STA 601 Homework 10

Melanie Lai Wai

due November 30th, 2017

Exercise 11.5

Part (a)

Model :

$$Y_i \mid \theta_i, x_i \sim Poi(\theta_i x_i)$$
$$\theta_1, ..., \theta_6 \mid a, b \sim Gamma(a, b)$$
$$a \sim Gamma(1, 1)$$
$$b \sim Gamma(10, 1)$$

Y_i can be interpreted as the number of occurrences of the birth defect in county i in a five-year period. θ_i is the rate of the disease in county i and x_i is the population of county i, such that the expected number of occurences of the birth defect is $\theta_i x_i$. The sampling distribution of the disease rate is a gamma distribution with parameters a, b, with mean a/b. a/b can be interpreted as the expectation of θ_i, the rate of the disease in county i, while y_i/x_i is the observed disease rate.

Part (b)

$$\begin{aligned} p(\theta_1, ..., \theta_6 \mid a, b, x, y) &\propto p(y \mid \theta_1, ..., \theta_6, x) p(\theta_1, ..., \theta_6 \mid a, b) p(a) p(b) \\ &= \prod_{i=1}^{i=6} \frac{e^{-\theta_i x_i} (\theta_i x_i)^{y_i}}{y_i!} \prod_{i=1}^{i=6} \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-b\theta_i} e^{-a} \frac{1}{\Gamma(10)} b^9 e^{-b} \\ &= \prod_{i=1}^{i=6} e^{\theta_i x_i} (\theta_i x_i)^{y_i} \theta_i^{a-1} e^{-b\theta_i} \\ &= \prod_{i=1}^{i=6} e^{-(b+x_i)\theta_i} \theta_i^{a+y_i-1} \end{aligned}$$

The expression above is a product of Gamma kernels. Hence the full conditional distribution for a single θ_i is a $Gamma(a + y_i, b + x_i)$ for each county i.

Part (c)

$$\begin{aligned}
\frac{p(a^*, b^*, \boldsymbol{\theta} \mid y)}{p(a, b, \boldsymbol{\theta} \mid y)} &= \frac{p(y \mid a^*, b^*, \boldsymbol{\theta})}{p(y \mid a, b, \boldsymbol{\theta})} \frac{p(\boldsymbol{\theta} \mid a^*, b^*)}{p(\boldsymbol{\theta} \mid a, b)} \frac{p(a^*)p(b^*)}{p(a)p(b)} \\
&= \frac{p(\boldsymbol{\theta} \mid a^*, b^*)}{p(\boldsymbol{\theta} \mid a, b)} \frac{p(a^*)p(b^*)}{p(a)p(b)} \text{ because } \boldsymbol{Y} \text{ depends on } a, b \text{ through } \boldsymbol{\theta} \\
&= \frac{\prod_{j=1}^{j=6} dgamma(\theta_i \mid a^*, b^*)}{\prod_{j=1}^{j=6} dgamma(\theta_i \mid a, b)} \frac{dgamma(a^*)dgamma(b^*)}{dgamma(a)dgamma(b)}
\end{aligned}$$

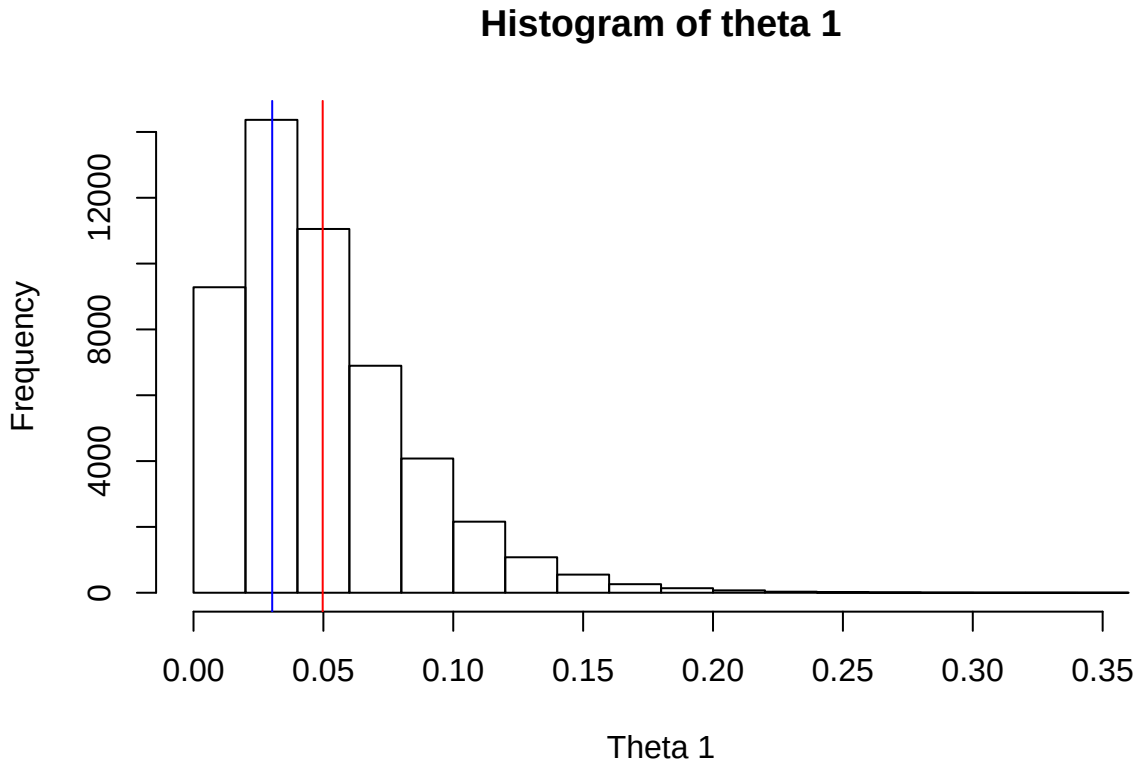
Part (d)

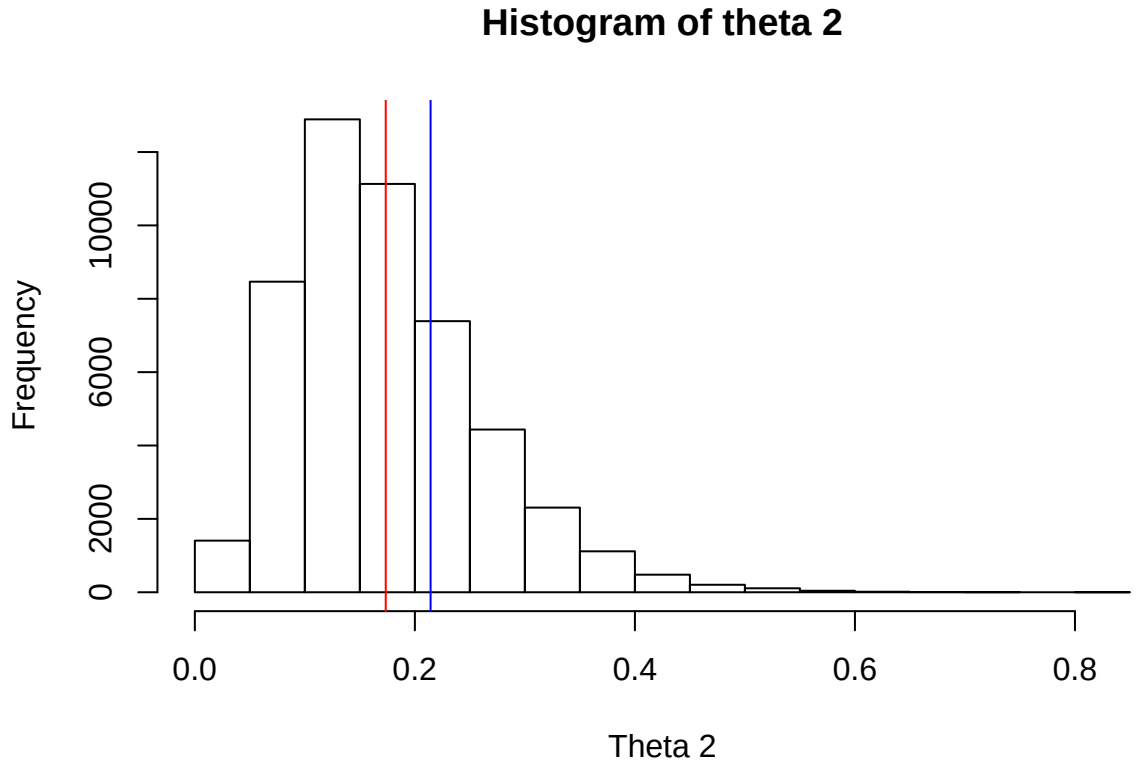
Metropolis-Hastings Algorithm: code is at the end of the exercise.

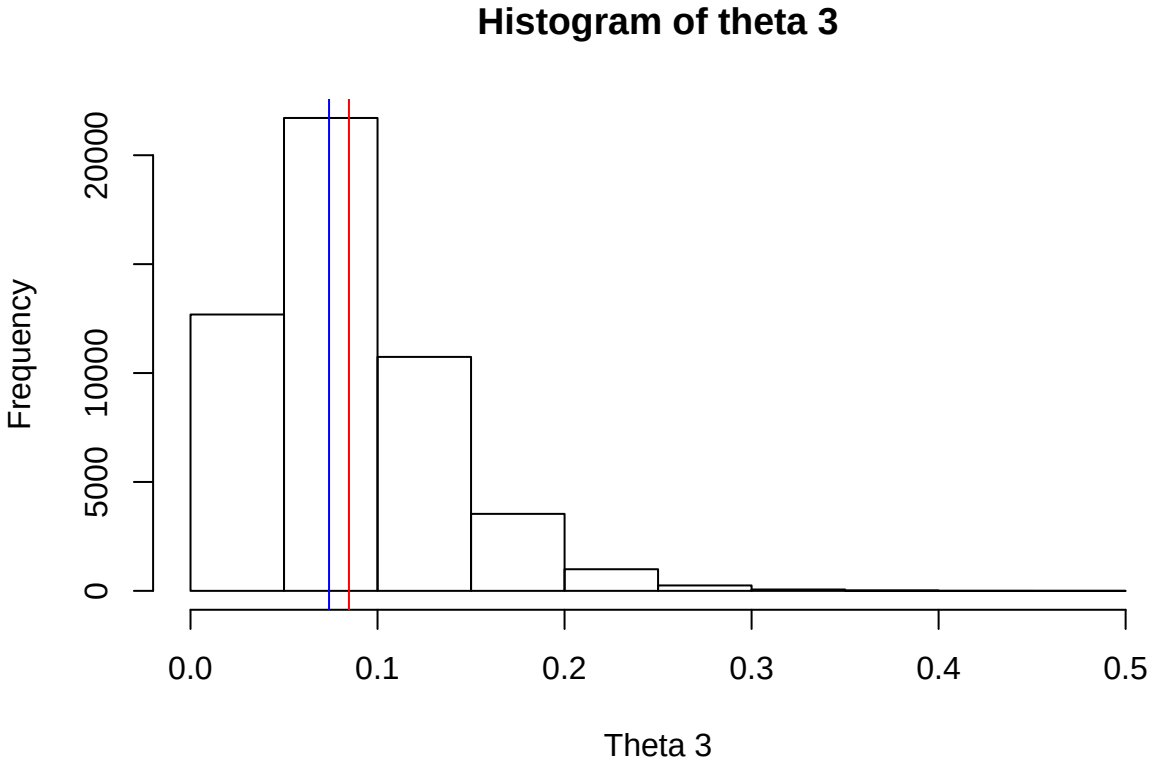
Part (e): Posterior Inference

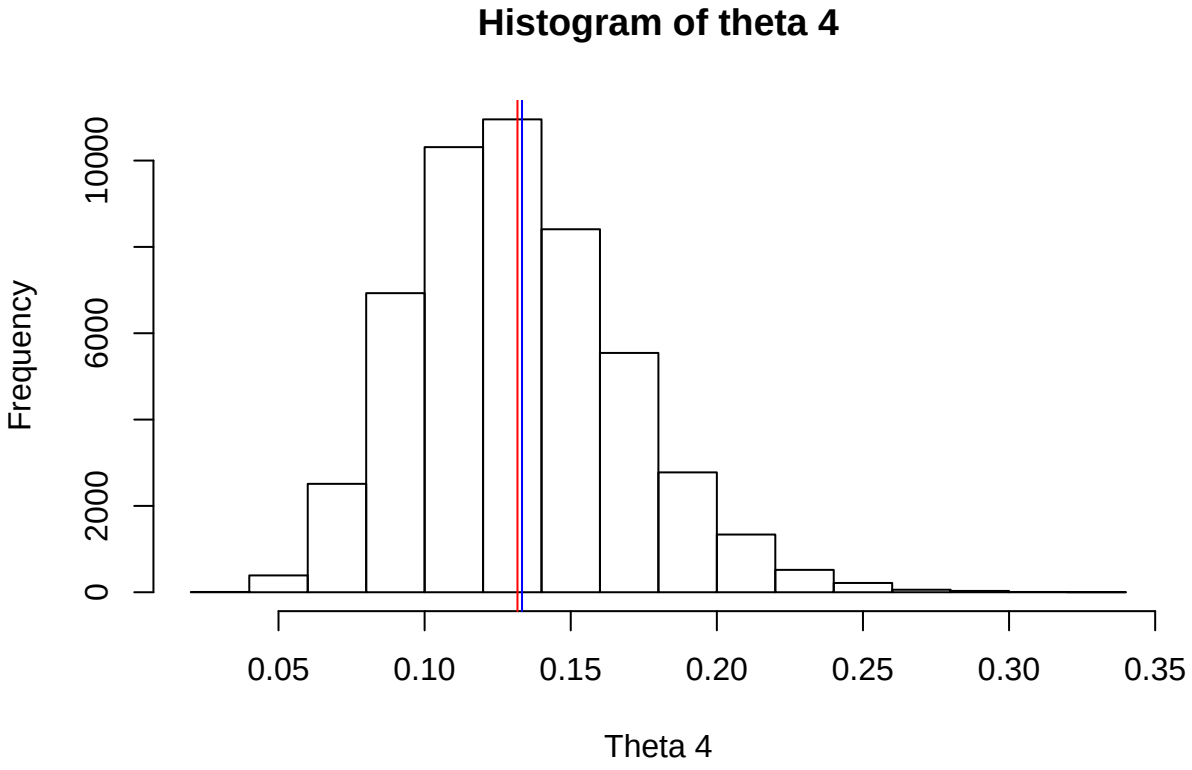
Part (e)(i)

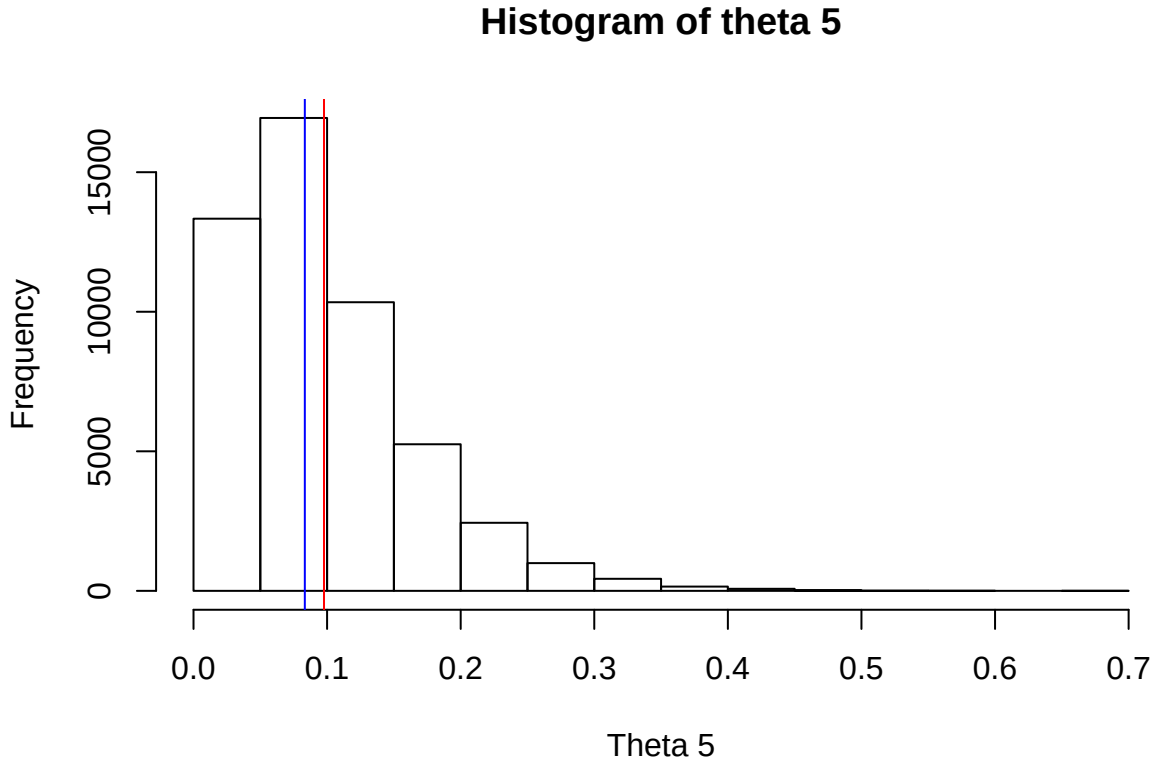
For all counties, the observed values of y_i/x_i (in blue) are relatively close to the posterior mean value (in red). The observed rate and posterior mean of the disease rate for county 4 are especially close to each other. It seems that the Poisson hierarchical model is a good fit for the data.

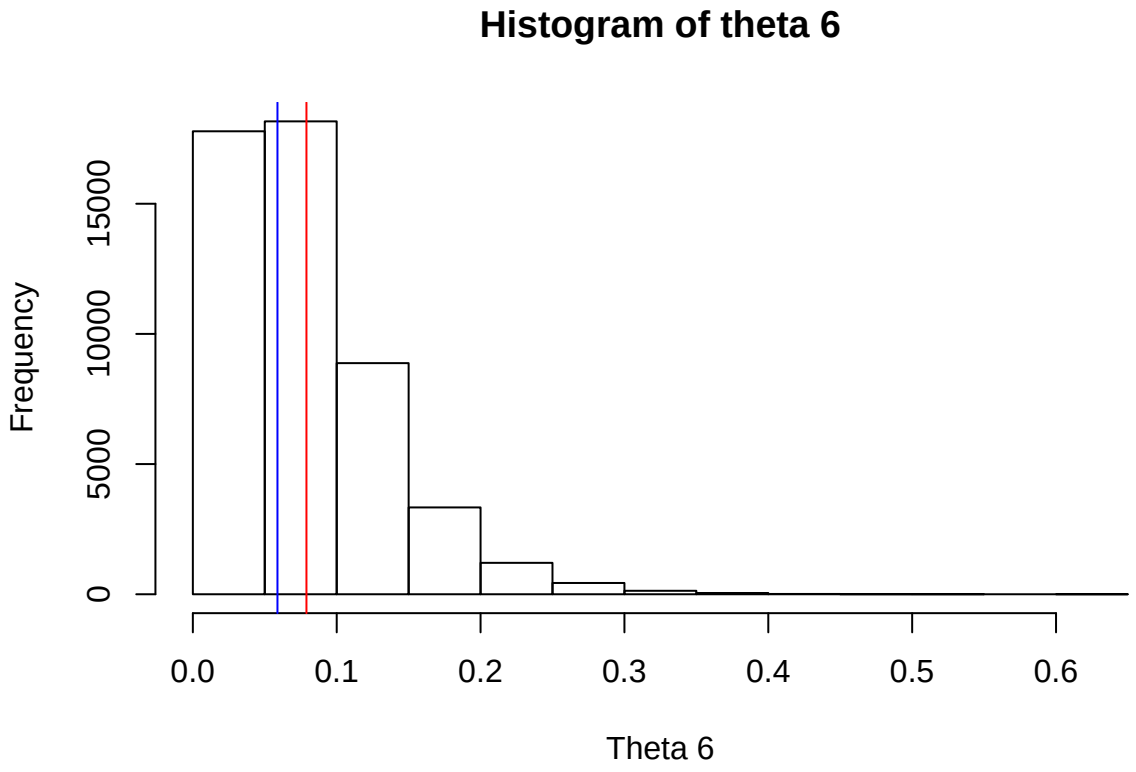



Histogram of theta 2
Frequency
0
2000
6000
10000
0.0
0.2
0.4
0.6
0.8
Theta 2


Histogram of theta 3
Frequency
0
5000
10000
20000
0.0
0.1
0.2
0.3
0.4
0.5
Theta 3

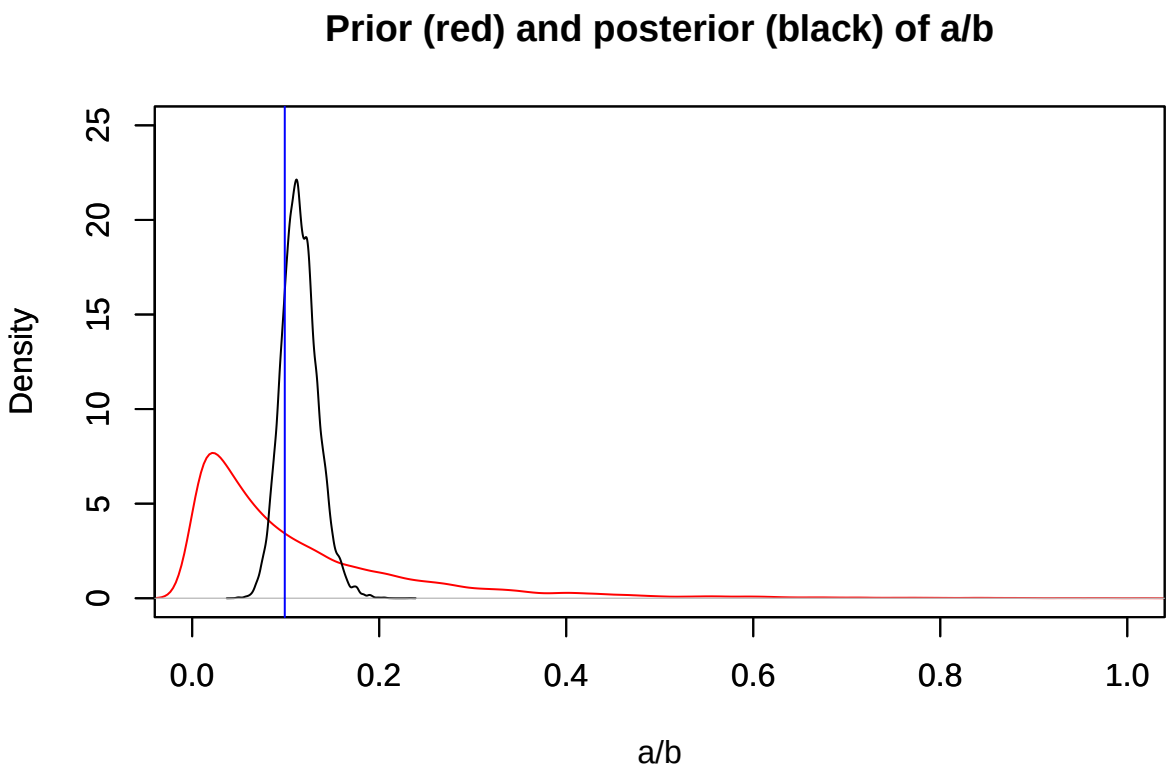

Histogram of theta 4
Frequency
0
2000
6000
10000
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Theta 4


Histogram of theta 5
Frequency
0
5000
10000
15000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Theta 5



Part (e)(ii)

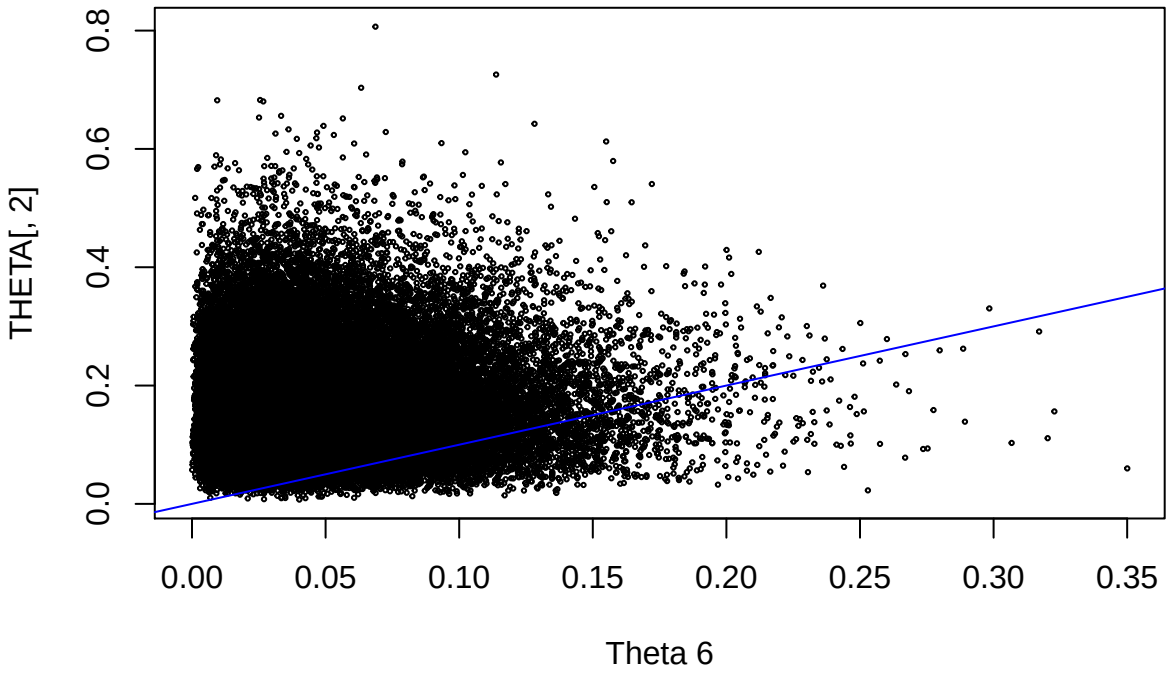
The prior density for a/b is flatter than the posterior and has a mean close to 0. After observing the disease rates in each county, the posterior distribution of a/b has a sharper peak than the prior with mean around 0.115. The observed average for y_i/x_i is 0.099, slightly lower than the posterior mean.

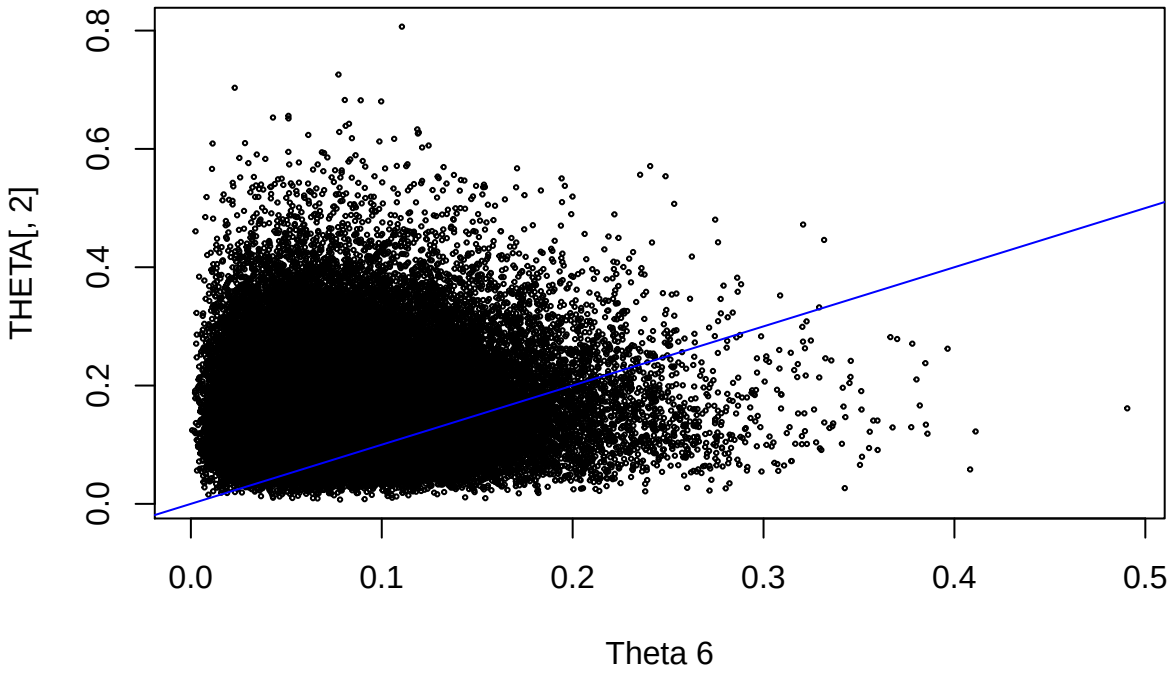


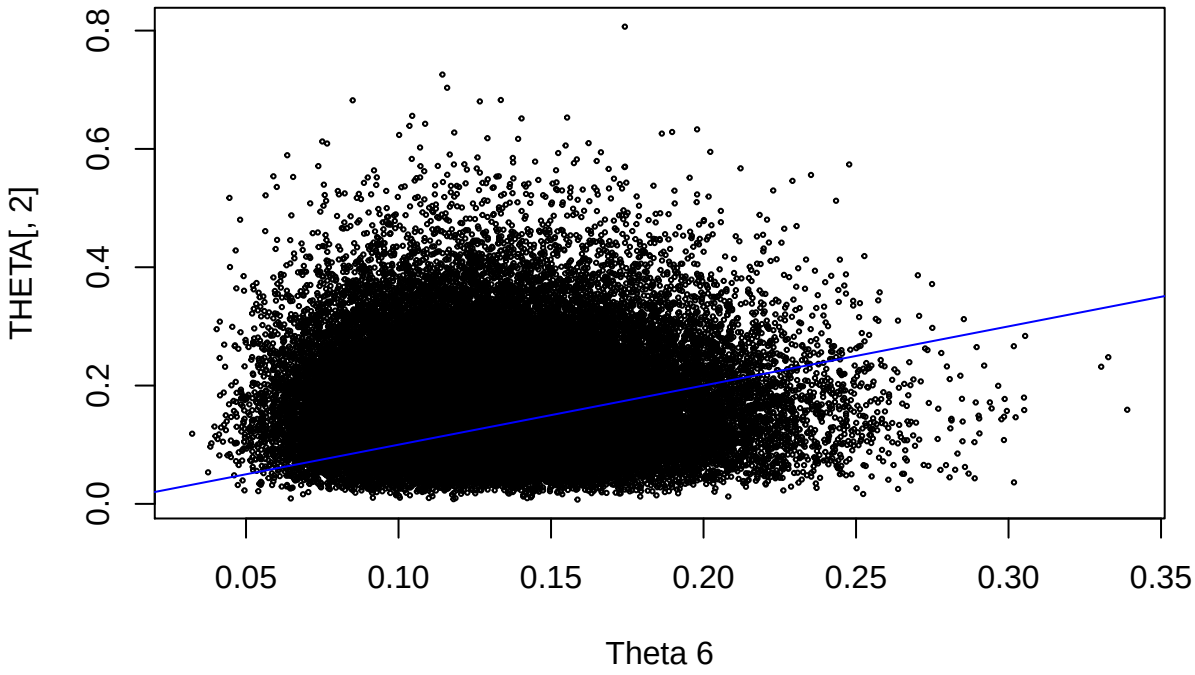
Part (e)(iii)

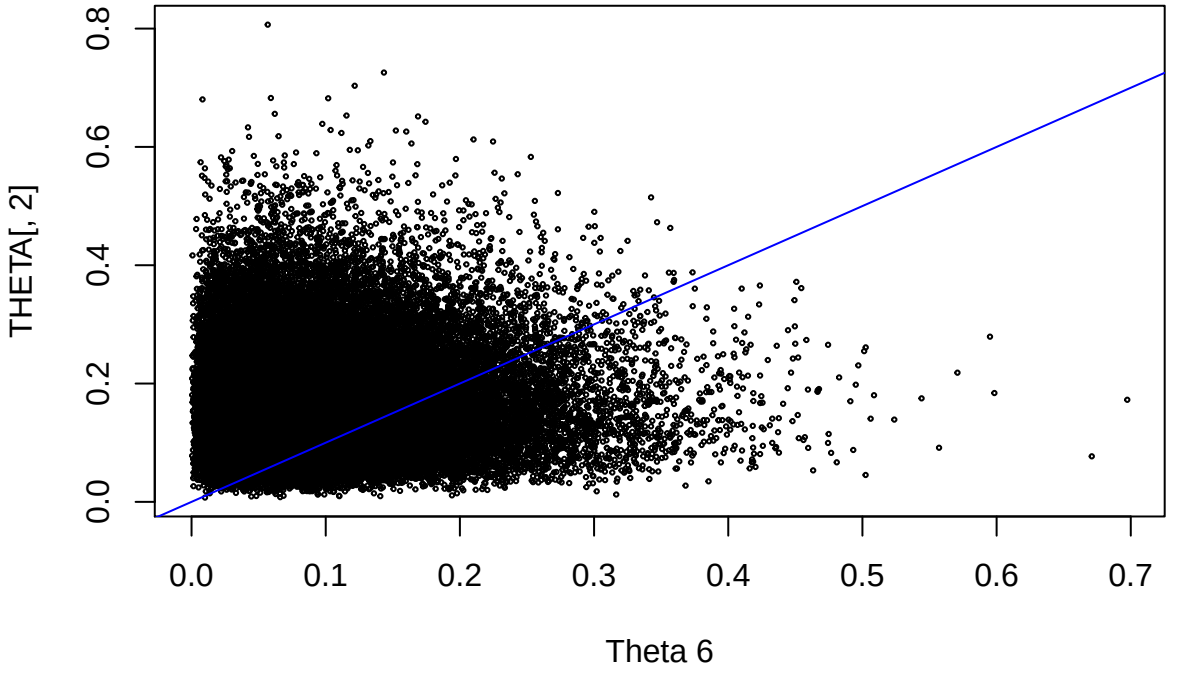
From the plots of θ_2 against each of the other θ_i's, it seems that the majority of points lie above the 45 degree line, indicating that θ_2 may be larger than the other θ_i's. From the table, we see that it is very likely that θ_2 is larger than the others. Comparing those probabilities to the observed y_i/x_i, we see that the rate of disease in county 2 is indeed higher than all of the other counties.

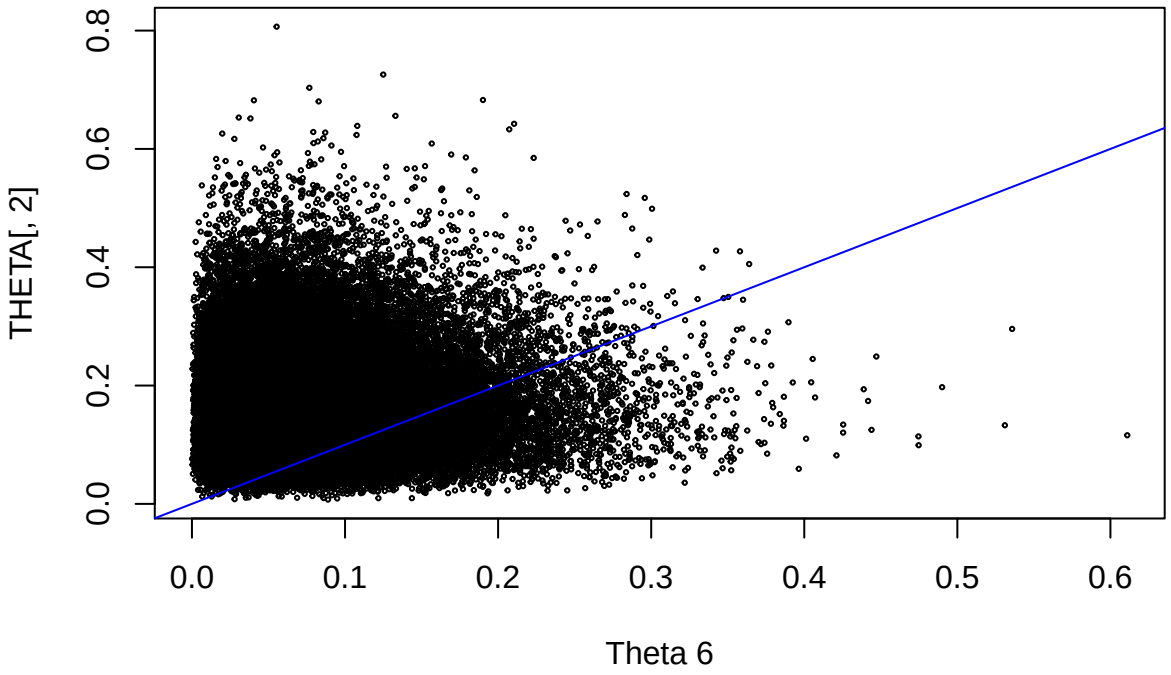
However, the posterior probability that θ_2 is the highest among all counties is 0.54, indicating that the chance that county 2 has the highest disease rate is actually not as high as we initially thought. If we based our conclusions on the observed data only, we might mistakenly believe that there is a strong association between the toxic soil samples in county 2 and the disease rate, when in fact there is only a 54% chance that county 2 will have the highest disease rate.


THETA[, 2]
0.0
0.2
0.4
0.6
0.8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Theta 6


THETA[, 2]
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
0.5
Theta 6


THETA[, 2]
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Theta 6


THETA[, 2]
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Theta 6



i	Prob. Theta 2 is bigger than Theta i
1	0.94044
2	NA
3	0.83218
4	0.64230
5	0.77792
6	0.84316

R code

```
library(coda)
library(knitr)

## Part (d)
## Initial Values
Y <- c(1,3,2,12,1,1)  # number of occurrences of disease
X <- c(33,14,27,90,12,17) # population of each county
n <- 50000 # number of MCMC samples
m <- length(Y) # number of counties
THETA <- matrix(NA, n, m)
A <- B <- NULL
theta <- Y/X # observed disease rates
a <- 1 # expected prior value of a
b <- 10 # expected prior value of b
delta.a <- 0.05*sd(Y) # tuning parameters
delta.b <- 0.05*sd(X)

## MCMC
set.seed(2)
for (s in 1:n)
{
  # Metropolis step: update a and b
  ## Propose new value for a and b from symmetric distribution from Exercise 10.1
  a.temp <- rnorm(1, a, delta.a)
  ifelse(a.temp<0, a.star <- -1*a.temp, a.star <- a.temp)
  b.temp <- rnorm(1, b, delta.b)
  ifelse(b.temp<0, b.star <- -1*b.temp, b.star <- b.temp)

  ## Compute acceptance ratio
  log.ratio = sum(dgamma(theta, a.star, b.star, log = TRUE) - dgamma(theta, a, b, log = TRUE)) +
    dgamma(a.star, 1, 1, log=TRUE) + dgamma(b.star, 10, 1, log = TRUE) -
    dgamma(a, 1, 1, log=TRUE) - dgamma(b, 10, 1, log = TRUE)

  ## Compare log.u from uniform (1,1) distribution with log.ratio
  ## and accept proposed a's and b's if log u < log.ratio
  if (log(runif(1, 1, 1)) < log.ratio)
  {
    A[s] <- a <- a.star
    B[s] <- b <- b.star
  }
  else
  {
    A[s] <- a
    B[s] <- b
  }

  # Gibbs step: update theta
  for (j in 1:m)
  {
    theta[j] <- rgamma(1, a + Y[j], b + X[j])
    THETA[s,j] <- theta[j]
```

```
  }
}

## Save some THETA output
THETA.THIN <- matrix(NA, ncol = m)
for (s in 1:n)
{
  if (s%%10==0) {THETA.THIN <- rbind(THETA.THIN, THETA[s,])}
}
THETA.THIN <- THETA.THIN[-1,]

## Diagnostic Tests
plot(THETA.THIN[,1], type="l")
acf(THETA.THIN[,1])

## Effective Sample Sizes
effectiveSize(THETA[,1])
effectiveSize(THETA[,1])
effectiveSize(THETA[,2])
effectiveSize(THETA[,3])
effectiveSize(THETA[,4])
effectiveSize(THETA[,5])
effectiveSize(THETA[,6])
effectiveSize(A)
effectiveSize(B)

## Autocorrelation
acf(THETA[,1])
acf(THETA[,2])
acf(THETA[,3])
acf(THETA[,4])
acf(THETA[,5])
acf(THETA[,6])
acf(A)
acf(B)

## (e)(i) compare posterior distributions of theta_i with observed y_i/x_i
for (i in 1:m)
{
hist(THETA[,i], main = paste("Histogram of theta", i),
     xlab = paste("Theta", i))
abline(v=mean(THETA[,i]), col = "red")
abline(v=Y[i]/X[i], col = "blue")
}

## (e)(ii) compare posterior distribution of a/b with prior distribution and observed y_i/x_i
## Prior for a/b
x <- seq(1,100, by=0.01)
## Gamma (1,1) for a and Gamma(10,1) for b
joint.prior.samps = c()
joint.prior.samps = rgamma(10000, 1, 1)/rgamma(10000, 10, 1)
post.ab = A/B
```

```r
## plot distributions
plot(density(joint.prior.samps), col = "red",
     xlim = c(0,1), ylim = c(0,25),
     main = "Prior (red) and posterior (black) of a/b",
     xlab = "a/b")
par(new = TRUE)
plot(density(post.ab),
     xlim = c(0,1), ylim = c(0,25),
     main = "",
     xlab = "")
abline(v=mean(Y/X), col = "blue")

## (e)(iii) Plot samples of theta_2 vs theta_j for j not equal to 2
prob.twoisbigger = c()
for (j in c(1,3,4,5,6))
{
  prob.twoisbigger[j] = mean(THETA[,2]>THETA[,j])
  plot(THETA[,j], THETA[,2], cex=0.3,
       xlab = paste("Theta", i))
  abline(0, 1, col = "blue")
}
kable(prob.twoisbigger, col.names = "Prob. Theta 2 is bigger than Theta i",
    rownames = "i", row.names = TRUE)

v <- c()
for (i in 1:n)
{
v[i] <- THETA[i,2]==max(THETA[i,])
}

prob.twoismax = mean(v)
```