}\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

$$P(\mu|\mathbf{y}) \propto \exp\left[\frac{-(\mu - \mu_0)^2}{2\tau_0^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{(\mu - \mu_0)^2}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\mu + \mu^2\right)\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{\mu^2}{\tau_0^2} + \frac{n\mu^2}{\sigma^2} - \frac{2\mu\mu_0}{\tau_0^2} - \frac{2\mu\sum_i y_i}{\sigma^2} + \frac{\mu_0^2}{\tau_0^2} + \frac{\sum_i y_i^2}{\sigma^2}\right)\right]$$

$$= \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_i y_i}{\sigma^2}\right) + \frac{\mu_0^2}{\tau_0^2} + \frac{\sum_i y_i^2}{\sigma^2}\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\mu^2 - 2\mu\frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_i y_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} + \frac{\frac{\mu_0^2}{\tau_0^2} + \frac{\sum_i y_i^2}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\left(\mu - \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2 + \frac{\frac{\mu_0^2}{\tau_0^2} + \frac{\sum_i y_i^2}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} - \left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\left(\mu - \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2\right]\right) \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\left(\frac{\frac{\mu_0^2}{\tau_0^2} + \frac{\sum_i y_i^2}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right) - \left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\left(\mu - \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2\right]\right) h(y)$$

$$P(\mu|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]\left[\left(\mu - \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)^2\right]\right)$$

Thus, the posterior is distributed $N(\mu_1, \tau_1^2)$, where $\mu_1 = \dfrac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$ and $\tau_1^2 = \left(\dfrac{1}{\tau_0^2} + \dfrac{n}{\sigma^2}\right)^{-1}$

That is,

$$\mu|\mathbf{y} \sim N\left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \left[\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right]^{-1}\right).$$

2. Let $x_i$, $i = 1, \ldots, n$ be iid exponential with rate $\lambda$. Find a conjugate prior for the exponential likelihood.

$$\mathcal{L}(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \lambda \exp\left(-\lambda x_i\right)$$

$$= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

$$= \lambda^n \exp\left(-\lambda n\bar{x}\right)$$

$$\pi(\lambda) \propto c(\lambda)^\eta \exp\left[w(\lambda)\nu\right]$$

$$= \lambda^\eta \exp\left(-\lambda\nu\right) \qquad \text{Let } \eta = \alpha - 1$$

$$= \underbrace{\lambda^{\alpha-1} \exp\left(-\lambda\nu\right)}_{Gamma(\alpha,\nu) \text{ kernel}}$$

$$P(\mathbf{x}|\lambda) = \mathcal{L}(\lambda|\mathbf{x})\pi(\lambda)$$

$$\propto \lambda^n \exp\left(-\lambda n\bar{x}\right) \lambda^{\alpha-1} \exp\left(-\lambda\nu\right)$$

$$= \lambda^{n+\alpha-1} \exp\left(-\lambda n\bar{x} - \lambda\nu\right)$$

$$= \underbrace{\lambda^{n+\alpha-1} \exp\left(-\left[n\bar{x} + \nu\right]\lambda\right)}_{Gamma\left(n+\alpha, n\bar{x}+\nu\right) \text{ kernel}}$$

The prior, distributed $Gamma\left(\alpha, \nu\right)$, is conjugate to the posterior, distributed $Gamma\left(n + \alpha, n\bar{x} + \nu\right)$

3. Let $Z$ be a geometric random variable with probability of success $\theta$. Here note that $Z$ is the number of failures until the first success. Thus the parameterization of the geometric we'll use is

$$p(Z|\theta) = (1 - \theta)^Z \theta.$$

Find Jeffreys' prior for the geometric likelihood.

$$\mathcal{L}(Z|\theta) = (1 - \theta)^Z \theta$$

$$\ell(Z|\theta) = Z \log(1 - \theta) + \log(\theta)$$

$$\frac{\partial \ell}{\partial \theta} = -\frac{Z}{(1 - \theta)} + \frac{1}{\theta}$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{Z}{(1 - \theta)^2} - \frac{1}{\theta^2}$$

$$-E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = -E\left[-\frac{Z}{(1 - \theta)^2} - \frac{1}{\theta^2}\right]$$

$$= \frac{E[Z]}{(1 - \theta)^2} + \frac{1}{\theta^2}$$

$$= \frac{1 - \theta}{\theta} \frac{1}{(1 - \theta)^2} + \frac{1}{\theta^2}$$

$$= \frac{\theta + (1 - \theta)}{\theta^2(1 - \theta)}$$

$$= \theta^{-2}(1 - \theta)^{-1}$$

$$\left[J(\theta)\right]^{\frac{1}{2}} = \theta^{-1}(1-\theta)^{-\frac{1}{2}}$$

$$P(\theta|Z) = \left[(1-\theta)^Z \theta\right]\left[(1-\theta)^{-\frac{1}{2}}\theta^{-1}\right]$$

$$= (1-\theta)^{Z-\frac{1}{2}+1-1}\,\theta^{1-1}$$

$$= \theta^{1-1}\,(1-\theta)^{Z+\frac{1}{2}-1}$$

$$\theta|Z \sim Beta\left(1, Z+\frac{1}{2}\right)$$

## Analysis Exercises

4. Table contains data collected from three recent surveys regarding whether or not Americans approve of the job Congress is doing. Each sample was taken from registered voters (data courtesy of `realclearpolitcs.com`).

Table 1: Recent Congressional Approval Polls

| Poll | Date | Sample Size | Number who Approve |
|---|---|---|---|
| Economist/YouGov | 1/12 - 1/15 | 1289 | 142 |
| CNN | 12/6 - 12/9 | 919 | 175 |
| Monmouth University | 11/9 - 11/12 | 716 | 165 |

(a) For the Monmouth University poll above, conduct an analysis of the proportion of registered voters who approve of the job Congress is doing using the standard uniform prior. In your analysis, provide a brief discussion of the results along with an estimate of the posterior mean, 95% credible interval, and plot of the posterior distribution.

$$\mathcal{L}(x|\theta) \propto \theta^x(1-\theta)^{n-x} \qquad \pi(\theta) = 1$$

$$P(\theta|x) = \theta^x(1-\theta)^{n-x} \qquad \theta|x \sim Beta(x+1, n-x+1)$$
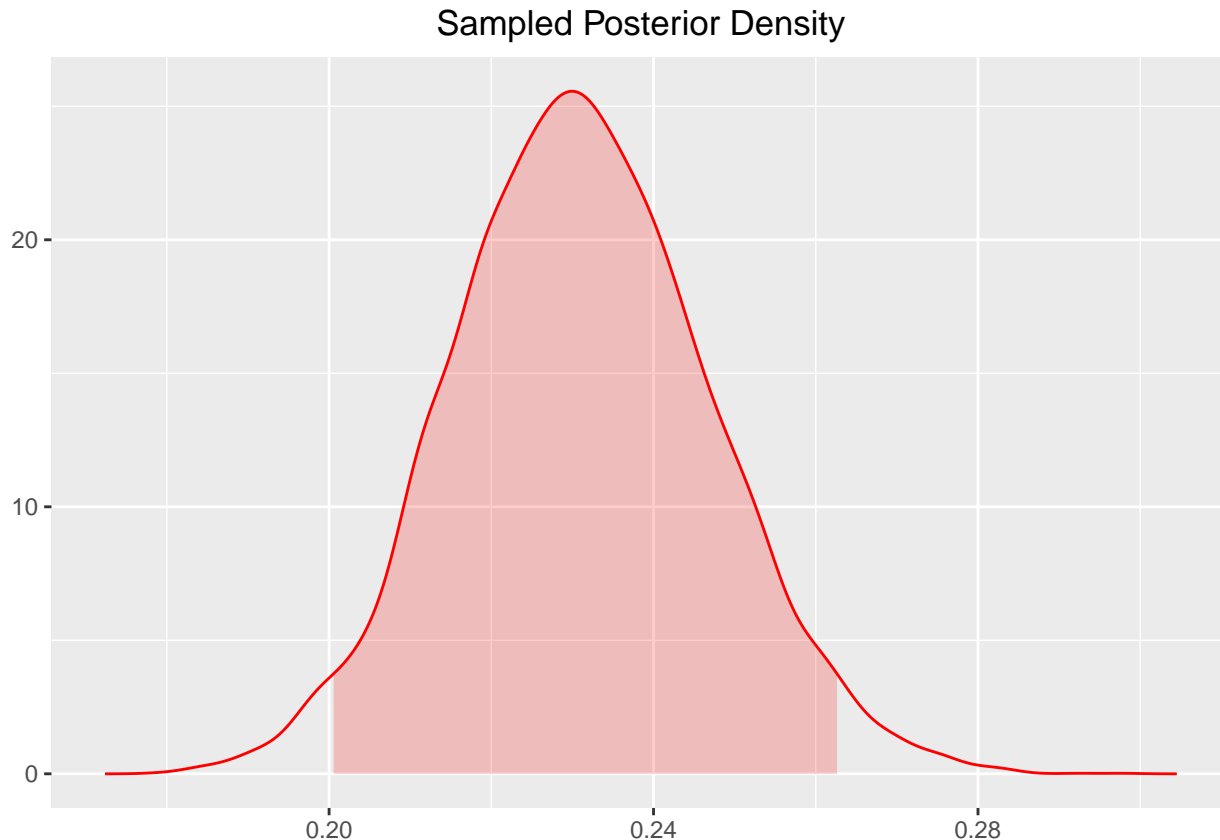
```
X <- 165
n <- 716
Nsim <- 10000

alpha1 = X + 1
Beta1 = n - X + 1

bs1  <- matrix( rbeta(Nsim, alpha1 , Beta1 ), , 1)

# posterior mean
mean( bs1 )
```

```
## [1] 0.2310757
```

```
# 95% credible interval
quantile(bs1, probs = c(0.5, 0.025, 0.975))
```

```
##        50%       2.5%      97.5%
## 0.2306799 0.2004849 0.2628391
```

### Sampled Posterior Density



The mean of the distribution is nearly identical to if we simply took the sample proportion, $\dfrac{165}{716} = 0.2304469$. This is probably reflecting that the standard uniform prior does not add any information to the estimate.

(b) Now analyze the CNN poll using the posterior distribution from part (a) as the new prior. As before, your analysis should give a brief discussion of the results along with an estimate of the posterior median, 95% credible interval, and plot of the posterior distribution.

$$\mathcal{L}(y|\theta) \propto \theta^y(1-\theta)^{m-y} \qquad\qquad \pi(\theta) = \theta^x(1-\theta)^{n-x}$$

$$P(\theta|y) = \theta^{x+y+1-1}(1-\theta)^{m+n-(x+y)+1-1} \qquad \theta|y \sim Beta(x+y+1, m+n-(x+y)+1)$$

```
Y <- 175
m <- 919

alpha2 = X + Y + 1
Beta2 = m + n - ( X + Y ) + 1

set.seed(1)
```
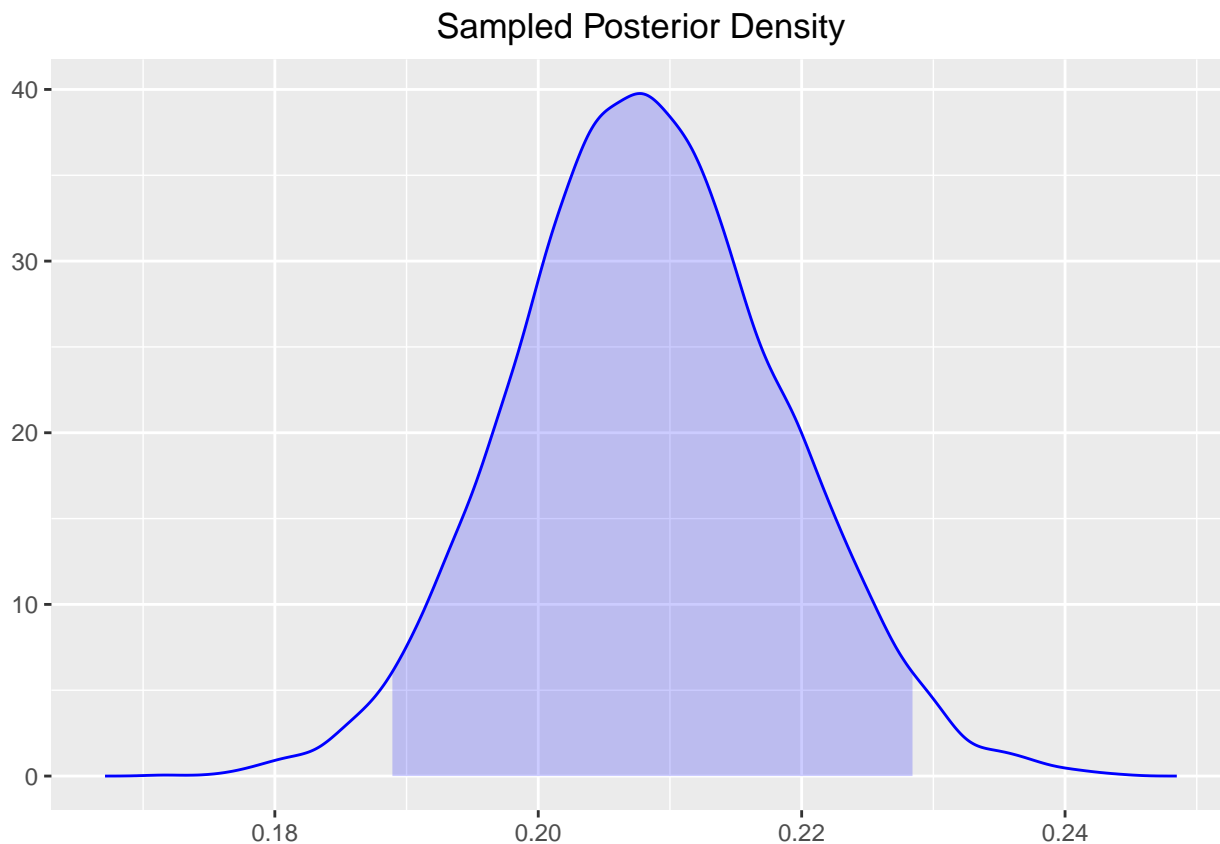
4

```
bs2 <- rbeta(Nsim, alpha2 , Beta2  )

median(bs2)
```

```
## [1] 0.2080713
```

```
# 95% credible interval
quantile(bs2, probs = c(0.5, 0.025, 0.975))
```

```
##       50%      2.5%     97.5%
## 0.2080713 0.1889030 0.2284873
```

### Sampled Posterior Density



For the CNN survey, the sample proportion is $\frac{175}{919} = 0.1904244$. While relatively close to the 19%, we note that the median is approximately 1.8 percentage points higher than the 19%. The use of the prior from the Monmouth University survey is providing some information that maybe the true value of the parameter is slightly higher than what the CNN survey is estimating.

The Credible interval is slightly tigher than before, probably reflecting the sample sizes were "combined." The previous one was approximately $\pm$ 3% while this one is about $\pm$ 2%.

(c) First analyze the Economist/YouGov poll using a standard uniform prior. Determine the posterior mean, 95% credible interval, and plot of the posterior distribution. Next, analyze the Economist/YouGov poll using the posterior distribution from part (b) as the prior. As before, your analysis should give a brief discussion of the results along with an estimate of the posterior mean, 95% credible interval, and plot of the posterior distribution. Compare the results of the analysis using a noninformative prior to the results using an informative prior be sure to present a graph showing both posteriors.

$$\mathcal{L}(z|\theta) \propto \theta^z (1-\theta)^{\ell-z} \qquad \pi(\theta) = 1$$

$$P(\theta|z) = \theta^z (1-\theta)^{\ell-z} \qquad \theta|z \sim Beta(z+1, \ell-z+1)$$

```
Z <- 142
l <- 1289

alpha3 =  Z + 1
Beta3 = l  -  Z   + 1

set.seed(1)
bs3 <-  rbeta(Nsim, alpha3 , Beta3  )

mean(bs3)
```
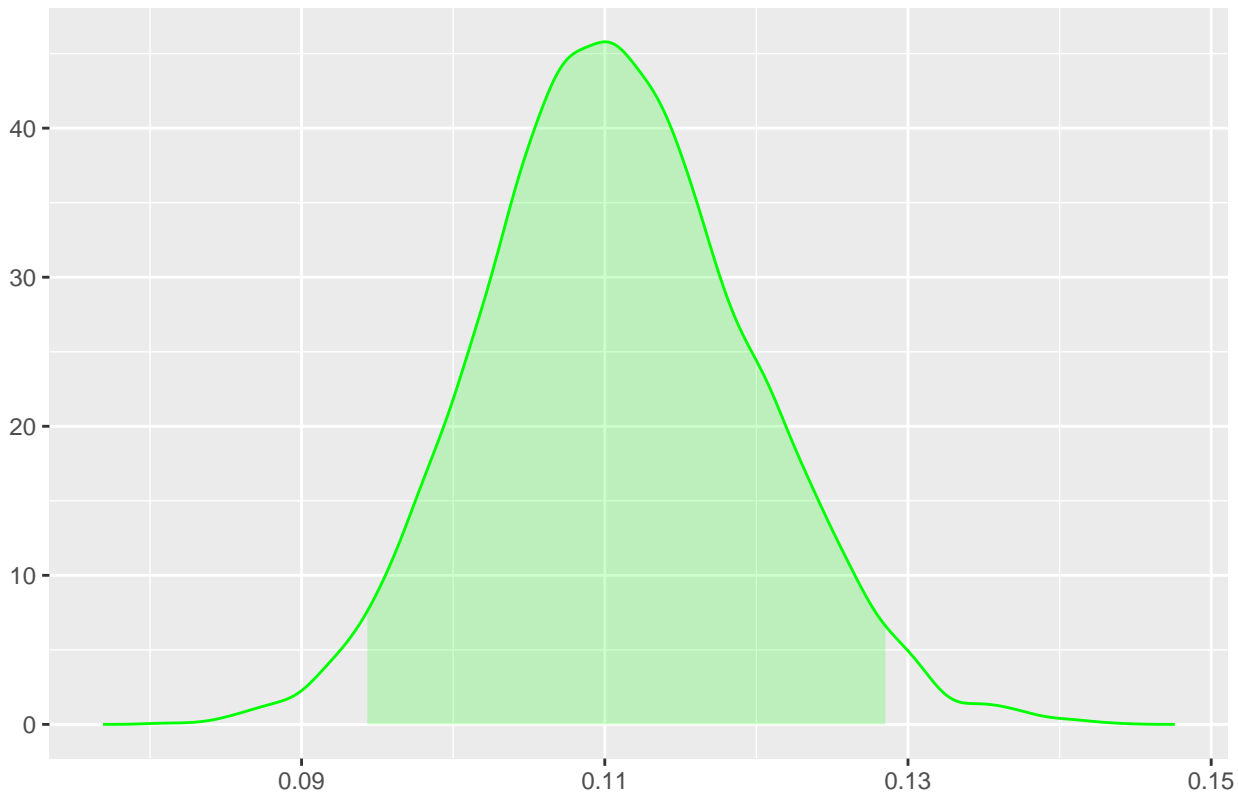
```
## [1] 0.1107962
```

```
# 95% credible interval
quantile(bs3, probs  = c(0.5, 0.025, 0.975))
```

```
##         50%        2.5%       97.5%
## 0.11048096 0.09422936 0.12858257
```

### Sampled Posterior Density



The discussion is similar to using the uniform prior on the Monmouth University survey. The mean of the distribution is nearly identical to if we simply took the sample proportion, $\frac{142}{1289} = 0.1101629$. This is probably

reflecting that the standard uniform prior does not add any information to the estimate.

However, we not that the estimate itself is considerably different than from the Monmouth University. We might put more trust in this survey because the sample size is almost twice as large. The credible interval is roughly $(\hat{\theta} - 0.016, \hat{\theta} + 0.018)$, smaller than the previous two.

$$\mathcal{L}(z|\theta) \propto \theta^z (1-\theta)^{\ell-z} \qquad\qquad \pi(\theta) = \theta^{x+y+1-1}(1-\theta)^{m+n-(x+y)+1-1}$$

$$P(\theta|y) = \theta^{x+y+z+1-1}(1-\theta)^{\ell+m+n-(x+y+z)+1-1} \qquad \theta|y \sim Beta(x+y+z+1, \ell+m+n-(x+y+z)+1)$$
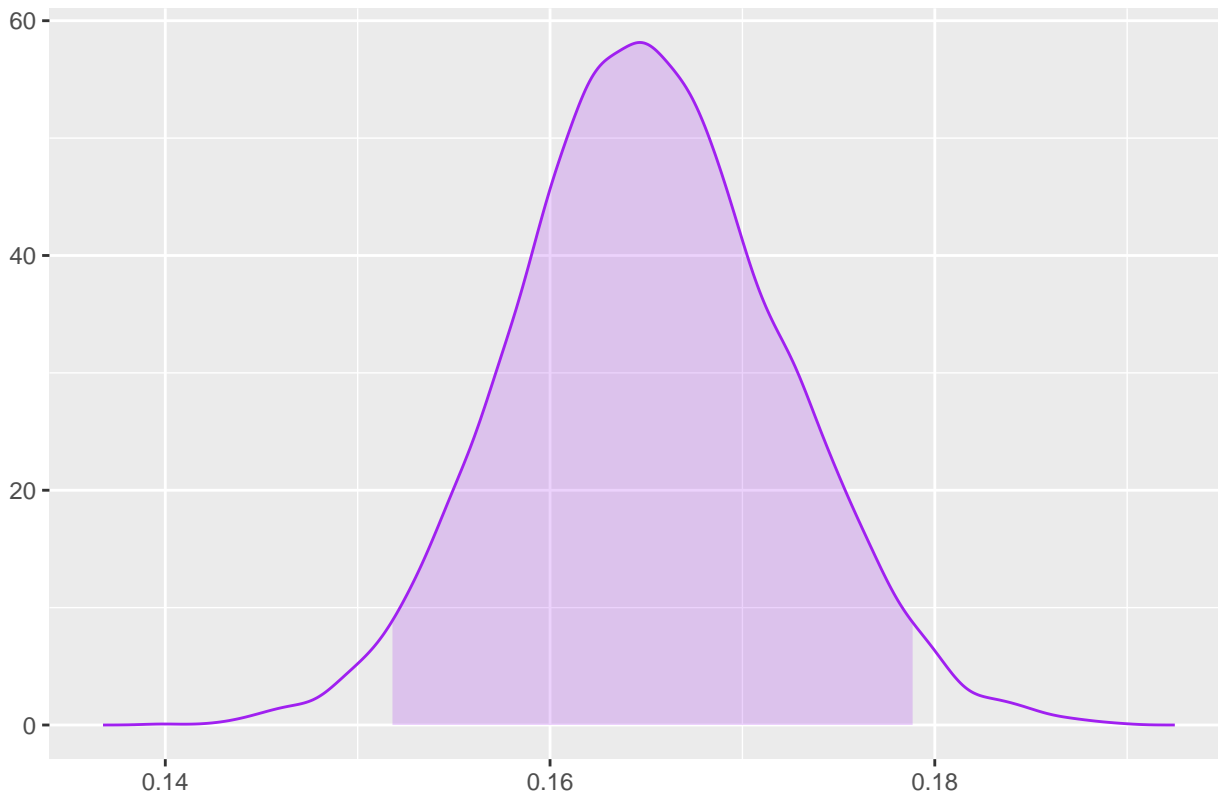
```
Z <- 142
l <- 1289

alpha4 = X + Y + Z + 1
Beta4 = l + m + n - ( X + Y + Z ) + 1

set.seed(1);  bs4   <-  rbeta(Nsim, alpha4 , Beta4 ) ;  mean(bs4)
```

```
## [1] 0.1650903
```

```
# 95% credible interval
quantile(bs4, probs = c(0.5, 0.025, 0.975))
```

```
##       50%       2.5%      97.5%
## 0.1649175 0.1518032 0.1788690
```



Sampled Posterior Density

In this last one, we have borrowed strength from the previous two surveys. It is as if we simply aggregated the data into one analysis to begin with.

The population parameter is most likely not nearly as high as the first two surveys suggest, and probably not nearly as low as the Economist/YouGov estimates. That is the model's, with the previous prior, estimate is about 5.4 percentage points higher than the estimate of the Economist/YouGov model that only used the standard uniform. Again, the credible interval shrinks as the "sample size" increases. In this analysis it is ± 1.3%.

We wonder, because there is a time component, that if combining the strength of the surveys is the wisest thing to do, but for our simple model, it seems sufficient enough.

All four plots
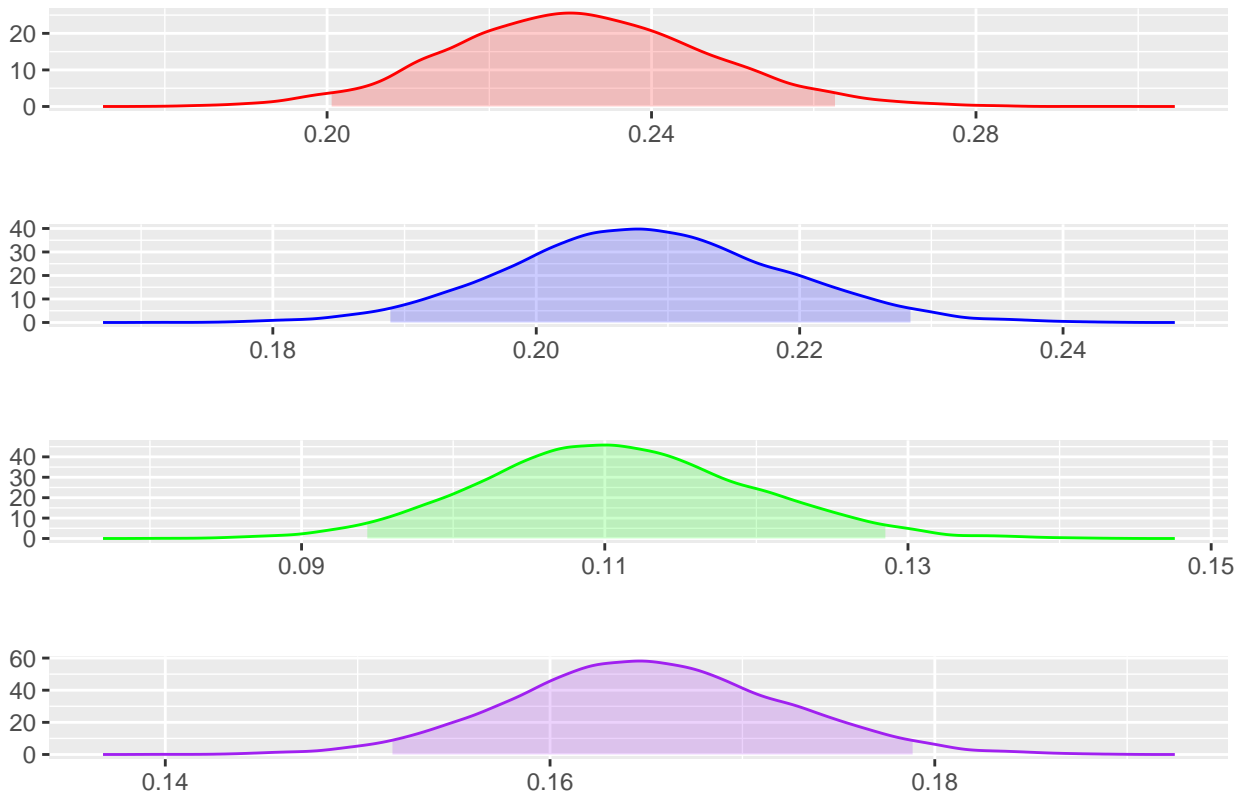
2. Consider data from a skin cancer prevention study found in the file `skin.txt` alongside this assignment. Let $x_i$ denote the number of skin cancers found on subject $i$ from the study. We can assume that $x_i$ are iid $Pois(\lambda)$ for $i = 1, \ldots, n$. Find Jeffrey's prior for $\lambda$ and use it it to determine the posterior distribution of $\lambda | x_1, \ldots, x_n$. The variable `numsc` contains the count of skin cancers found on each subject. Using this, calculate the posterior mean and 95% credible interval of the rate of skin cancers. Also provide a plot of the posterior distribution.

$$\mathcal{L}(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!}$$

$$= \exp(-n\lambda) \frac{\lambda^{n\bar{x}}}{\prod_i x_i!}$$

$$\propto \exp(-n\lambda) \lambda^{n\bar{x}}$$

$$\ell(\mathbf{x}|\lambda) \propto -n\lambda + n\bar{x}\log(\lambda)$$

$$\frac{\partial \ell}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{n\bar{x}}{\lambda^2}$$

$$-E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = -E\left[-\frac{n\bar{x}}{\lambda^2}\right]$$

$$= \frac{\sum_i E[x_i]}{\lambda^2}$$

$$= \frac{n\cancel{\lambda}}{\lambda^{\cancel{2}}}$$

$$= \frac{n}{\lambda}$$

$$\left[J(\theta)\right]^{\frac{1}{2}} = \left[\frac{n}{\lambda}\right]^{\frac{1}{2}}$$

$$\propto \frac{1}{\sqrt{\lambda}}$$

$$P(\lambda|\mathbf{x}) \propto \left[\exp(-n\lambda) \lambda^{n\bar{x}}\right]\left[\lambda^{-\frac{1}{2}}\right]$$

$$= \exp(-n\lambda) \lambda^{n\bar{x} - \frac{1}{2} + 1 - 1}$$

$$= \exp(-n\lambda) \lambda^{n\bar{x} + \frac{1}{2} - 1}$$

$$\lambda|\mathbf{x} \sim Gamma\left(n\bar{x} + \frac{1}{2}, n\right)$$

```
set.seed(1)
gs <- rgamma(Nsim , sum( skin[,1] ) + .5 ,  nrow(skin) )
quantile(gs, prob  = c(0.5, 0.025, 0.975))
```

```
##      50%     2.5%    97.5%
## 2.892049 2.810743 2.973529
```

```
#mean
mean(gs)
```

```
## [1] 2.892471
```

9

Sampled Posterior Density