Lab 3

Senan Hogan-H.

8 March 2018

Completed with James Kinney, around 50/50 work completed between us.

Question 1.

a. Define $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$.

$$\sum_{i=1}^{n}(x_i - \mu)^2 =$$

$$\sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2) =$$

$$\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 =$$

$$\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 + n\bar{x}^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 =$$

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 + n\bar{x}^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 =$$

$$\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) + n(\bar{x}^2 - 2\mu\bar{x} + \mu^2) =$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 =$$

$$ns^2 + n(\bar{x} - \mu)^2$$

$$(1)$$

b. $\sigma_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$ is a convex function of the prior variance and the sample variance, σ_0^2 and $\frac{\sigma^2}{n}$ respectively.

Similarly $\mu_n = \sigma_n^2(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2})$ is a convex function of the prior and sample means, μ_0 and $\bar{x}$ respectively. These two values are scaled by their variances.

This is like the beta-binomial conjugate posterior where the posterior parameters α and β are convex functions (sums in that case) of the prior and data. In this case, the weight of the posterior variance coming from the sample is based on the sample variance and n with increased weight from larger n and smaller σ^2. The weight of the posterior mean is also based on n and the prior and sample variances with increased weight going to the distribution (prior or sample) with smaller variance.

c. Start with a normal prior, $f(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, where $\mu_0 = \sigma_0 = 1$. Suppose that the real distribution has a mean and standard deviation of 3, so that $\bar{x} = 3$.

```
x_bar <- sd <- 3
mu_0 <- 1
sd_0 <- 1
```

```r
prob <- seq(0, 6, .01)
# First prior
x_0 <- dnorm(prob, mu_0, sd_0)

# Update with data sample of 10
n <- 10

sd_10 <- ((n/(sd^2) + 1/(sd_0^2)))^(-0.5)
mu_10 <- sd_10^2*((mu_0/(sd_0^2)) + (n*x_bar)/sd^2)

x_10 <- dnorm(prob, mu_10, sd_10)

# Update with another data sample of 100
n <- 100

sd_100 <- ((n/(sd^2) + 1/(sd_0^2)))^(-0.5)
mu_100 <- sd_100^2*((mu_0/(sd_0^2)) + (n*x_bar)/sd^2)

x_100 <- dnorm(prob, mu_100, sd_100)

data.frame(x_0, x_10, x_100, prob) %>% ggplot(aes(x= prob)) +
  geom_line(aes(y= x_0, colour = 'Prior Distribution')) +
  geom_line(aes(y= x_10, colour = 'Posterior Distrubtion after n=10 sample')) +
  geom_line(aes(y= x_100, colour = 'Posterior Distrubtion after n=100 sample')) +
  labs(title = "Density for Normal, mean of 3, Updated with two Samples",
       x = "mu", y = "Density") + theme_classic()
```
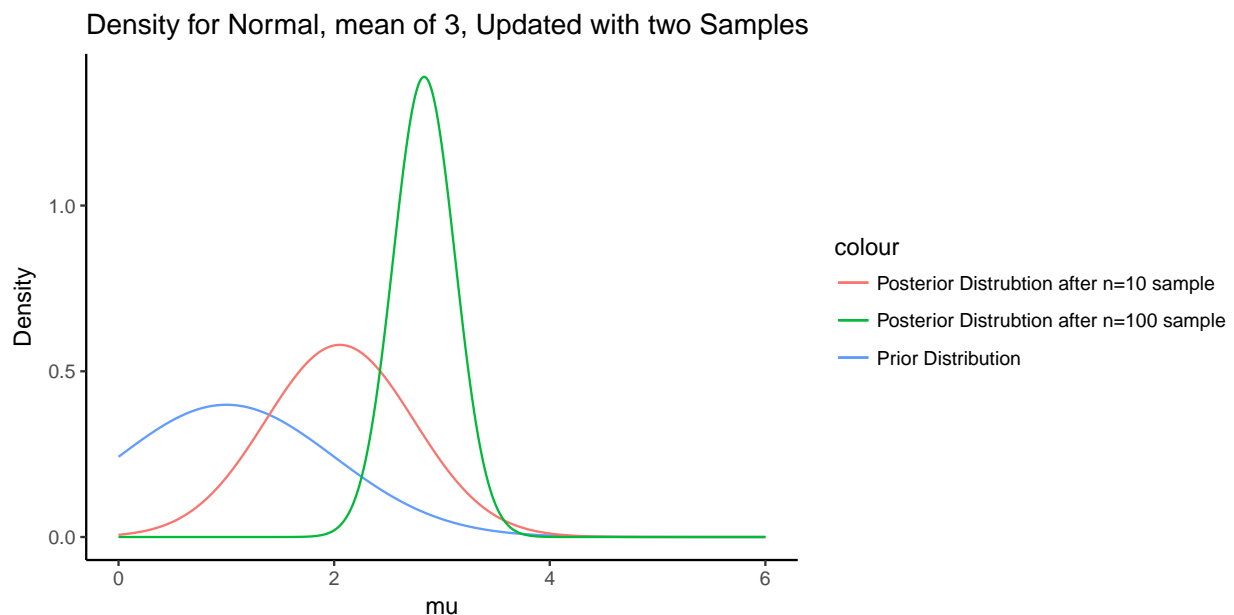

d. Use a prior of the form $\mathcal{G}(\alpha_0, \beta_0)$ where $\alpha_0, \beta_0 > 0$ to estimate λ.

```r
prob <- seq(0, 5, .01)

alpha_0 <- 1
beta_0 <- 1
```
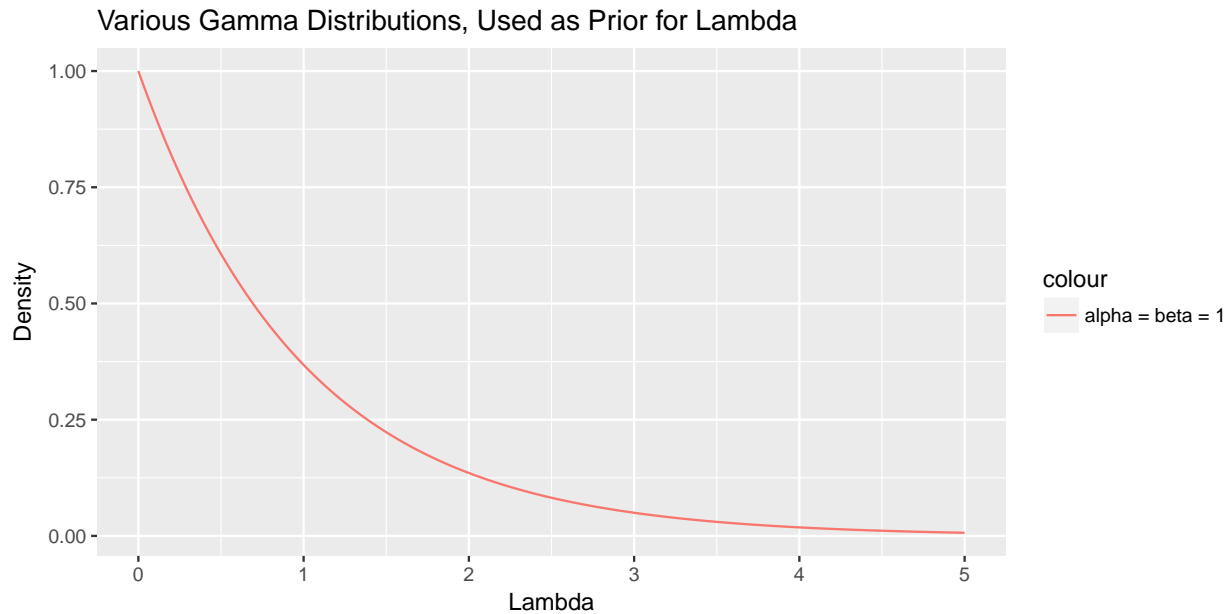
```
x_0 <- dgamma(prob, alpha_0, beta_0)

data.frame(x_0, prob) %>%
  ggplot(aes(x = prob)) +
  geom_line(aes(y = x_0, colour = 'alpha = beta = 1')) +
  labs(title = "Various Gamma Distributions, Used as Prior for Lambda",
       x = "Lambda", y = "Density")
```

Various Gamma Distributions, Used as Prior for Lambda

This would say that our best guess of the precision is 0 and that we believe the probabilty of the precision decreases as λ increases. It also says that we are close to positive that $\lambda < 5$.

```
alpha_0 <- 2
beta_0 <- 3
x_0 <- dgamma(prob, alpha_0, beta_0)

data.frame(x_0, prob) %>%
  ggplot(aes(x = prob)) +
  geom_line(aes(y = x_0, colour = 'alpha = 2 beta = 3')) +
  labs(title = "Various Gamma Distributions, Used as Prior for Lambda",
       x = "Lambda", y = "Density")
```

## Various Gamma Distributions, Used as Prior for Lambda
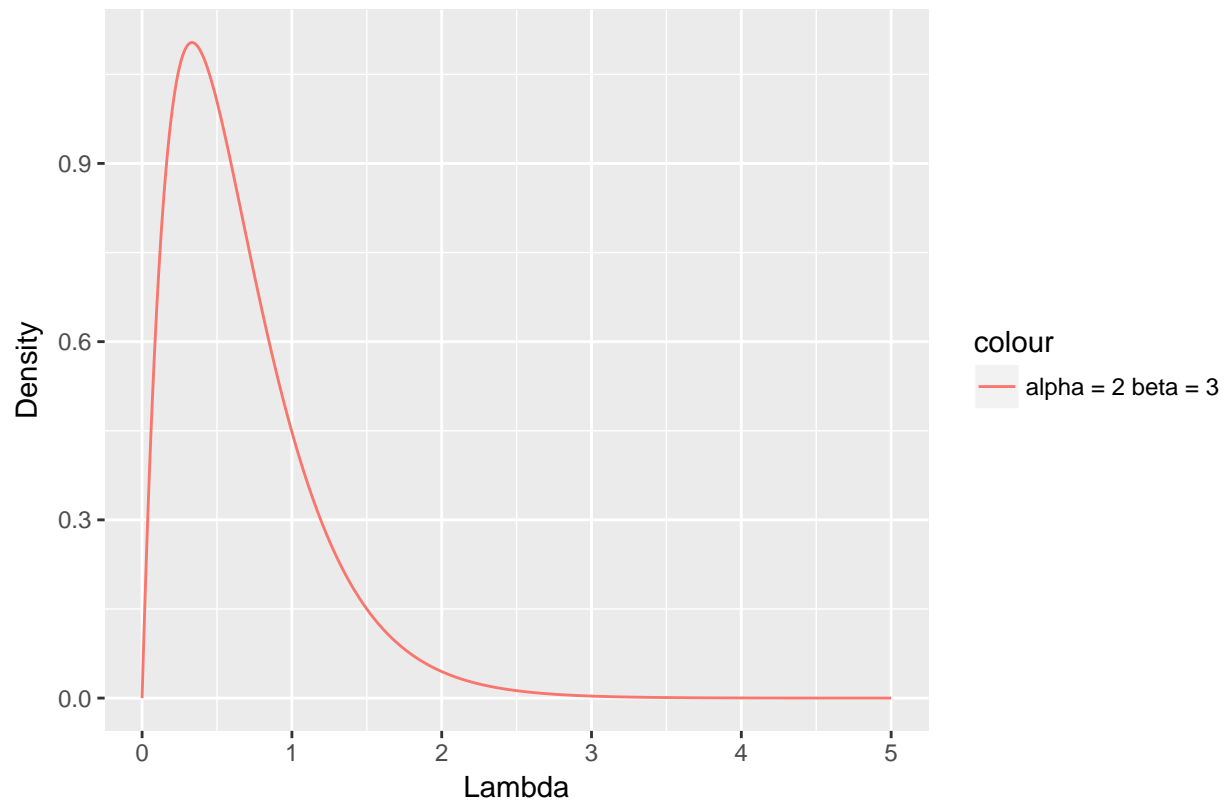
This says that our best guess of λ is close to $\frac{1}{3}$ and we believe it to be within $(0, 2)$ with high probability. We also believe that there is probability 0 of $\lambda = 0$, which seems like a good idea because we don't believe that the variance is infinite.

```
alpha_0 <- 50
beta_0 <- 25
x_0 <- dgamma(prob, alpha_0, beta_0)

data.frame(x_0, prob) %>%
  ggplot(aes(x = prob)) +
  geom_line(aes(y = x_0, colour = 'alpha = 50 beta = 25')) +
  labs(title = "Various Gamma Distributions, Used as Prior for Lambda",
       x = "Lambda", y = "Density")
```

Various Gamma Distributions, Used as Prior for Lambda
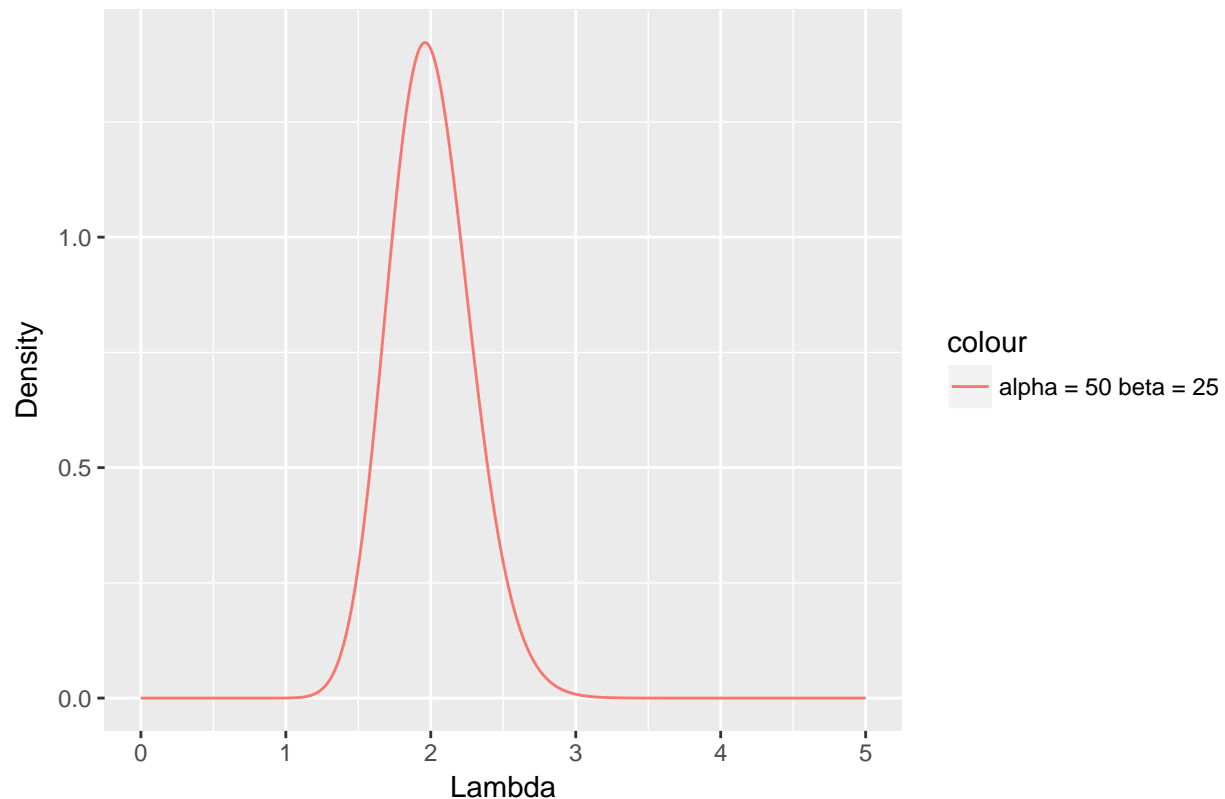
This prior would suggest that our best guess of λ is slightly less than 2, and we are almost postive that $\lambda \in [1,3]$.

    e. THIS QUESTION IS ACTUALLY NOT POSSIBLE WITH GIVEN PRIOR. CHANGE PROBABILITY TO 0.95.

For the prior, want $\sigma^2 = 4$, $P(\sigma^2 > 2) = 0.95$.

So that $\frac{\alpha_0 - 1}{\beta_0} = 4$, $pbeta(2, \alpha_0, \beta_0) = 1 - 0.95 = 0.05$

```r
x <- 2

vector <- seq(1, 3, by = 0.00001)
i <- 1

# (alpha_0 - 1)/beta_0 = 4, so beta_0 = (alpha_0 - 1)/4
alpha_0 <- vector[i]
beta_0 <- (alpha_0 - 1)/4

epsilon <- 0.0001
while (abs(pgamma(x, alpha_0, beta_0) - 0.05) > epsilon){
  #print(i)
  i <- i + 1
  alpha_0 <- vector[i]
  beta_0 <- (alpha_0 - 1)/4
}
rm(vector, i)
alpha_0
```

```
## [1] 1.174
beta_0
```

```
## [1] 0.04341
pgamma(x, alpha_0, beta_0) # roughly = 0.05
```

```
## [1] 0.0499
```

f.

```
Normal_estimator <- function(x.data, mu_0, kappa_0, alpha_0, beta_0){
  # returning posterior mean for mu.  First define the prior hyperparameters.
  n <- length(x.data)
  x_bar <- mean(x.data)
  s <- sd(x.data)

  # Then define some posterior hyperparameters.
  mu_n <- (kappa_0*mu_0 + n*x_bar)/(kappa_0 + n)
  kappa_n <- kappa_0 + n
  alpha_n <- alpha_0 + n/2
  beta_n <- beta_0 + (n/2)*s^2 + (kappa_0*n*(x_bar - mu_0)^2)/(2*(kappa_0 + n))

  # Sample from posterior for lambda_star
  lambda.data <- c()
  for (i in c(1:10000)){
    lambda.data <- c(lambda.data,
                 rgamma(1, alpha_n, beta_n))
  }
  lambda_star <- mean(lambda.data)

  # sample fron posterior for mu_star
  mu.data <- c()
  for (i in c(1:10000)){
    mu.data <- c(mu.data,
                 rnorm(1, mu_n, (kappa_n*lambda.data[i])))
  }

  # posterior mean
  mu_star <- mean(mu.data)

  # 95% CI for mu
  mu_CI <- quantile(mu.data, probs = c(0.025, 0.975))

  # 95% CI for new observation
  N <- 100
  new.data <- rnorm(N, mu_star, ((kappa_0 + N)*lambda_star ))
  new_CI <- quantile(new.data, probs = c(0.025, 0.975))
  return(list(mu_star, mu_CI, new_CI))
}

data <- rnorm(10, 12, 1)
Normal_estimator(data , 12, 1, 6, 4)
```

```
## [[1]]
## [1] 12.01
```

```
## 
## [[2]]
##    2.5%  97.5%
## -17.33  41.12
## 
## [[3]]
##    2.5%  97.5%
## -223.6  194.6
```

g. Cars had lower fuel efficiency in 1974. I am guessing the average fuel efficiency is 12 miles per gallon. I feel like this is a pretty good guess, but I am not so confident that I would rule out 10 or 14 miles per gallon. Thus, I want about 95% within 2 of the mean. This translates to $1.96\sigma = 2 \implies \sigma \approx 1 \implies \sigma^2 = 1$. Again, I'm fairly confident of that variance. Thus I want $\frac{\beta}{\alpha - 1} = 1$. Let $\mu_0 = 20$, $\kappa_0 = 1$, $\alpha_0 = 30$, and $\beta_0 = 30$.

First lets plot the $\mathcal{G}(\alpha, \beta)$.

```
prob <- seq(0, 7, .01)
alpha_0 <- 30
beta_0 <- 30
x_0 <- dgamma(prob, alpha_0, beta_0)

data.frame(x_0, prob) %>%
  ggplot(aes(x = prob)) +
  geom_line(aes(y = x_0, colour = 'alpha = 30, beta = 30')) +
  labs(title = "Various Gamma Distributions, Used as Prior for Lambda",
       x = "Lambda", y = "Density")
```
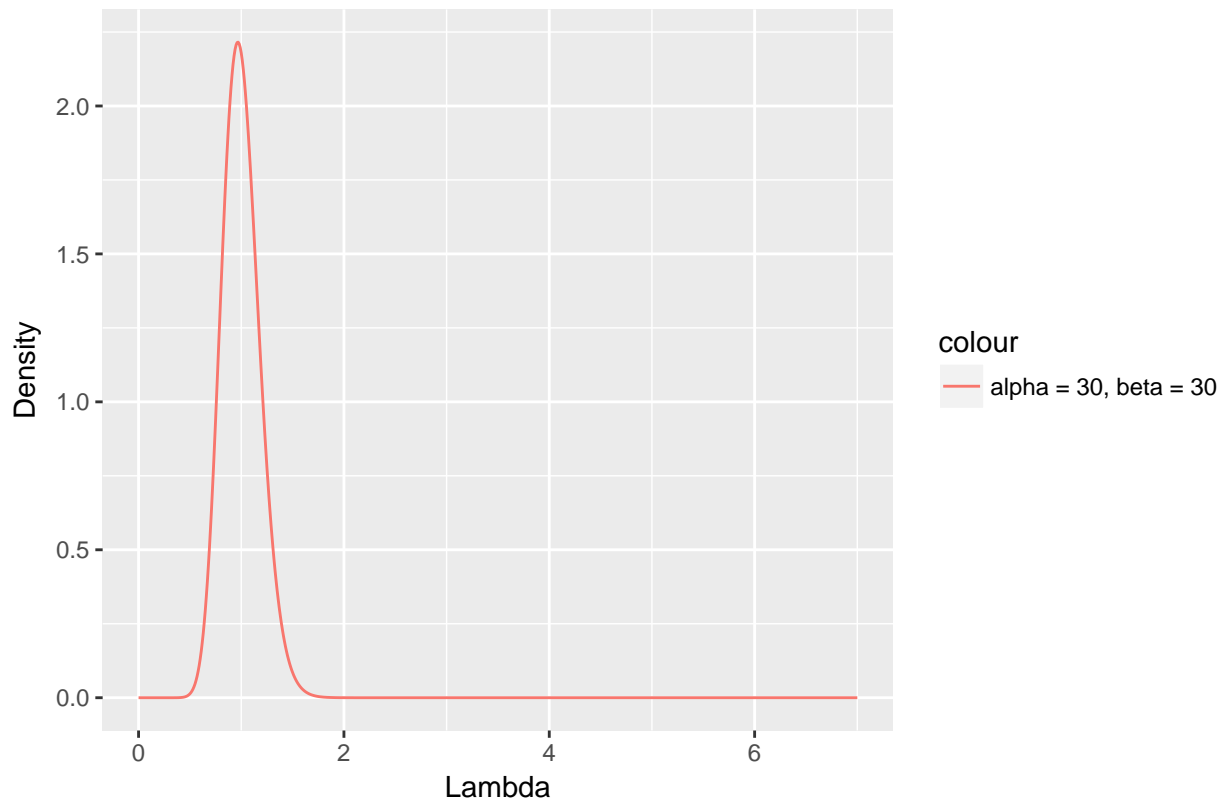
This says that we are confident that the precision is less than $\frac{3}{2}$ and greater than $\frac{1}{2}$. This corresponds to $\sigma^2 \in [\frac{2}{3}, 2]$

```r
data(mtcars)
mpg = mtcars[,1]
missingObs = sample(seq(1,32,1), 1)
mpg_minus = mpg[-missingObs]
missing = mpg[missingObs]
estimateMPG = Normal_estimator(mpg_minus, 20, 1, 30, 30)
estimateMPG[3]
```

```
## [[1]]
##   2.5%  97.5%
##  5.282 34.998
```

```r
missing
```

```
## [1] 10.4
```

This contains the missing point.

Consulted the following resource in working on this exercise: https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf