Lisa Over
Final Exam Problem 3
May 5, 2015

CODE

```
#Select file and read data
data = read.csv(file.choose(), header=TRUE)
attach(data)

#Create x vector from “data”
x = X4

#Set Gibbs parameters
p = 0.5
b = 1
m = mean(x)
Y = m + b
N = 25000
lag = 1
burnin = 0
```

#Gibbs function receives six parameters: "x" for the sample data values, "p" for the probability that data value, xi, comes from distribution 1, "Y" for the mean of the distribution with mean (b+m) with b=1, "N" for number of realizations, "lag" for determining how many realizations to skip between saves, and "burnin" for determining how many realizations to skip before starting to save. Gibbs generates N independent z vectors of indicator variables, N independent means from a Poisson distribution, and N independent probabilities that are each a probability of drawing from distribution 1.

```
gibbs <- function(x,p,Y,N,lag,burnin) {

	#obtain length of x
	n = length(x)

	#Set N to be N*lag+burnin
	N <- N*lag + burnin

	#Initialize vectors to hold the z vector realizations, m2 realizations, and p
realizations
	zs = NULL
	Ys = NULL
	ps = NULL

	for(i in 1:N) {
```

Lisa Over
Final Exam Problem 3
May 5, 2015

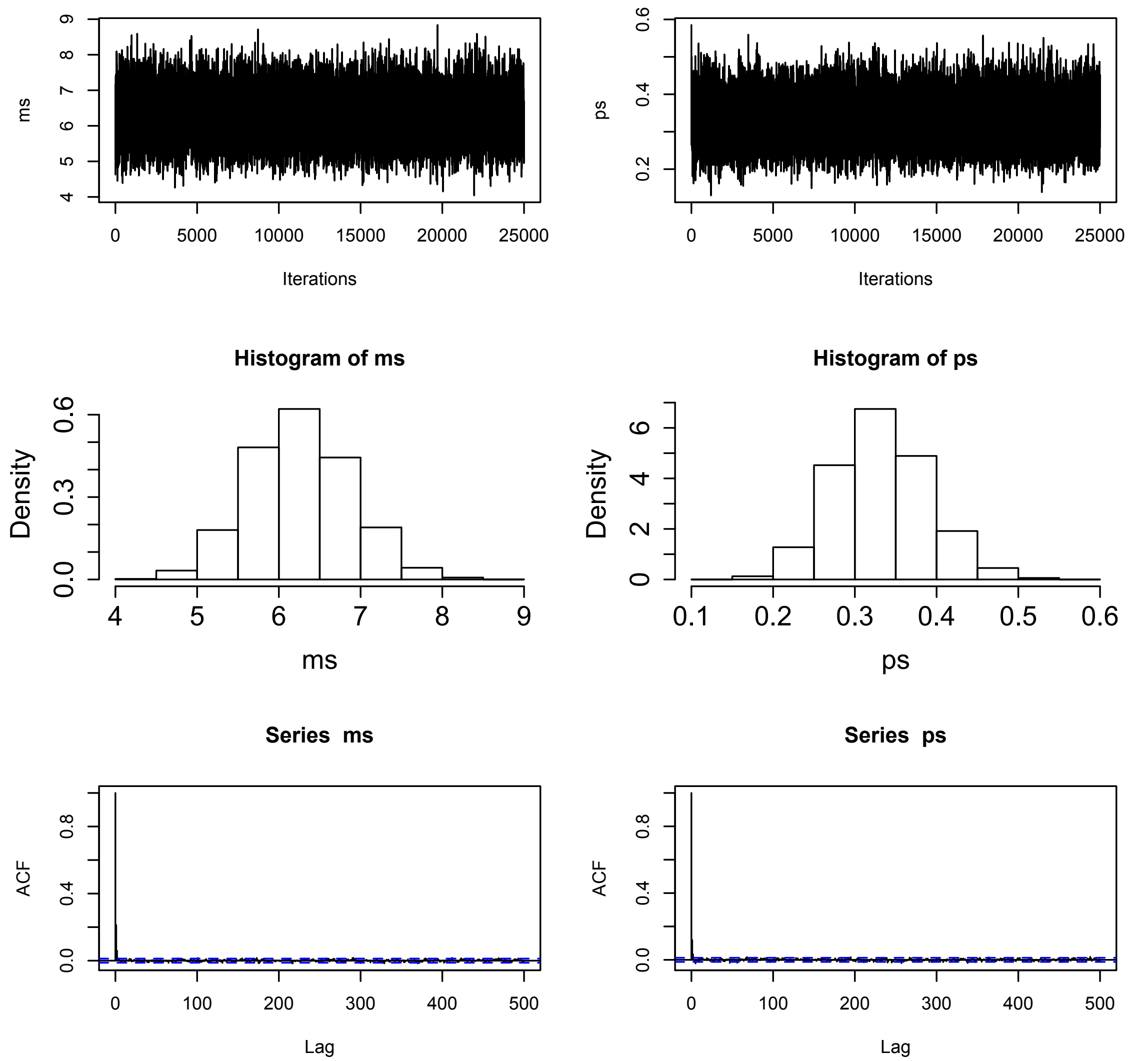
```
	#Calculate the vector of probabilities where each value represents the probability that the corresponding value in the data vector x was drawn from distribution 1, the lower mean distribution
	probz1 = (1-p)*exp(-1)/factorial(x)
	probz2 = p*Y^x*exp(-Y)/factorial(x)
	probz = probz1 + probz2
	pvec = probz1/probz

	#Calculate the z vector of latent indicator values where each value is either 0 or 1 to indicate which distribution the corresponding value in the data vector x is from - "0" indicates distribution 2 and "1" indicates distribution 1
	z = rbinom(n,1,pvec)

	#Compute a mean for the mixed model from a gamma distribution using the indicator variable vector z and the data vector x.
	Y = rgamma(1,(sum(x)-sum(x*z)+1),(n-sum(z)+1/10))

	#Compute a p for the probability that the data value comes from distribution 1
	p = rbeta(1,n-sum(z)+1,sum(z)+1)

	#if i is greater than burnin and if i is a multiple of the lag, store z, Y, and p
  if(i > burnin) {
	  if(i %% lag == 0) {
		zs <- c(zs,z)
		Ys <- c(Ys,Y)
		ps <- c(ps,p)
   }
  }
}

  vectors <- list("zs" = zs, "Ys" = Ys, "ps" = ps)
  return(vectors)

}

vectors = gibbs(x,p,Y,N,lag,burnin)

#vectors$zs represents Gibbs sampler realizations for the z vector
#vectors$Ys represents Gibbs sampler realizations for the mean+1 of distribution 2
#vectors$ps represents Gibbs sampler realizations for the probability that an x value comes from distribution 1
zs = vectors$zs
Ys = vectors$Ys
```

Lisa Over
Final Exam Problem 3
May 5, 2015

```
ps = vectors$ps

#Since Y = m + 1, subtract 1 from Ys values to create a vector of m values
ms = Ys - 1

par(mfrow=c(3,2)) #split plotting window into 2 rows and 2 columns
ts.plot(ms,xlab="Iterations")
ts.plot(ps,xlab="Iterations")
hist(ms,probability=T, cex.lab=1.5, cex.axis=1.5)
hist(ps,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(ms,lag.max=500)
acf(ps, lag.max=500)

#Convert z to an n column matrix
zmat = matrix(zs, ncol=length(x), byrow=TRUE)
#Get column means of matrix zmat
zimeans = colMeans(zmat)

plot(x,zimeans)

mean(ms)
mean(ps)
quantile(ms, 0.025)
quantile(ms, 0.975)
quantile(ps, 0.025)
quantile(ps, 0.975)
```

Lisa Over
Final Exam Problem 3
May 5, 2015

RESULTS



Lisa Over
Final Exam Problem 3
May 5, 2015

In the plot below, the proportion of 1s for each x_i, i.e., the means of the 25,000 z_i values generated for each x_i, is plotted against the corresponding x_i value in the sample data. The proportion of 1s indicates the probability that the corresponding value was drawn from distribution 1, the lower mean distribution. The x values that could only be drawn from distribution 1 are those forming a horizontal line at z=1. The x values that could only be drawn from distribution 2, the higher mean distribution, are those forming a horizontal line at z=0. The x values that correspond to $0 < z < 1$ are those that could have been drawn from either distribution 1 or 2. There is a negative relationship between the proportion of 1s and the x values. As the x values increase, they become less likely to have come from distribution 1 and more likely to have come from distribution 2.

The mean of the distribution of m values (calculated as Y – 1) is 6.26. The mean of the distribution of p values is 0.33. Therefore, the predicted mean of distribution 2 is 6.26 with a 95% credible interval of (5.09, 7.50), and the predicted proportion of values drawn from distribution 1 is 0.33 with a 95% credible interval of (0.2256, 0.4507). Therefore, the predicted proportion of values drawn from distribution 2 is 0.69 with a 95% credible interval of (0.5493, 0.7744).