

STAT3120 Applied Bayesian Methods

Semester 2, 2016

Benjamin G. Moran; c3076448@uon.edu.au

20th September 2016

Due Friday 16th September 2016 (on-campus by 3pm, distance by 5pm)

Late assignments will not be accepted without prior written permission of the lecturer.

Total marks: 45

Marking Criteria:

1. Answers must be written in clear English with all appropriate working and/or supporting computer output shown.
 2. Raw computer output without explanatory text is unacceptable.
 3. Students are required to understand the content in Weeks 2-7 to answer these questions.
 4. As part of your workings you should include applicable R-code.
-

Question 1 Total: 12 marks

- (a) Consider a sample of $n = 15$ data points, y , from the Normal distribution with unknown mean μ and variance σ^2 . For this sample we wish to allow the 14th and 15th data points, (y_{14}, y_{15}) , to have an inflated variance, namely $10\sigma^2$.

- (i) Write down the likelihood $p(y \mid \mu, \sigma^2)$, including all constants. [2 marks]

Answer: The likelihood is:

$$\begin{aligned} p(y \mid \mu, \sigma^2) &= \prod_{i=1}^{13} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \times \prod_{i=14}^{15} (20\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{20\sigma^2}(y_i - \mu)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{13}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{13} (y_i - \mu)^2\right\} \times (20\pi\sigma^2)^{-\frac{2}{2}} \exp\left\{-\frac{1}{20\sigma^2} \sum_{i=14}^{15} (y_i - \mu)^2\right\} \\ &= 20^{-1} (2)^{-\frac{13}{2}} (\pi\sigma^2)^{-\frac{15}{2}} \exp\left\{-\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 - \sum_{i=14}^{15} (y_i - \mu)^2}{20\sigma^2}\right\} \\ &\propto (\sigma^2)^{-\frac{15}{2}} \exp\left\{-\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 - \sum_{i=14}^{15} (y_i - \mu)^2}{20\sigma^2}\right\} \end{aligned}$$

To perform inference on $p(\mu, \sigma^2 | y)$ we partition using $p(\mu | y)$ and $p(\sigma^2 | y, \mu)$.

- (ii) Mathematically derive the density $p(\sigma^2 | y, \mu)$ assuming the non-informative prior $p(\sigma^2) \propto 1/\sigma^2$. Derive $E(\sigma^2 | y, \mu)$ and $V(\sigma^2 | y, \mu)$. [4 marks]

Answer: We can express the joint posterior density as:

$$\begin{aligned} p(\sigma^2 | y, \mu) &\propto p(y | \mu, \sigma^2) p(\sigma^2) \\ &\propto (\sigma^2)^{-\frac{15}{2}} \exp \left\{ -\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 - \sum_{i=14}^{15} (y_i - \mu)^2}{20\sigma^2} \right\} \times \frac{1}{\sigma^2} \\ &\propto (\sigma^2)^{(-\frac{15}{2}+1)} \exp \left\{ -\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 - \sum_{i=14}^{15} (y_i - \mu)^2}{20\sigma^2} \right\} \end{aligned}$$

Therefore, the density of $p(\sigma^2 | y, \mu)$ has the form of an Inverse Gamma, such that:

$$\sigma^2 | y, \mu \sim I.G. \left(\alpha = \frac{15}{2}, \beta = \frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 + \sum_{i=14}^{15} (y_i - \mu)^2}{20} \right)$$

We know that the mean of an Inverse Gamma is given by $\frac{\beta}{\alpha-1}$ and the variance by $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$. Therefore

$$\begin{aligned} E(\sigma^2 | y, \mu) &= \frac{\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 + \sum_{i=14}^{15} (y_i - \mu)^2}{20}}{\frac{15}{2} - 1} \\ &= \frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 + \sum_{i=14}^{15} (y_i - \mu)^2}{130} \\ V(\sigma^2 | y, \mu) &= \frac{\left(\frac{10 \sum_{i=1}^{13} (y_i - \mu)^2 + \sum_{i=14}^{15} (y_i - \mu)^2}{20} \right)^2}{\left(\frac{15}{2} - 1 \right)^2 \left(\frac{15}{2} - 2 \right)} \\ &= \frac{3200 \left(10 \sum_{i=1}^{13} (y_i - \mu)^2 + \sum_{i=14}^{15} (y_i - \mu)^2 \right)}{1859} \end{aligned}$$

- (b) A biologist believes that lifetimes in days of fruitflies follow an exponential distribution. The density for a single observation y drawn from an exponential distribution is

$$p(y; \lambda) = \lambda e^{-\lambda y}, \quad y > 0.$$

For data, the biologist observes 100 randomly selected fruitflies and records how many days they live.

- (i) Write the likelihood of λ if there are 100 data values, represented symbolically as y_1, \dots, y_{100} . What parametric family is the conjugate prior for the exponential likelihood? (Just name the family.) [2 marks]

Answer: For the likelihood function of an exponential distribution is:

$$\begin{aligned} L(\lambda; y_n) &= \prod_{i=1}^n \lambda e^{-\lambda y_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n y_i} \end{aligned}$$

with $n = 100$ we get:

$$\begin{aligned} L(\lambda; y_{100}) &= \lambda^{100} e^{-\lambda \sum_{i=1}^{100} y_i} \\ &= \lambda^{100} e^{-100\lambda \bar{y}} \end{aligned}$$

The conjugate prior for the exponential likelihood is Gamma distribution.

- (ii) The biologist states that a prior distribution with mean 3 days and standard deviation 2 would reflect their previous knowledge about fruitfly lifetimes. What parameters should they specify in their prior? (numeric answer; show your work) [2 marks]

Answer: We know that the mean and variance of $Gamma(a, b)$ are given by

$$Mean = \frac{a}{b}; \quad Variance = \frac{a}{b^2}$$

So we simply need to solve the simultaneous equations

$$\begin{aligned} \frac{a}{b} &= 3 \implies a = 3b \\ \frac{a}{b^2} &= 2 \end{aligned}$$

If we sub the value for a into the second equation, we get:

$$\frac{3b}{b^2} = 2 \implies b = \frac{3}{2} \implies a = \frac{9}{2}$$

which leaves us with a prior of $Gamma(\frac{9}{2}, \frac{3}{2})$.

- (iii) Based on the likelihood and the prior that you have specified, find the posterior density $p(\lambda | y_1, \dots, y_{100})$. Name it and give the numeric values of the parameters. [2 marks]

Answer: Suppose that we have data $X_i \sim Exp(\lambda)$, and that the prior distribution for λ is $Gamma(\alpha, \beta)$. Then the posterior distribution of λ is given as

$$\begin{aligned} p(\lambda | \vec{x}, \alpha, \beta) &\propto \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) \lambda^{\alpha-1} \exp(-\beta\lambda) \\ &= \lambda^{n+\alpha-1} \exp\left(-\lambda \left(\beta + \sum_{i=1}^n x_i\right)\right) \\ &= \lambda^{n+\alpha-1} \exp(-\lambda(\beta + n\bar{x})) \end{aligned}$$

which is equivalent to a $Gamma(\alpha + n, \beta + n\bar{x})$. Subbing in the values we determined above gives us:

$$\begin{aligned} p(\lambda | \vec{y}, 9/2, 3/2) &\propto \lambda^{100} \exp(-100\lambda \bar{y}) \lambda^{7/2} \exp(-3/2\lambda) \\ &= \lambda^{100+7/2} \exp(-\lambda(3/2 + 100\bar{x})) \\ &= \lambda^{207/2} \exp(-\lambda(3/2 + 100\bar{x})) \end{aligned}$$

which is equivalent to $Gamma(207/2, 3/2 + 100\bar{x})$.

Question 2 Total: 13 marks

The table below is adapted from the text BDA. It consists of the numbers of fatal accidents and approximate numbers of passenger miles (in 100 millions) flown each year from 1976 to 1985.

	Fatal Accidents	Passenger Miles
1976	24	3863
1977	25	4300
1978	31	5027
1979	31	5481
1980	22	5814
1981	21	6033
1982	26	5877
1983	20	6223
1984	16	7433
1985	22	7107

- (a) Assume the numbers of fatal accidents are independently and identically distributed. Suggest an appropriate distribution to model the number of fatal accidents. [2 marks]

Answer: An appropriate model to use would be a Poisson model. We would have a parameter θ representing the *expected* number of accidents occurring in a year and data y_i representing the *actual* number of accidents each year. So, $y_i \mid \theta \sim \text{Pois}(\theta)$. An appropriate distribution that we can use to model this would be a Gamma distribution:

$$y_i \mid \theta \sim \text{Gamma}(\alpha, \beta) = \text{Gamma}(n\bar{y}_i, n) = \text{Gamma}(238, 10)$$

- (b) Set a suitable prior distribution for the mean parameter in (a) and determine its posterior distribution based on the data in the table above. Plot this posterior density. [3 marks]

Answer: From the data we can see that there have been 238 fatal accidents in $n = 10$ years. If we assume a non-informative $\text{Gamma}(1, 0)$ prior, then the posterior distribution for the mean parameter is

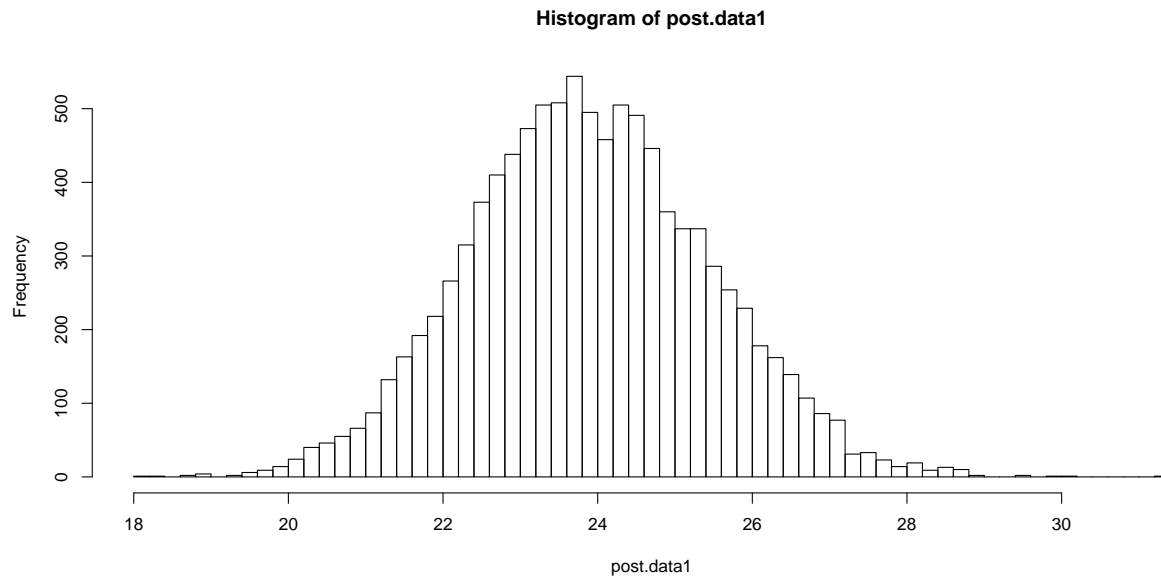
$$\text{Gamma}_{\text{post}}(\alpha, \beta) = \text{Gamma}(1 + n\bar{y}_i, 0 + n) = \text{Gamma}(239, 10)$$

```
fatal.accidents <- c(24,25,31,31,22,21,26,20,16,22)
miles.flown <- c(3863,4300,5027,5481,5814,6033,5877,6223,7433,7107)
mu.y <- mean(fatal.accidents)

a0 <- 1; b0 <- 0; n <- 10

post.data1 <- rgamma(10000,a0 + n*mu.y, b0 + n)

hist(post.data1,50)
```



(c) Find a point estimate and 95% interval for the average yearly number of fatal accidents. [2 marks]

Answer: We do this by going back to the Poisson model.

```
theta <- rgamma(1000,n*mu.y)/(n)
theta.hat <- rpois(1000,theta)
quantile(theta.hat,c(0.025,0.975))
## 2.5% 97.5%
## 14 35
mean(theta.hat)
## [1] 24.047
```

Here we can see an estimate of 24.047 for the mean, with a 95% CI of 14, 35.

(d) Part (a) has ignored the information about passenger miles. Perhaps a better model here would take account that each fatal accident count is proportional to the number of miles flown. Set up a model of accident counts that accounts for the number of miles flown. [2 marks]

Answer: Again, we have a Poisson model with an appropriate Gamma distribution. By accounting for the number of passenger miles flown (x_i), our model becomes $y_i | x_i, \theta \sim \text{Pois}(x_i \theta)$, which leads to a conjugate prior Gamma of:

$$y_i | \theta \sim \text{Gamma}(\alpha, \beta) = \text{Gamma}(n\bar{y}_i, n\bar{x}_i) = \text{Gamma}(238, 5715.8 * 10^{12})$$

(e) Using this new model, set a prior distribution for the mean parameter and determine its posterior distribution. Plot this density and compare it with that in (b). [4 marks]

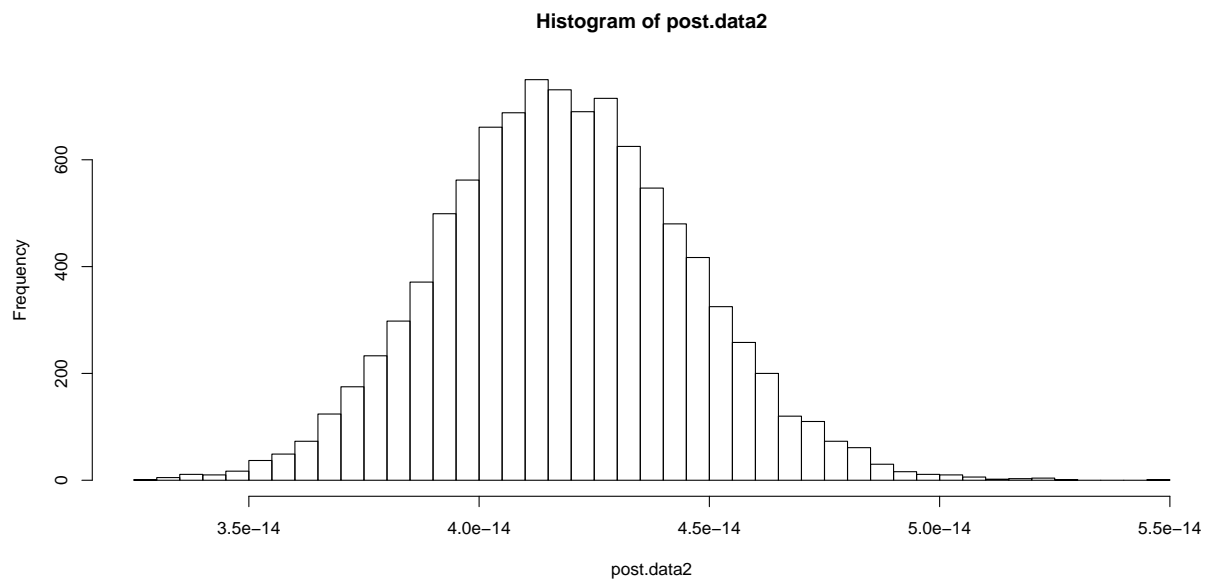
Answer: Again, assuming a non-informative conjugate prior of $\text{Gamma}(1, 0)$, our posterior becomes:

$$y_i | \theta \sim \text{Gamma}_{\text{post}}(\alpha, \beta) = \text{Gamma}(1 + n\bar{y}_i, 0 + n\bar{x}_i) = \text{Gamma}(239, 5715.8 * 10^{12})$$

```

mu.miles <- mean(miles.flown*10^(11))
mu.miles
## [1] 5.7158e+14
a0 <- 1; b0 <- 0; n <- 10
post.data2 <- rgamma(10000,a0 + n*mu.y, b0 + n*mu.miles)
hist(post.data2,50)

```



```

qgamma(0.025,a0 + n*mu.y, b0 + n*mu.miles)
## [1] 3.668037e-14
qgamma(0.975,a0 + n*mu.y, b0 + n*mu.miles)
## [1] 4.727878e-14

theta <- rgamma(1000,n*mu.y)/(mu.miles*n)
theta.hat <- rpois(1000,theta*mu.miles)
quantile(theta.hat,c(0.025,0.975))
## 2.5% 97.5%
## 14 34
mean(theta.hat)
## [1] 23.691

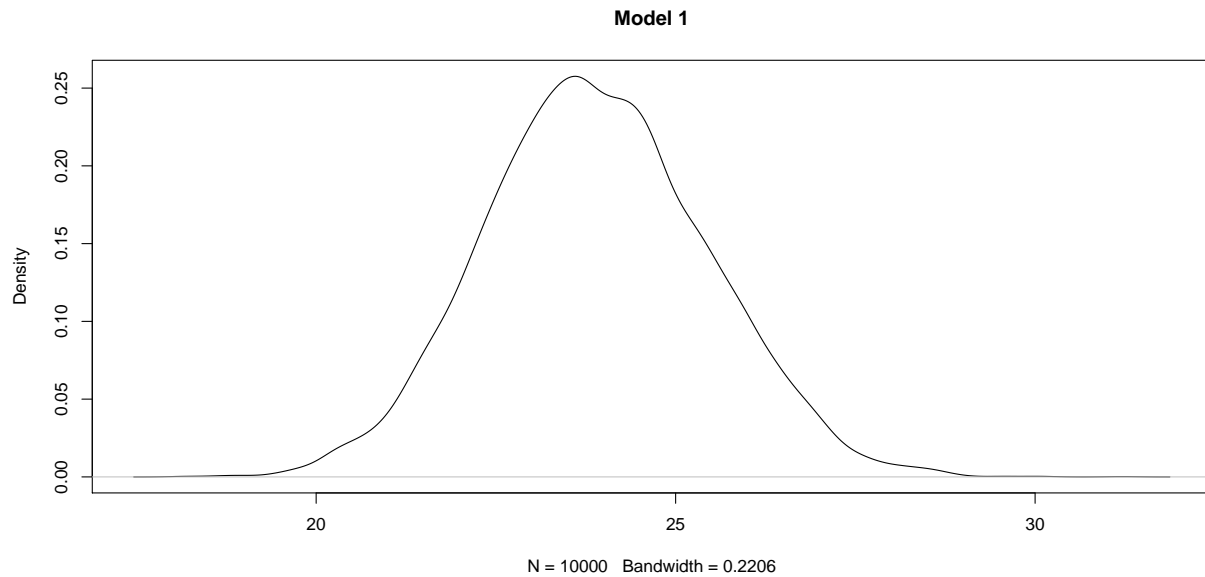
```

Here we can see an estimate of 23.691 for the mean, with a 95% CI of 14, 34. Note that these are virtually identical to the estimates arrived at by our first model. The similarities between the two outputs are made clearer when we plot their densities.

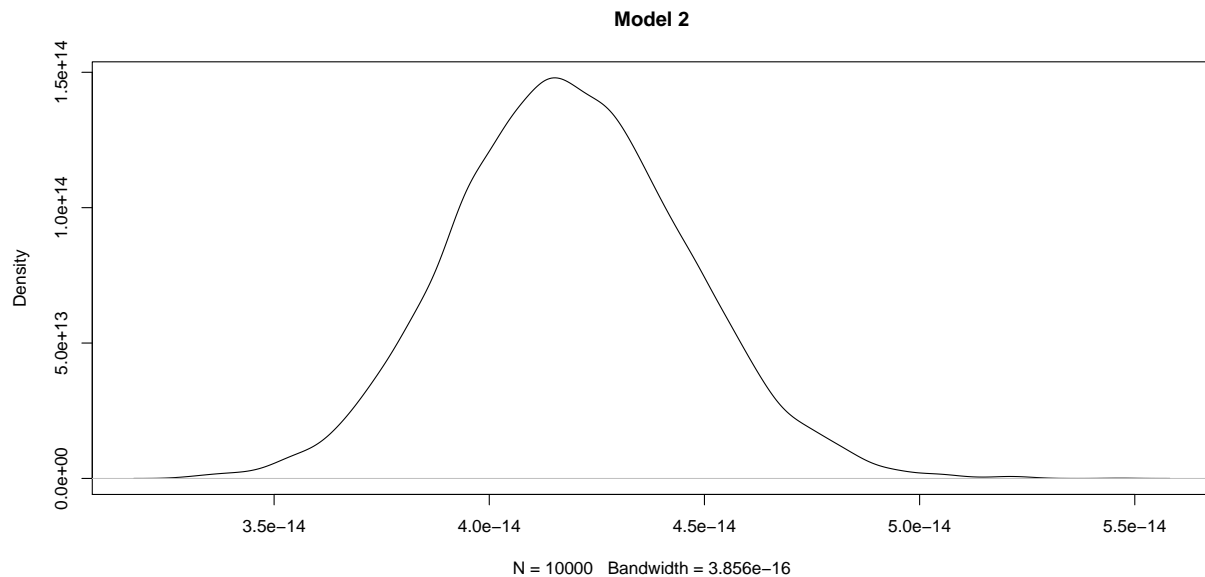
```

plot(density(post.data1), main = "Model 1")

```



```
plot(density(post.data2), main = "Model 2")
```



Both seem to be very similar. Along with the similarities in the estimates, we should conclude that the new model has not improved our ability to estimate the true parameters of the data.

Question 3 [Total: 20 marks]

Consider the example on corn yields but with an extended data set, with sample sizes, sample means and sample variances as given below. We have five different corn growers giving individual yields for a particular

new type of genetically engineered corn that has three seasons per year. The corn is distributed by a research station to the growers. Growers give yields for the most recent seasons in tons/hectare. Summary statistics for the data are given below.

Table 2: Corn Yield Data by Grower

	1	2	3	4	5
n	16.0	19.0	14.0	12.0	8.0
y	15.3	16.2	16.4	13.2	13.5
s ²	8.2	12.3	7.9	5.2	6.2

- (a) By adapting code from the week 7 lecture, fit a hierarchical Normal model to this data, using the Empirical Bayes approach. Provide estimates and inference for each unknown true grower mean yield, including a plot showing all estimated posterior densities, one for each unknown mean. [7 marks]

Answer: The lecture material presents us with 3 possible models, however only one is represented by the code in those same notes: the model that makes no assumptions about individual grower variances (3). So, the model we must fit to our data is a non-standard posterior distribution that is constructed by multiplying a Normal prior with the likelihood for μ_j in the form of a t-density.

$$\begin{aligned}
 p(\mu_j | y_j, \mu, \tau^2) &\propto p(y_j | \mu_j, \mu, \tau^2) p(\mu_j | \mu, \tau^2) \\
 &= p(\mu_j | \mu, \tau^2) \int p(\mu_j | y_j, \sigma_j^2, \mu, \tau^2) (\sigma_j^2 | \mu, \tau^2) \\
 &\propto p(N(\mu, \tau^2)) p(t_{n_j-1}(\bar{y}_j, s_j^2)/n_j)
 \end{aligned}$$

For the choice of hyperparameters μ and τ , we again proceed in line with the notes and calculate the average yield across all growers, along with the standard deviation of that average yield. This gives us our empirical priors, which we can use to perform the estimation.

```

x <- seq(5, 25, by=0.001)
n <- c(16, 19, 14, 12, 8)
y.bar <- c(15.3, 16.2, 16.4, 13.2, 13.5)
sig2 <- c(8.2, 12.3, 7.9, 5.2, 6.2)
mu <- mean(y.bar) # 14.92
tau <- sqrt(var(y.bar)) # 1.49566...
par(mfrow = c(1,2))
pdf.data <- matrix(nrow=5,ncol=20001)
for(k in 1:5){
  px <- -(x-mu)^2/(2*tau^2)-3/2*log(1+(3*(x-y.bar[k])^2)/(2*sig2[k]))
  pdf.data[k,] <- px
  if(k==1){
    plot(x,exp(px),type='l',col=k, main = "PDFs")
  } else {
    lines(x,exp(px),type='l', col = k)
  }
}
cdf.data <- matrix(nrow=5,ncol=20001)
for(k in 1:5) {
  px <- -(x-mu)^2/(2*tau^2) -3/2*log(1+(3*(x-y.bar[k])^2)/(2*sig2[k]))
  cdf.data[k,] <- exp(px)/sum(exp(px))
}

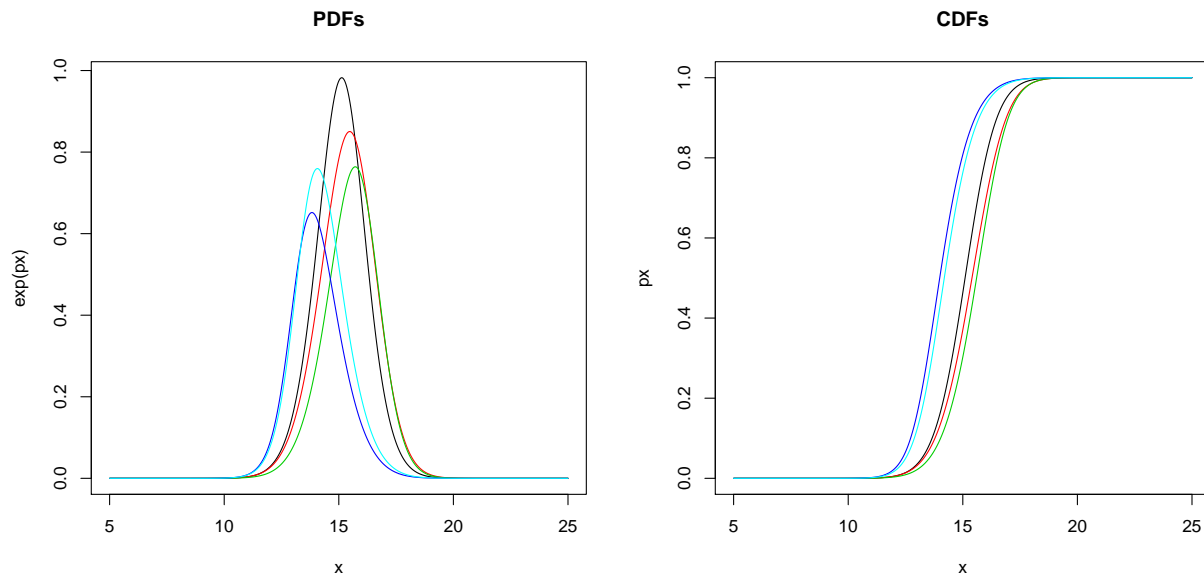
```



```

for(k in 1:5) {
  px <- cumsum(cdf.data[k,])
  if(k==1) {
    plot(x,px,type='l', col=k, main = "CDFs")
  } else {
    lines(x,px,type='l', col = k)
  }
}

```



We can generate MCMC samples using the following code (Note: I have not included the code for each grower in order to save space):

```

cdf1 <- cumsum(cdf.data[1,])
u=runif(1000,0,1)
mu1=0

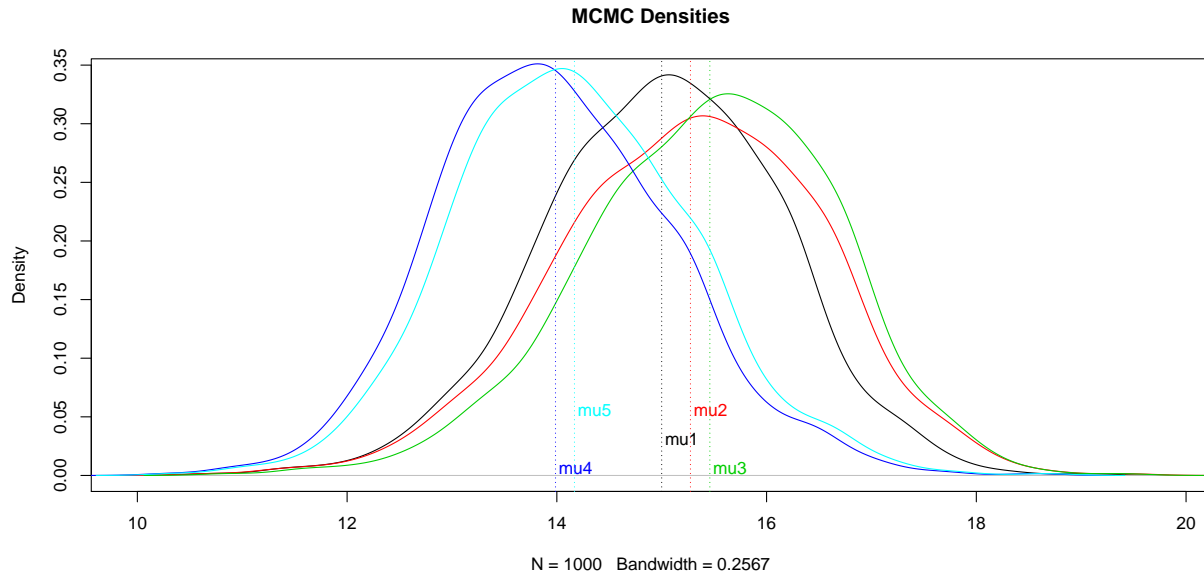
for(j in 1:1000){
  st=0;k=1
  while(st==0) {
    if(cdf1[k]<u[j]&cdf1[k+1]>u[j]) {
      mu1[j] <- x[k]
      st <- 1
    } else {
      k <- k+1
    }
  }
}

```

Now, using our MCMC samples, we can calculate the posterior mean and CrI for each grower:

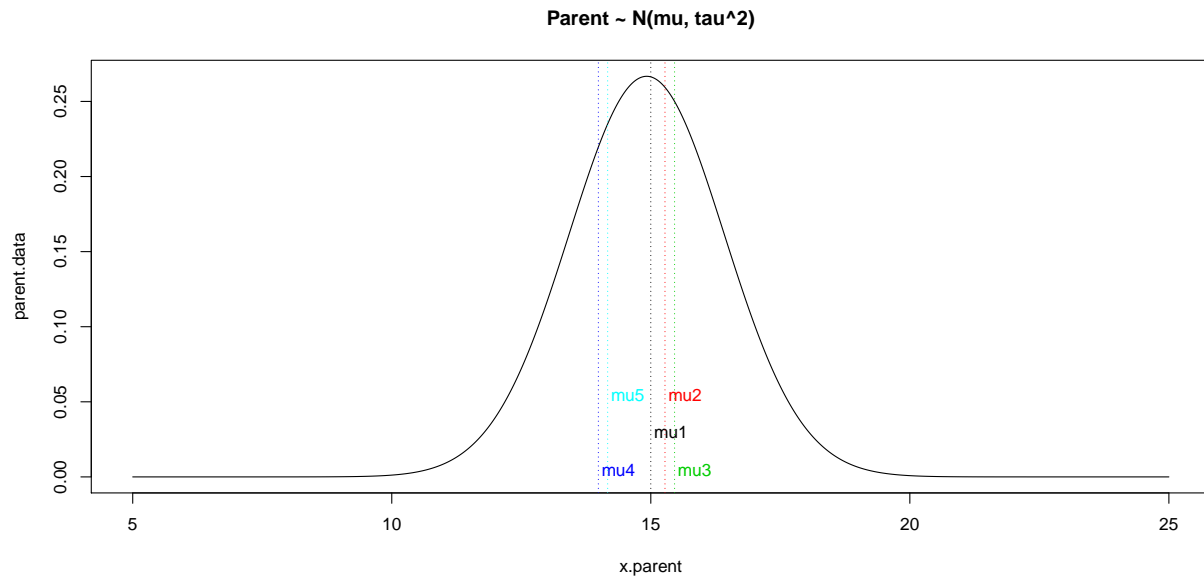
Grower	Mean	2.5%	97.5%	Range
1	14.99997	12.75093	17.21930	4.468375
2	15.27382	12.79392	17.63230	4.838375
3	15.45917	13.03092	17.66330	4.632375
4	13.98731	11.99695	16.41137	4.414425
5	14.16786	12.12192	16.55738	4.435450

This suggests an order (from the grower with the largest mean output to the lowest) of 3, 2, 1, 5, 4. Plotting the MCMC sample densities reaffirms this idea:



- (b) Plot the estimated parent density for the five grower means. At what percentile of this distribution does each grower's estimated posterior mean yield lie? Do they all fit in with the prior distribution? [2 marks]

```
x.parent <- seq(5,25,by=0.001)
parent.data <- dnorm(x.parent,mu,tau)
plot(x.parent,parent.data, type = 'l', main = "Parent ~ N(mu, tau^2)")
text(mean(mu1),0.025, "mu1", col = 1, adj = c(-0.1, -0.1))
abline(v=mean(mu1), lty=3, col = 1)
text(mean(mu2),0.05, "mu2", col = 2, adj = c(-0.1, -0.1))
abline(v=mean(mu2), lty=3, col = 2)
text(mean(mu3),0, "mu3", col = 3, adj = c(-0.1, -0.1))
abline(v=mean(mu3), lty=3, col = 3)
text(mean(mu4),0, "mu4", col = 4, adj = c(-0.1, -0.1))
abline(v=mean(mu4), lty=3, col = 4)
text(mean(mu5),0.05, "mu5", col = 5, adj = c(-0.1, -0.1))
abline(v=mean(mu5), lty=3, col = 5)
```



We can easily calculate the corresponding percentile of the parent distribution for each posterior mean by:

```
# Grower k
round(mean(x.parent<mean(muk))*100,2) # etc
```

This tells us that the posterior means for Growers 1 through 5 are:

Grower	1	2	3	4	5
Percentile	50	51.37	52.3	44.94	45.84

- (c) What is the posterior probability that each of the five growers has the largest mean yield? What is your conclusion in this case? [2 marks]

We can calculate these probabilities fairly simply, using the MCMC estimates for the posterior means of each grower and the following code:

```
# Grower k
length(muk[muk>mua&muk>mub&muk>muc&muk>mud])/n
# where k in {1:5} and a,b,c,d in {1:5}/{k}
```

Grower	1	2	3	4	5
Likelihood	0	0.005	0.994	0	0

This tells us that, using our estimates, Grower 3 has the largest mean yield with a probability of $\approx 99.4\%$.

- (d) What is the posterior probability that, for each pair of growers (grower i and i+k), the mean yield for grower(i) exceeds the mean yield for grower (i+k)? [2 marks]

Answer: Firstly, we can construct a matrix with all possible combinations of the probability statement outlined in the question. The following is example code (the full version is not reproduced to save space):

```
probs <- matrix(nrow = 5, ncol = 5)
n <- 1000
probs[1,1] <- length(mu1[mu1>mu1])/n # 0
probs[1,2] <- length(mu1[mu1>mu2])/n # 0
probs[1,3] <- length(mu1[mu1>mu3])/n # 0
probs[1,4] <- length(mu1[mu1>mu4])/n # 1
probs[1,5] <- length(mu1[mu1>mu5])/n # 1
# etc
```

Next, we can sort this data and display the results:

```
# Probability that mean of Grower A (Row) is greater than mean of Grower B (Column)
colnames(probs) <- c("1", "2", "3", "4", "5")
rownames(probs) <- c("1", "2", "3", "4", "5")
probs.m <- melt(probs)

probs.m <- as.data.frame(probs.m)
names(probs.m) <- c("GrowerA", "GrowerB", "Pr")
probs.m <- probs.m %>%
  arrange(GrowerA, desc(Pr))

kable(probs.m, row.names = FALSE, col.names = c("Grower A", "Grower B", "Pr(A>B)"))
```

Grower A	Grower B	Pr(A>B)
1	4	1.000
1	5	1.000
1	2	0.009
1	1	0.000
1	3	0.000
2	4	1.000
2	5	1.000
2	1	0.991
2	3	0.005
2	2	0.000
3	1	1.000
3	4	1.000
3	5	1.000
3	2	0.994
3	3	0.000
4	1	0.000
4	2	0.000
4	3	0.000
4	4	0.000
4	5	0.000
5	4	1.000
5	1	0.000
5	2	0.000
5	3	0.000
5	5	0.000

- (e) What is the most likely ordering for the 5 mean grower yields in your MC sample? What are the 2nd, 3rd and 4th most likely orderings? [4 marks]

Answer: We can evaluate likely orderings by calculating all permutations of $mean(\mu_a > \mu_b \& \mu_b > \mu_c \& \mu_c > \mu_d \& \mu_d > \mu_e), \forall a, b, c, d, e \in 1 : 5$. Doing so leaves us with:

```
# 1st: 3>2>1>5>4
mean(mu3>mu2&mu2>mu1&mu1>mu5&mu5>mu4)
## [1] 0.985
# 2nd: 2>3>1>5>4
mean(mu2>mu3&mu3>mu1&mu1>mu5&mu5>mu4)
## [1] 0.005
# 3rd: 3>1>2>5>4
mean(mu3>mu1&mu1>mu2&mu2>mu5&mu5>mu4)
## [1] 0.009
```

At this point we can see that the values add up to 1. Whilst other combinations are not impossible, the precision at which r has calculated these results has limited the number of possible combinations.

- (f) Based on (a)-(d), and any other evidence you can provide, where do you believe significant differences exist among the 5 grower means (if any)? [3 marks]

Answer: From the graph of MCMC densities, it is possible that we can discern two separate clusters of growers: Growers 4 and 5 in one; Growers 1,2, and 3 in the other. Also, all of the most likely permutations of posterior grower means only differ in the order that they list Growers 1, 2 and 3; Growers 4 and 5 are unchanged. This suggests to me that at the very least there is a significant difference between both groups. Secondly, within each group, the estimates produced by the MCMC sampling suggest very strongly that significant differences exist between each grower in the order determined to be most likely. For $3 > 2 > 1 > 5 > 4$, the probability assigned to each individual inequality (e.g. $3 > 2$) is $\geq 99\%$. From this we should conclude that significant differences exist between - and among - the 5 growers.

End of Assignment 2