Homework 2

Olivia Hackworth

3/11/2019

```
library(tidyverse)
```

A survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. Table 3.3 (BDA3) displayes the number of bicycles and other vehicles recorded in the study.

1. For this part, restrict your attention to the first four rows of the table: the data on residential streets.

a Let $y_1, ..., y_10$ and $z_1, ..., z_8$ be the observed proportion of traffic that was on bicycles in the residential streets with bike lanes and with no bike lanes, respectively. Set up a model so that the y_i's are independent and identically distributed given parameters θ_y and the z_i's are independent and identically distributed given parameters θ_z.

b Set up a prior distribution that is independent in θ_y and θ_z.

c Determine the posterior distribution for the parameters in your model and draw 1000 simulations from the posterior distribution (Hint: θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)

d Let $\mu_y = E(y_i|\theta_y)$ be the mean of the distribution of the y_i's; μ_y will be a function of θ_y. Similarly, define μ_z. Using your posterior simulations from 1c, plot a histogram of the posterior simulations of $\mu_y - \mu_z$, the expected difference in proportions in bicycle traffice on residential streets with and without bike lanes.

1

```
yvec = c(16/(58+16),9/(90+9),10/(48+10),13/(57+13),19/(103+19),20/(57+20),18/(86+18),17/(112+17),35/(273
```

```
zvec = c(12/(113+12),1/(18+1),2/(14+2),4/(44+4),9/(208+9),7/(67+7),9/(29+9),8/(154+8))
```

a

$y_i \sim Beta(\alpha_y, \beta_y)$
$z_i \sim Beta(\alpha_z, \beta_z)$

$p(y_1, ..., y_n|\theta_y) = \prod_{i=1}^{n} p(y_i|\theta_y)$
$\propto \prod_{i=1}^{n} y_i^{\alpha_y - 1}(1 - y_i)^{\beta_y - 1}$

$p(z_1, ..., z_n|\theta_z) = \prod_{i=1}^{n} p(z_i|\theta_z)$
$\propto \prod_{i=1}^{n} z_i^{\alpha_z - 1}(1 - z_i)^{\beta_z - 1}$

b

One option for a prior that is independent for θ_y and θ_z is Jeffreys' prior which is the Fisher information matrix raised to the 1/2. In this case, the prior is

$p(\alpha_i, \beta_i) = (\alpha_i + \beta_i)^{-5/2}$.

c

```
#make grid
alpha = seq(0.01,100, length.out = 1000)
beta = seq(0.01,100, length.out = 1000)
grid = expand.grid(x=alpha,y=beta)
dalpha = diff(alpha)[1]
dbeta = diff(beta)[1]


#posterior function

posty = function(alpha,beta){
  yvec = c(16/(58+16),9/(90+9),10/(48+10),13/(57+13),19/(103+19),20/(57+20),18/(86+18),17/(112+17),35/(1
  #p = prod((yvec)^(alpha-1)*(1-yvec)^(beta-1))*(alpha+beta)^(-5/2)
  p = exp(sum(dbeta(yvec,alpha,beta,log=T)) - 5/2* log(alpha+beta))
  return(p)
}


postz = function(alpha,beta){
  zvec = c(12/(113+12),1/(18+1),2/(14+2),4/(44+4),9/(208+9),7/(67+7),9/(29+9),8/(154+8))
  p = exp(sum(dbeta(zvec,alpha,beta,log=T)) - 5/2* log(alpha+beta))
  return(p)
}

#apply function to grid
unnormalized_posty = apply(grid,1,function(x) posty(x[1],x[2]))
unnormalized_postz = apply(grid,1,function(x) postz(x[1],x[2]))

#normalizing constants
c_y = sum(unnormalized_posty) * dalpha * dbeta
c_z = sum(unnormalized_postz) * dalpha * dbeta

#posterior distribution - normalized
normalized_posty = (1/c_y)*unnormalized_posty
normalized_postz = (1/c_z)*unnormalized_postz

#draw samples from posterior

sampy = grid[sample(1:nrow(grid),size = 1000, replace = T,prob=normalized_posty),]
head(sampy)
```

```
##               x         y
## 81025  2.412162  8.117297
## 142034 3.312973 14.222793
## 242063 6.215586 24.231802
## 183046 4.514054 18.326486
## 102023 2.211982 10.219189
## 140036 3.513153 14.022613
```

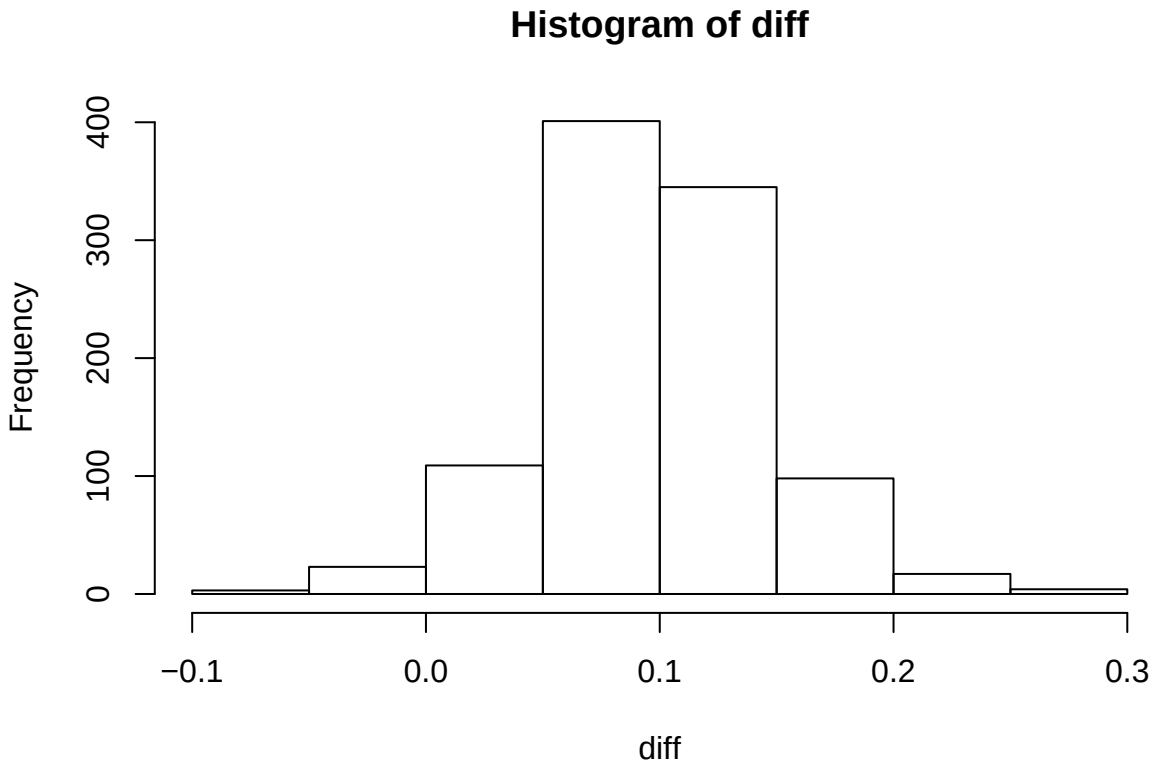
```
#mean(sampy[,1])
#mean(sampy[,2])

sampz = grid[sample(1:nrow(grid),size = 1000, replace = T,prob=normalized_postz),]
head(sampz)
```

```
##               x         y
## 453048 4.714234 45.350811
## 113016 1.511351 11.320180
## 76014  1.311171  7.616847
## 90014  1.311171  9.018108
## 217032 3.112793 21.729550
## 374046 4.514054 37.443694
```

```
#mean(sampz[,1])
#mean(sampz[,2])
```

d

```
mu_y = sampy[,1]/(sampy[,1]+sampy[,2])
mu_z = sampz[,1]/(sampz[,1]+sampz[,2])

diff = mu_y-mu_z
hist(diff)
```



2. For this problem, restrict your attention to the first two rows of the table: residential streets labeled as 'bike routes', which we will use to illustrate this computational exercise.

a Set up a model for the data in Table 3.3 so that, for j = 1,...,10, the observed number of bicycles at

location j is binomial with unknown probability θ_j ans samples size equal to the total number of vehicles (bicycles included) in that block. The parameter θ_j can be interpreted as the underlying or 'true' proportion of traffice at location j that is bicycles. Assign a beta population distribution for the parameters θ_j and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3 of BDA. Write down the joint posterior distribution.

b Compute the marginal posterior density of the hyperparameters and draw simulations from teh joing posterior distribution of the parameters and hyperparameters.

c Compare the posterior distributions of the parameters θ_j to the raw proportions, (number of bicycles/total number of vehicles) in location j. How do the inferences from the posterior distribution differ from the raw proportions?

d Give a 95% posterior interval for the average underlying proportion of traffice that is bicycles.

e A new city block is sampled at random and is a residential street with a bike route. In an hour of observation, 100 vehicles of all kinds go by. Give a 95% posterior interval for the number of those vehicles that are bicycles. Discuss how much you trust this interval in application.

f Was the beta distribution for the θ_j's reasonable?

2

Page 110!

a

```
yvec = c(16,9,10,13,19,20,18,17,35,55)

nvec = c(58+16,90+9,48+10,57+13,103+19,57+20,86+18,112+17,273+35,64+55)
nvec
```

```
##  [1]  74  99  58  70 122  77 104 129 308 119
```

The likelihood is binomial, the prior is beta, the hyperprior is Jeffreys' prior.

Likelihood: $p(y_j|\theta_j, n_j, \alpha, \beta) \propto \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$

prior: $p(\theta_j|\alpha, \beta) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}$

hyperprior: $p(\alpha, \beta) = (\alpha+\beta)^{-5/2}$

Joint posterior: $p(\alpha, \beta, \theta|y) \propto p(\alpha, \beta) * \prod_{j=1}^{J} p(\theta_j|\alpha, \beta) * \prod_{j=1}^{J} p(y_j|\theta_j, n_j, \alpha, \beta)$
$\propto (\alpha+\beta)^{-5/2} * \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} * \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$

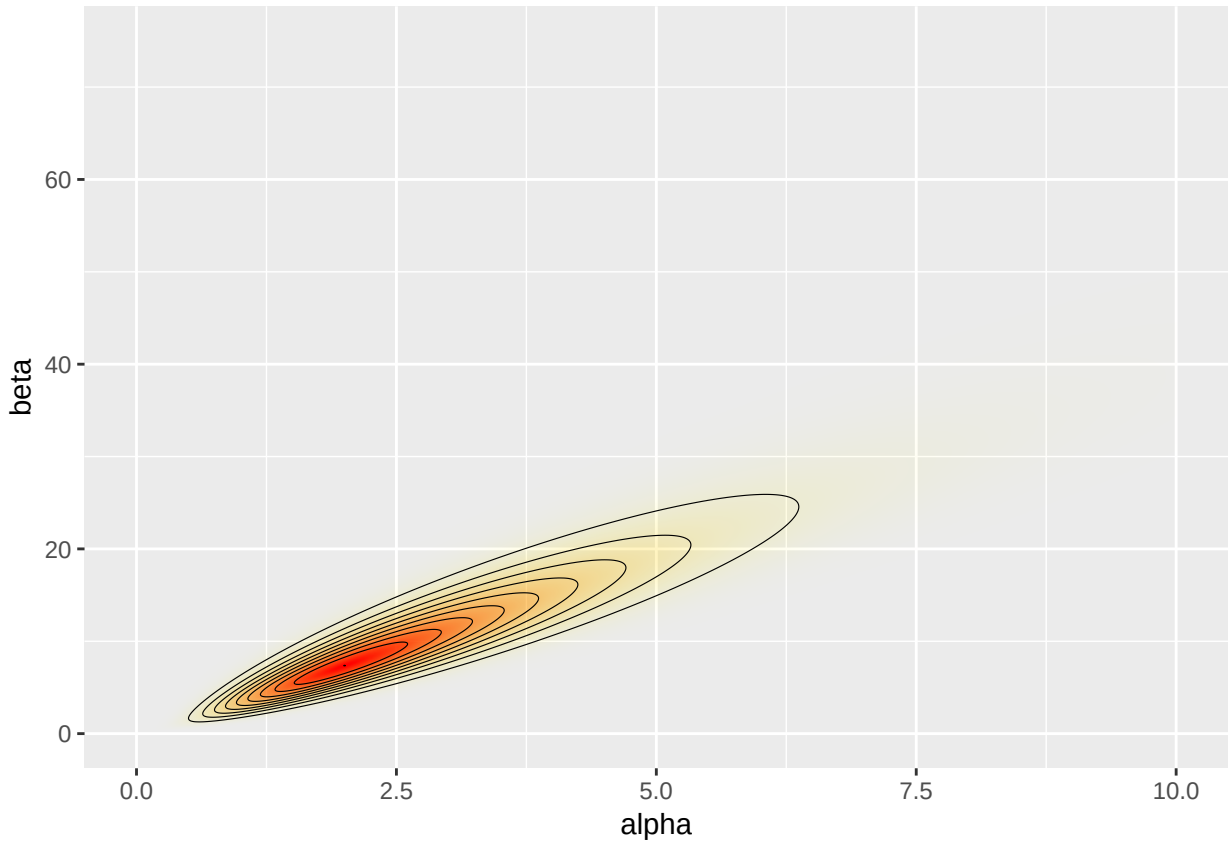
b

From BDA3 page 110:
$p(\alpha, \beta|y) \propto (\alpha+\beta)^{-5/2} * \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} * \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$

computing the marginal distribution of alpha and beta

```
alpha = seq(0.01,10, length.out = 1000)
beta = seq(0.01,75, length.out = 1000)
grid = expand.grid(x=alpha,y=beta)
```

```r
logmargposthyperfunc = function(a, b){
  #data
  y = c(16,9,10,13,19,20,18,17,35,55)
  n = c(58+16,90+9,48+10,57+13,103+19,57+20,86+18,112+17,273+35,64+55)

  #log of marginal posterior
  p = log(a+b)*(-5/2) +
  sum(lgamma(a+b)-lgamma(a)-lgamma(b)+lgamma(a+y)+lgamma(b+n-y)-lgamma(a+b+n))
  return(p)
  }

logmargposthyper = apply(grid,1,function(x) logmargposthyperfunc(x[1],x[2]))


#subtracting max value to avoid over/underflow in exponentation
df_marg = data.frame(grid, p = exp(logmargposthyper - max(logmargposthyper)))

probs = exp(logmargposthyper - max(logmargposthyper))

#plot

ggplot(data = df_marg, aes(x = x, y = y)) +
  geom_raster(aes(fill = p, alpha = p), interpolate = T) +
  geom_contour(aes(z = p), colour = 'black', size = 0.2) +
  coord_cartesian(xlim = c(0,10), ylim = c(0, 75)) +
  labs(x = 'alpha', y = 'beta') +
  scale_fill_gradient(low = 'yellow', high = 'red', guide = F) +
  scale_alpha(range = c(0, 1), guide = F)
```



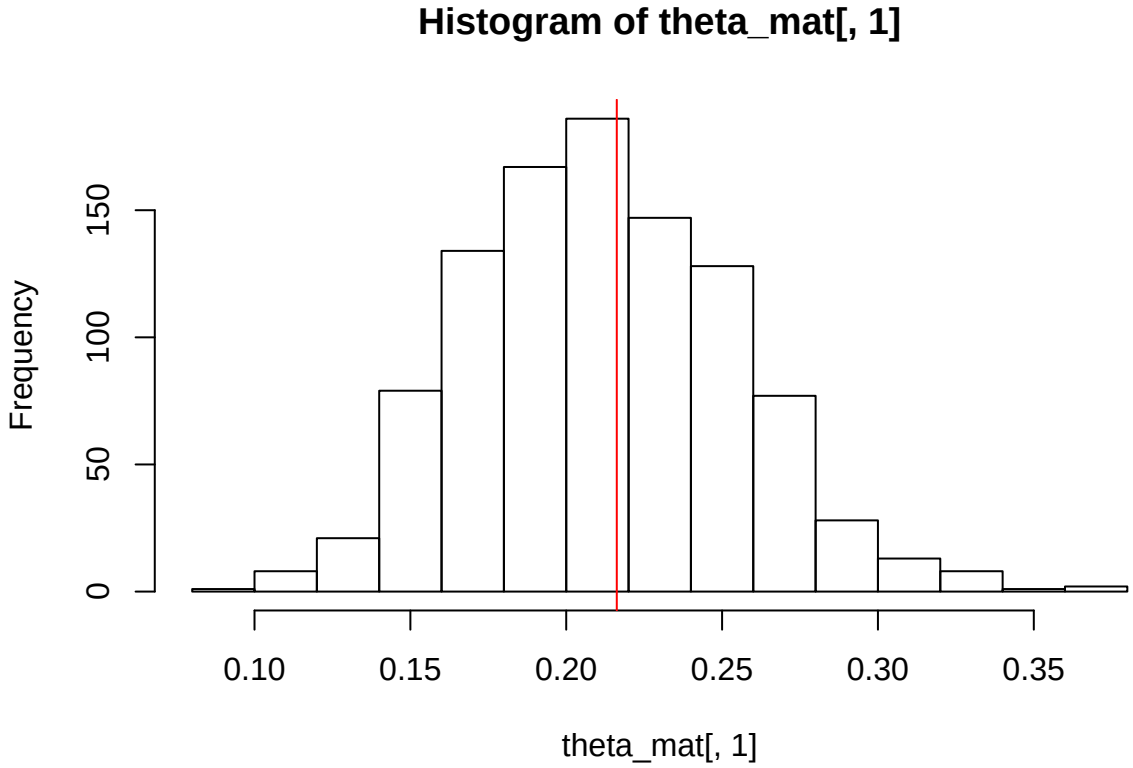
pull samples from marginal post of hyperparameters and then use those samples in a conjugate beta posterior to pull the posterior samples of the thetas

```
gridrows = sample(1:nrow(grid), size = 1000, replace = T, prob=probs)
alpha_beta_samples = grid[gridrows,] #x = alpha, y = beta

#pull theta samples
theta_1 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[1], x[2]+nvec[1]-yvec[1]))

theta_2 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[2], x[2]+nvec[2]-yvec[2]))

theta_3 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[3], x[2]+nvec[3]-yvec[3]))

theta_4 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[4], x[2]+nvec[4]-yvec[4]))

theta_5 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[5], x[2]+nvec[5]-yvec[5]))

theta_6 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[6], x[2]+nvec[6]-yvec[6]))

theta_7 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[7], x[2]+nvec[7]-yvec[7]))

theta_8 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[8], x[2]+nvec[8]-yvec[8]))

theta_9 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[9], x[2]+nvec[9]-yvec[9]))

theta_10 = apply(alpha_beta_samples, 1, function(x) rbeta(1,x[1]+yvec[10], x[2]+nvec[10]-yvec[10]))
```

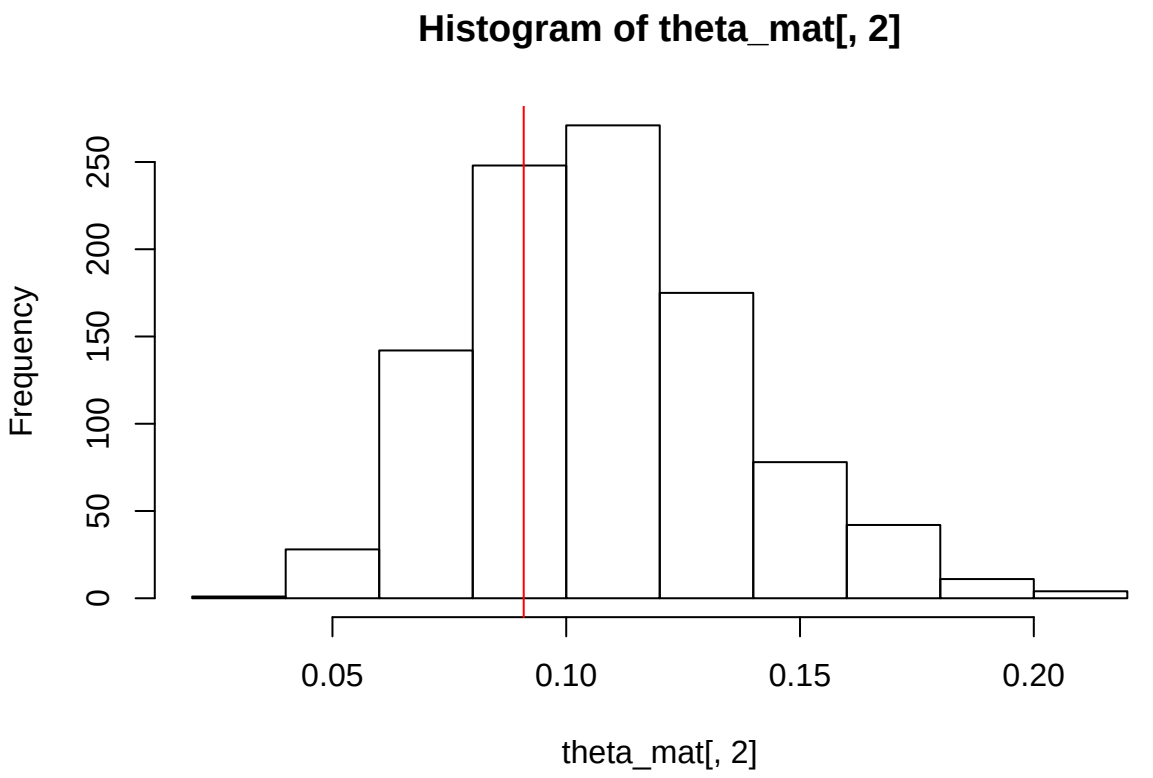
```
theta_mat = cbind(theta_1,theta_2,theta_3,theta_4,theta_5,theta_6,theta_7,theta_8,theta_9,theta_10)

joint_post_samps = cbind(alpha_beta_samples,theta_mat)
```

c

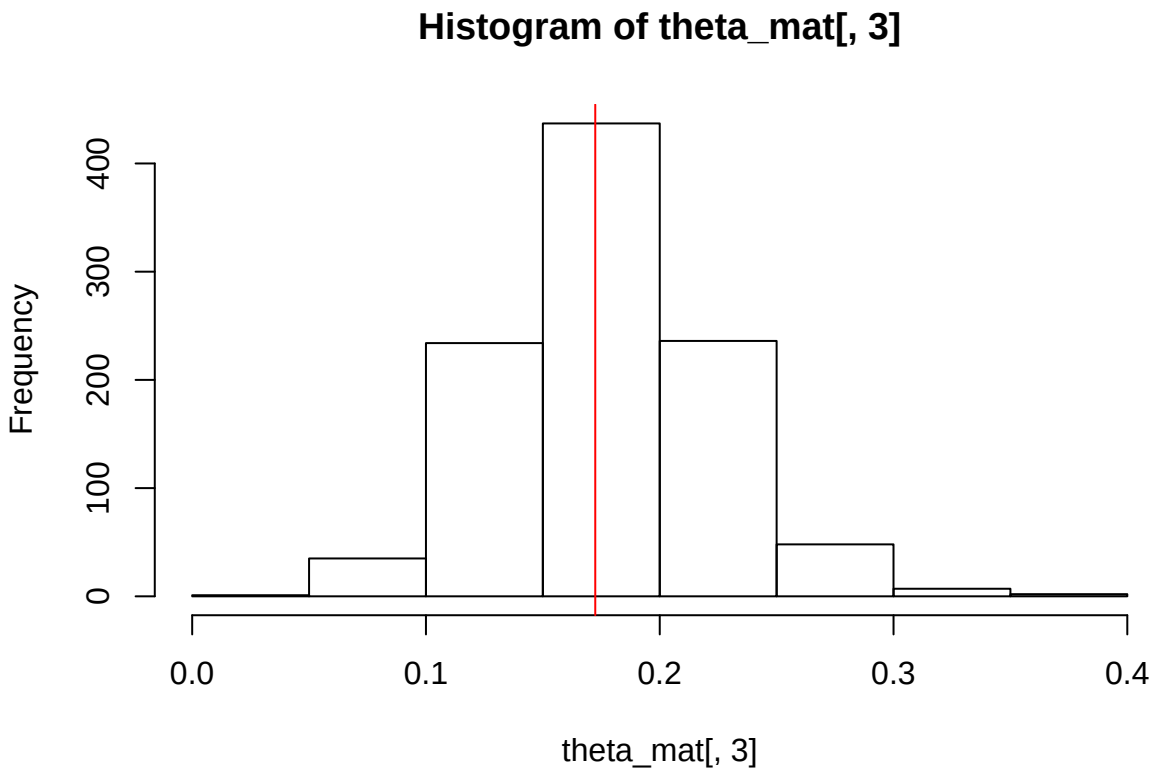
```
yvec = c(16/(58+16),9/(90+9),10/(48+10),13/(57+13),19/(103+19),20/(57+20),18/(86+18),17/(112+17),35/(273
#theta 1
plot.new()
hist(theta_mat[,1])
abline(v= yvec[1],col = "red")
```



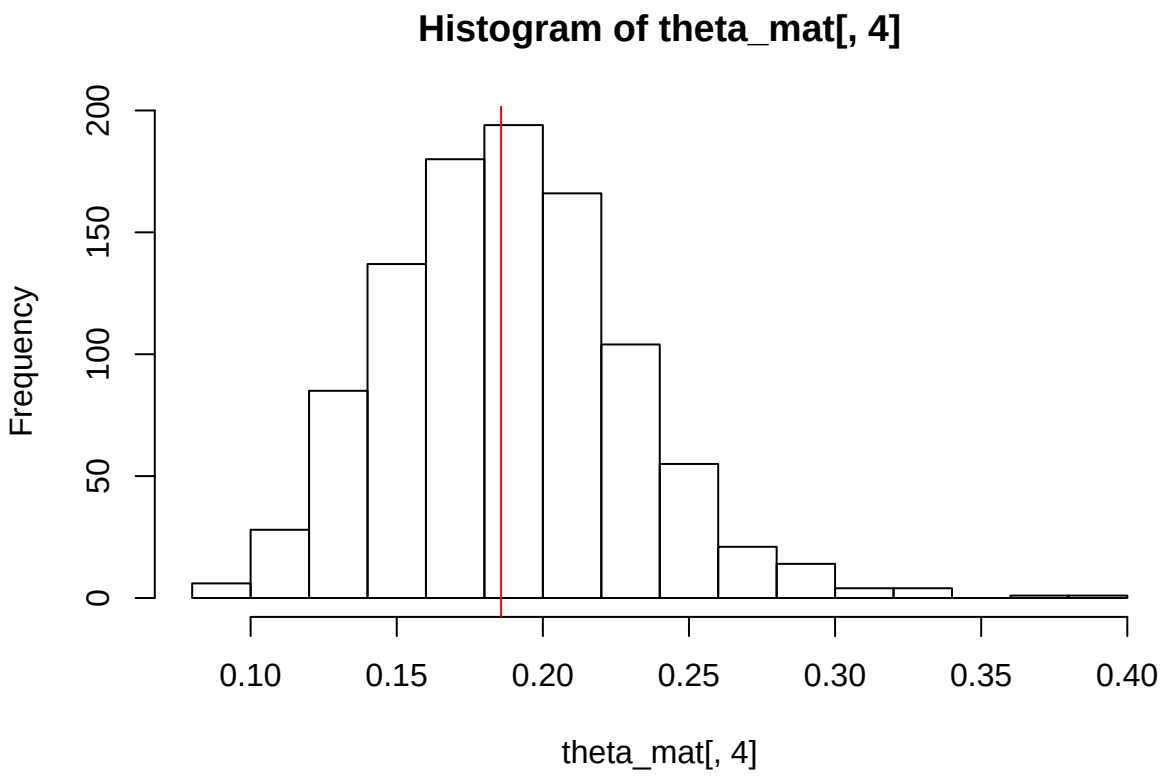

```
#theta 2
hist(theta_mat[,2])
abline(v = yvec[2],col = "red")
```



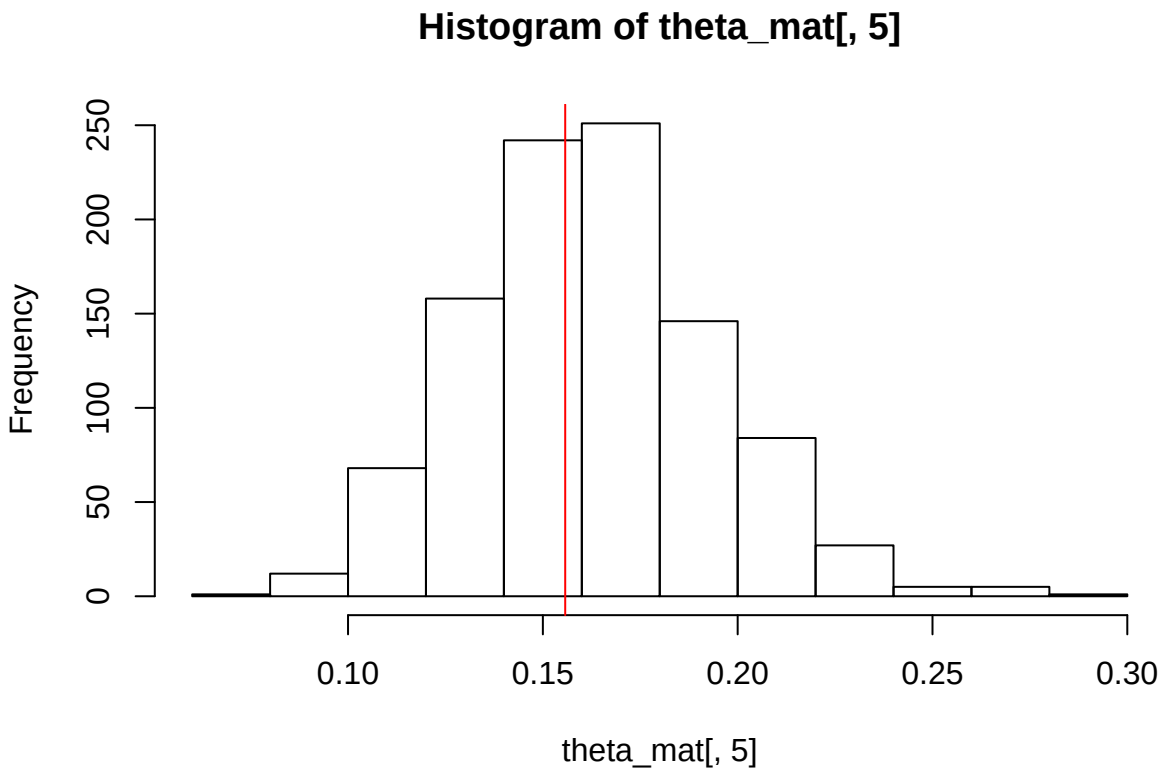

```
#theta 3
hist(theta_mat[,3])
abline(v = yvec[3],col = "red")
```



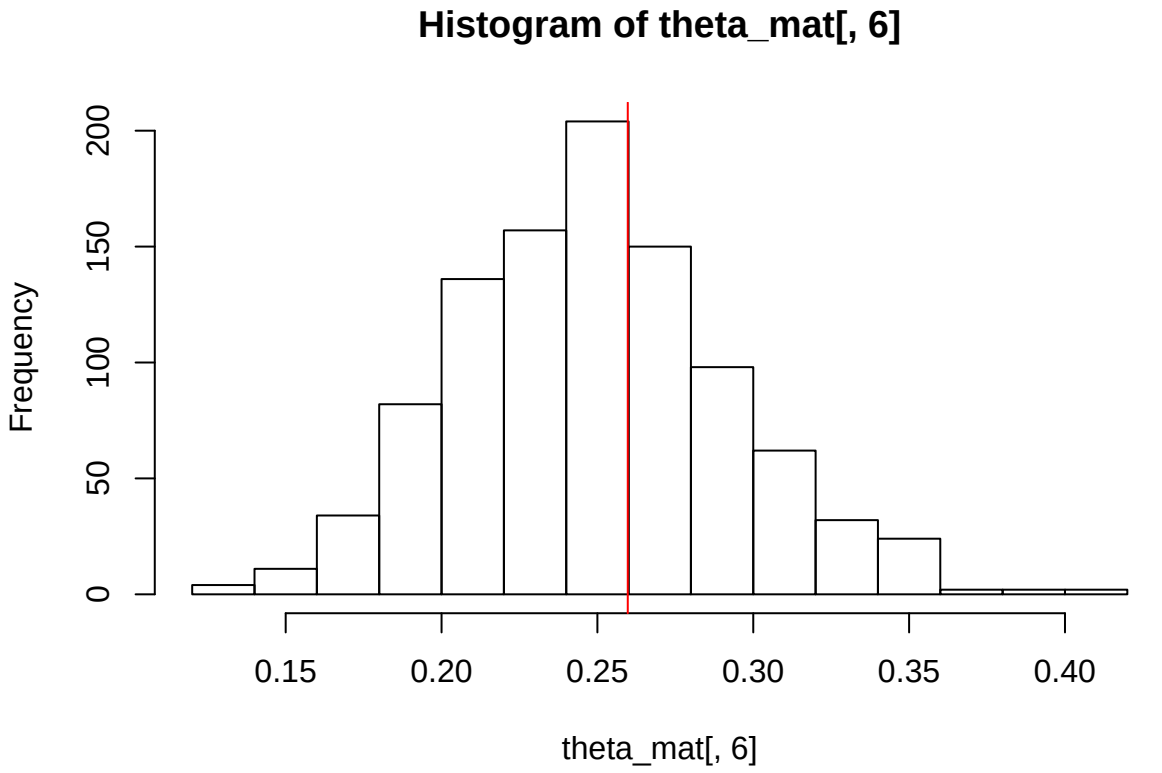
```r
#theta 4
hist(theta_mat[,4])
abline(v = yvec[4],col = "red")
```



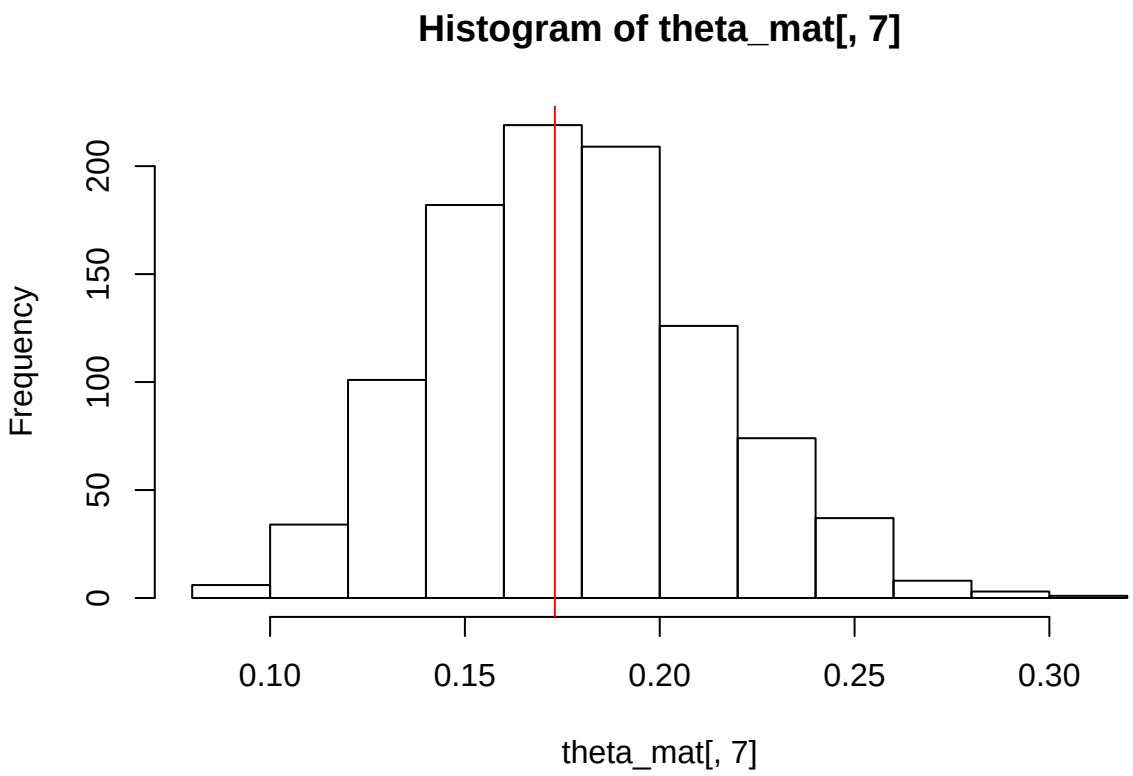

```r
#theta 5
hist(theta_mat[,5])
abline(v = yvec[5],col = "red")
```



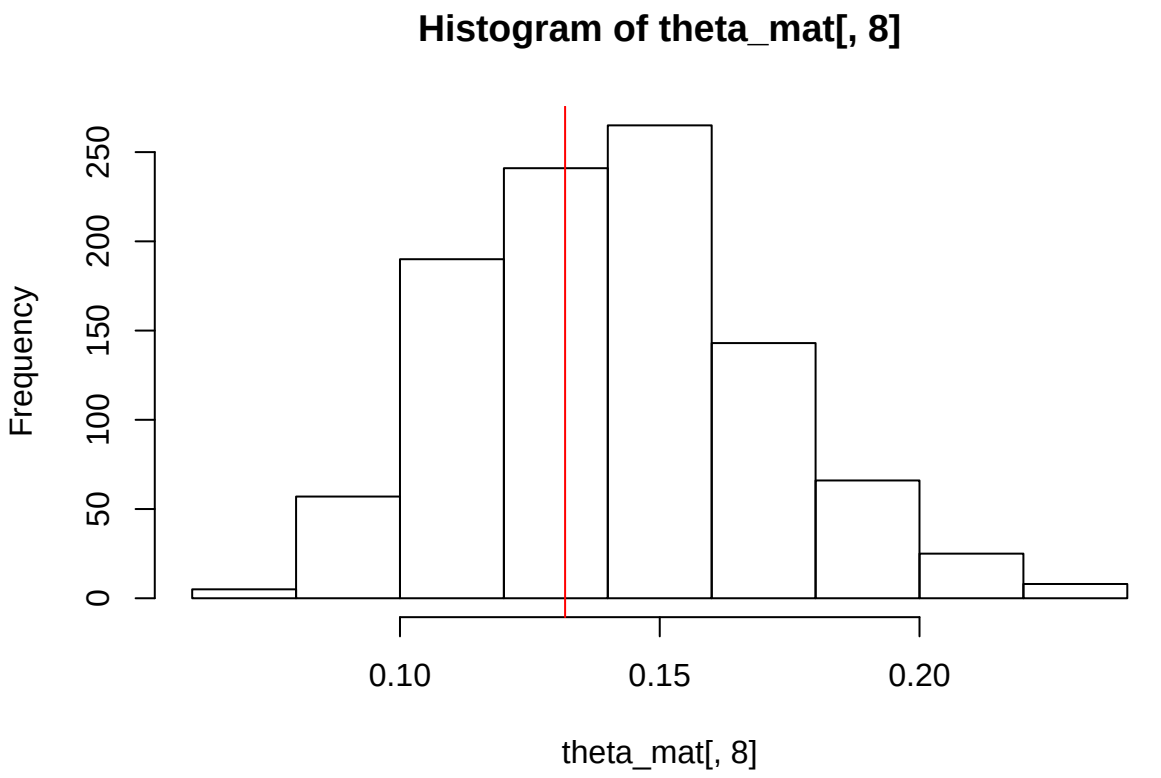

```
#theta 6
hist(theta_mat[,6])
abline(v = yvec[6],col = "red")
```



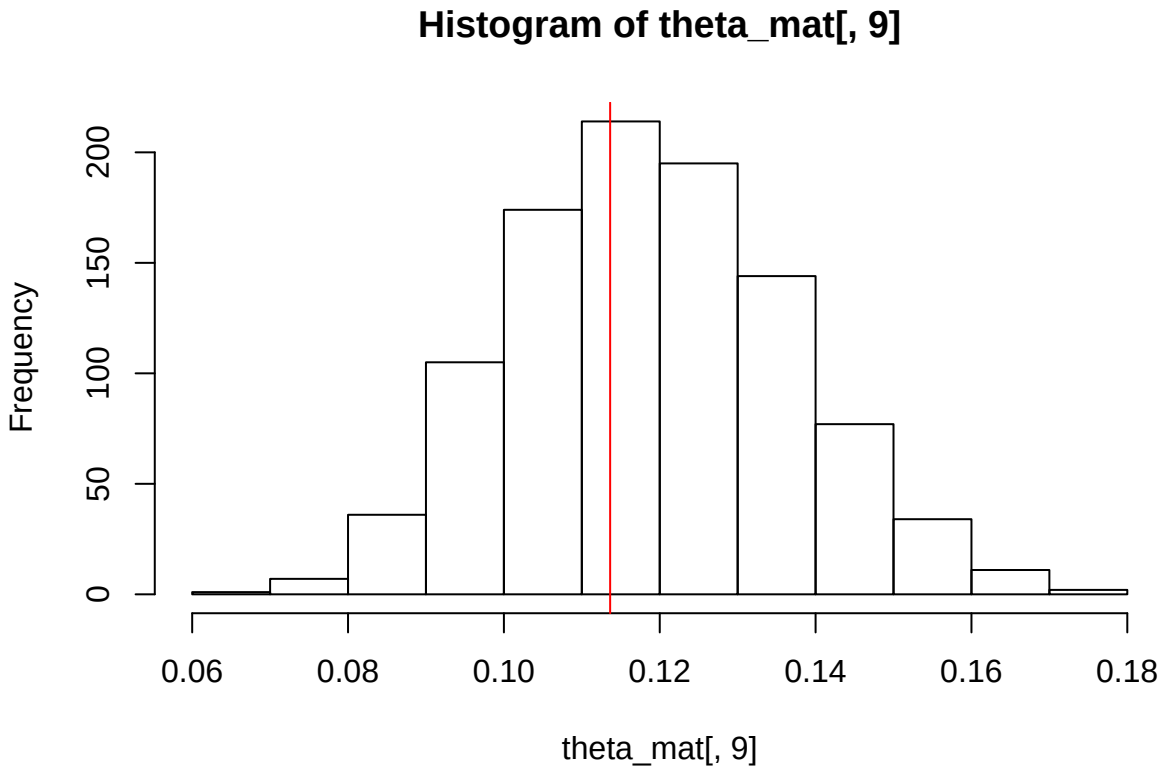
```r
#theta 7
hist(theta_mat[,7])
abline(v = yvec[7],col = "red")
```



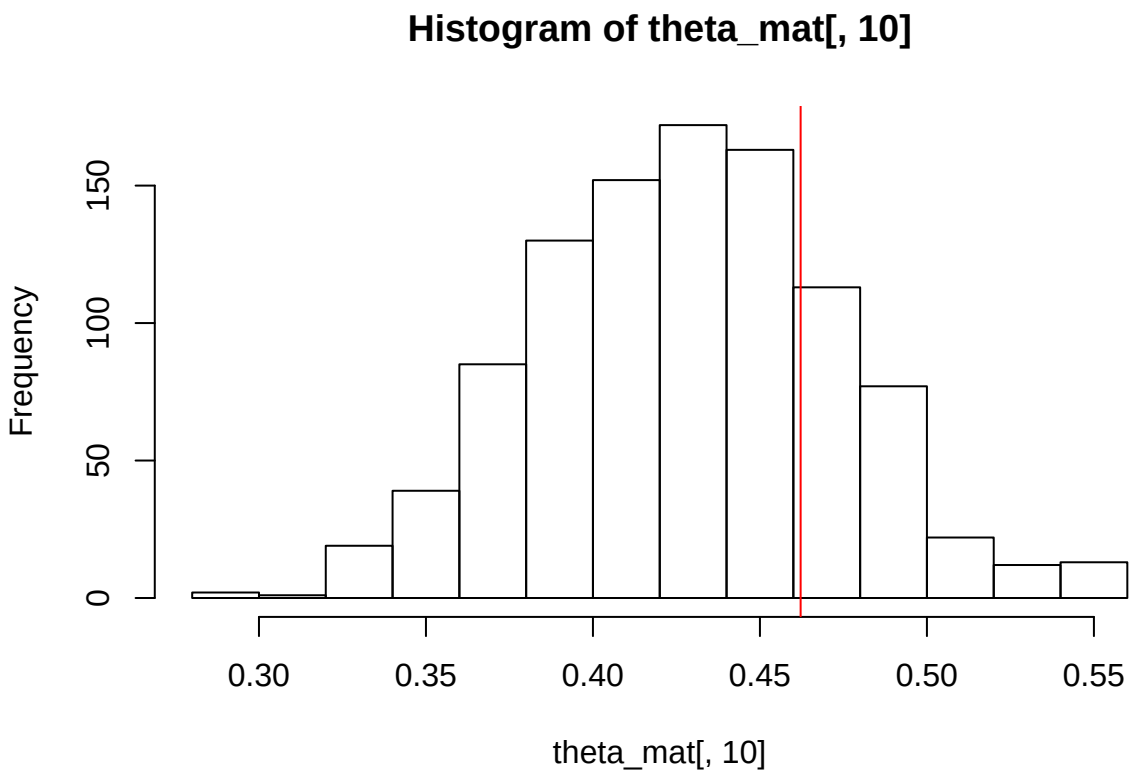

```r
#theta 8
hist(theta_mat[,8])
abline(v = yvec[8],col = "red")
```




```
#theta 9
hist(theta_mat[,9])
abline(v = yvec[9],col = "red")
```



```r
#theta 10
hist(theta_mat[,10])
abline(v = yvec[10],col = "red")
```



The observed thetas for each neighborhood are pretty close to the center of each histogram. The posterior inferences are similar to the raw values.

d

```r
avg_thetas = apply(theta_mat,1,mean)
quantile(avg_thetas,probs = c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.1732195 0.2207244
```

I'm not super confident about this strategy. I couldn't find an example for this task. The question asks for a posterior interval for the average theta, so I averaged each sample of thetas to create a vector of average posterior thetas. Then I simulated a 95% interval from those values.

e

following p118

```r
n_new = 100

#one posterior sample of alpha, beta
gridrow = sample(1:nrow(grid), size = 1, replace = T, prob=probs)
alpha_beta_sample = grid[gridrow,]
```

```r
#one prior draw of theta
theta_new = rbeta(1,alpha_beta_sample$x,alpha_beta_sample$y)

#predictive y draws
y_new = rbinom(10000, n_new, theta_new)


quantile(y_new,probs = c(0.025,0.975))
```

```
##  2.5% 97.5%
##    14    30
```

Above is my interval for bicycle counts. It is really variable because it depends on a single random draw of alpha and beta and then a single random draw of theta. I don't trust the interval because it is so variable. I've gotten intervals that don't even overlap, like (6,18) and (21,39).

Note: This problem was done incorrectly. I should have generated 10,000 draws of alpha, beta and theta to plug into rbinom to get 10,000 samples for y. The size in the gridrow sample and the number of draws from rbeta should be 10,000 instead of 1.

f

I think the beta distribution for the θ_j was a good choice. The histograms from part c of the posterior draws of the ten thetas seem to match the behavior of the observed thetas. The observed thetas were often close to the center of the histograms.