STATS 551 - HW2

Zhen Qin

1.

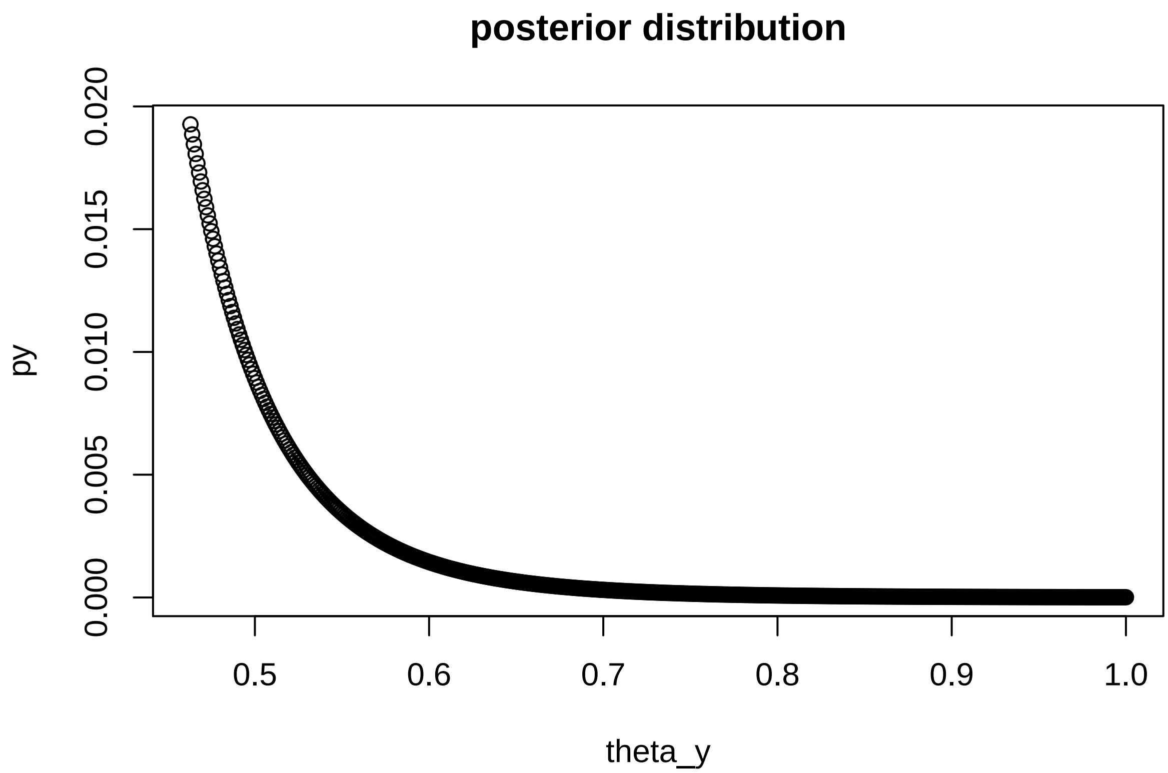
(a)y and z are the proportion so y and z are less than 1 and positive. The model is $y|\theta_y \sim Unif(0, \theta_y)$, $z|\theta_z \sim Unif(0, \theta_z)$. y_js are independent and identically distributed given parameters θ_y, z_js are independent and identically distributed given parameters θ_z.

(b)A prior distribution that is independent in θ_y, θ_z is a noninformative prior of uniform distribution, i.e. $\theta_y \sim Unif(0, 1)$, $\theta_z \sim Unif(0, 1)$.

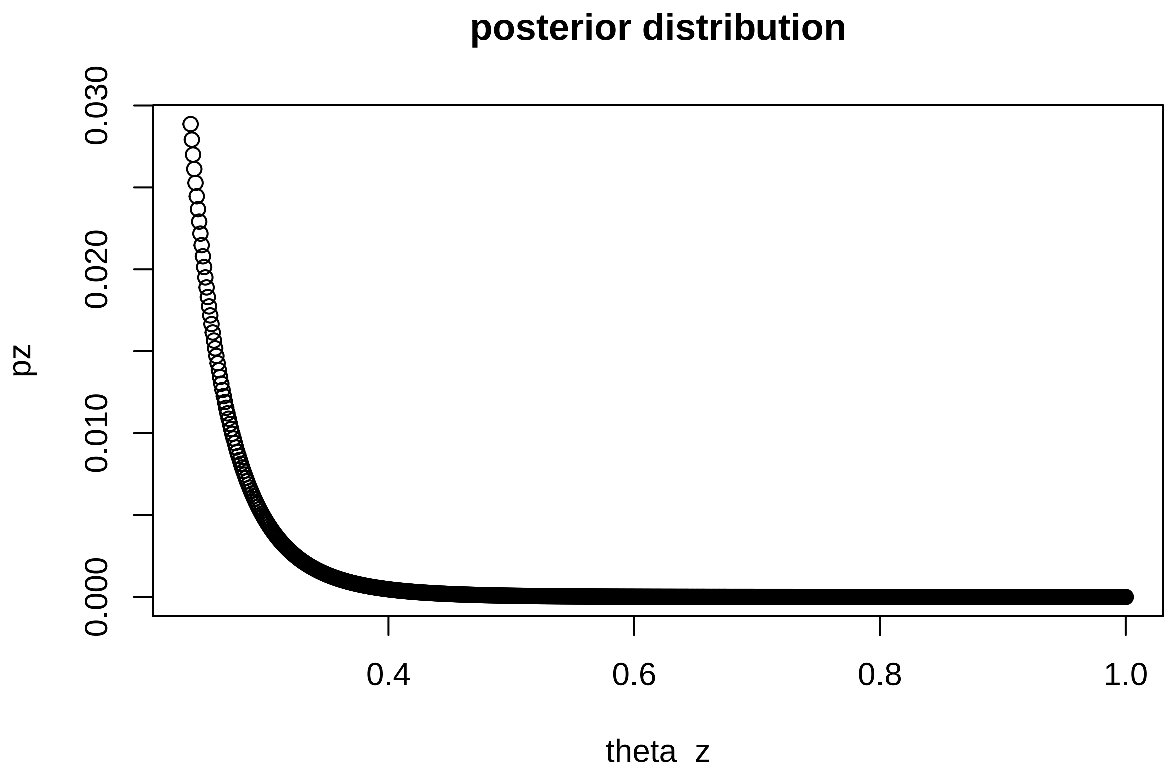
(c)After calculation, the posterier distribution is $p(\theta_y|y_1, ..., y_{10}) \propto \prod p(y_i|\theta_y)p(\theta_y) \propto \frac{1}{\theta_y^{10}} 1(1 > \theta_y > max(y_i) = 0.4621849)$, $p(\theta_z|z_1, ..., z_8) \propto \prod p(z_i|\theta_z)p(\theta_z) \propto \frac{1}{\theta_z^8} 1(1 > \theta_z > max(z_i) = 0.2368421)$.

(d)According to the uniform distribution, $\mu_y = E(y_i|\theta_y) = \theta_y/2$, $\mu_z = E(z_i|\theta_z) = \theta_z/2$

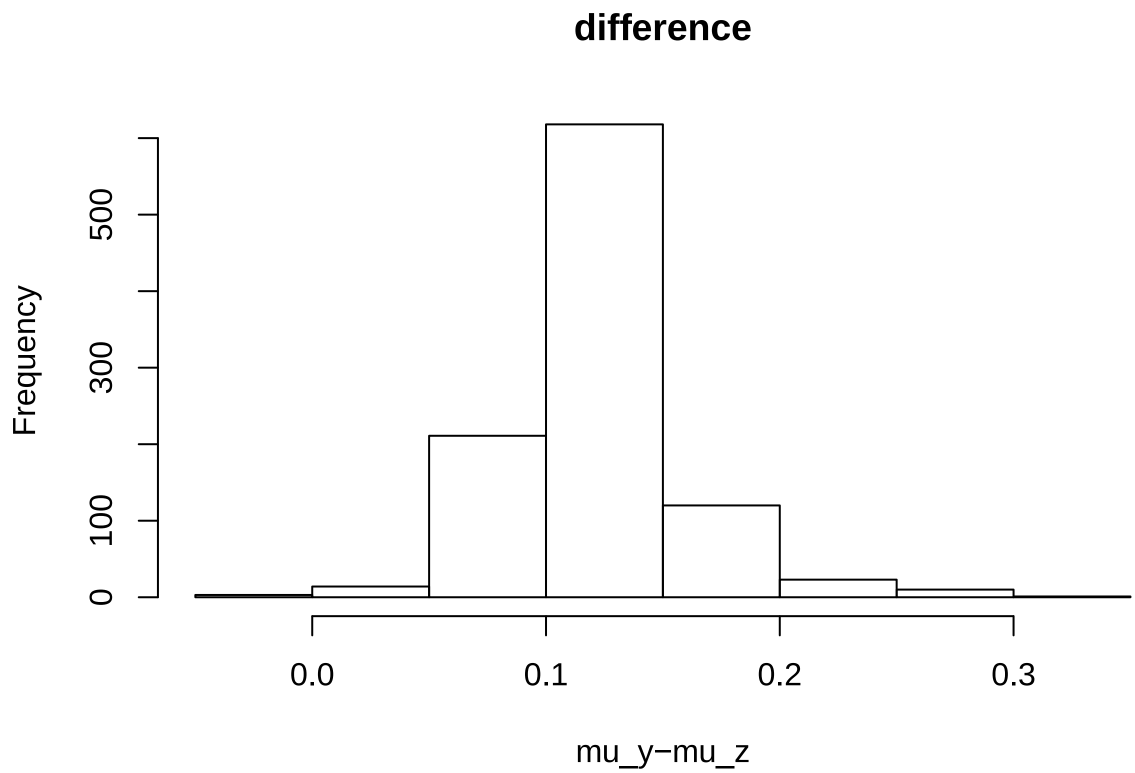
```r
library(ggplot2)
library(gridExtra)
library(tidyr)

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

znum=c(12,1,2,4,9,7,9,8)
zden=c(113,18,14,44,208,67,29,154)

y=ynum/(ynum+yden)
z=znum/(znum+zden)

inty=(max(y)^(-9)-1)/9
intz=(max(z)^(-7)-1)/7

ylim=(463:1000)/1000
zlim=(239:1000)/1000

py=1/ylim^10
pz=1/zlim^8

py=py/sum(py)
pz=pz/sum(pz)

plot(ylim,py,main = 'posterior distribution',xlab = 'theta_y')
```




```
plot(zlim,pz,main = 'posterior distribution',xlab = 'theta_z')
```




```
# simulation
ysim=sample(463:1000,1000,replace = T,prob = py)/1000
zsim=sample(239:1000,1000,replace = T,prob = pz)/1000

diffyz=(ysim-zsim)/2
hist(diffyz,xlab = 'mu_y-mu_z',main = 'difference')
```



2.

(a)

The model is $y_j|\theta_j \sim binomial(n_j, \theta_j)$. n_j is the total number of vehicles. $\theta_j \sim beta(\alpha, \beta)$. Suppose that (α, β) obey a noninformative hyperprior distribution, i.e. $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$. The joint posterior distribution is $p(\alpha, \beta, \theta_1, ..., \theta_n|y_1, ..., y_n) \propto (\alpha+\beta)^{-5/2}\prod_j \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} = (\alpha+\beta)^{-5/2}\prod_j \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$.

(b)

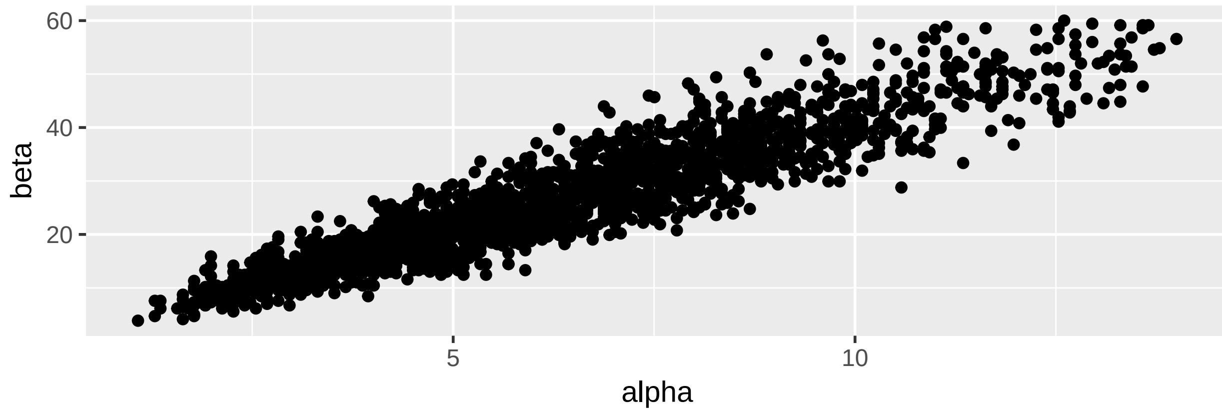
According to the beta distribution and integration of the density of the distribution, the marginal posterior density of the hyperparameters is $p(\alpha, \beta|obs) = \int p(\alpha, \beta, \theta_1, ..., \theta_{10}|y_1, ..., y_n)d\theta_1...d\theta_{10} \propto (\alpha+\beta)^{-5/2}\prod_j \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$.

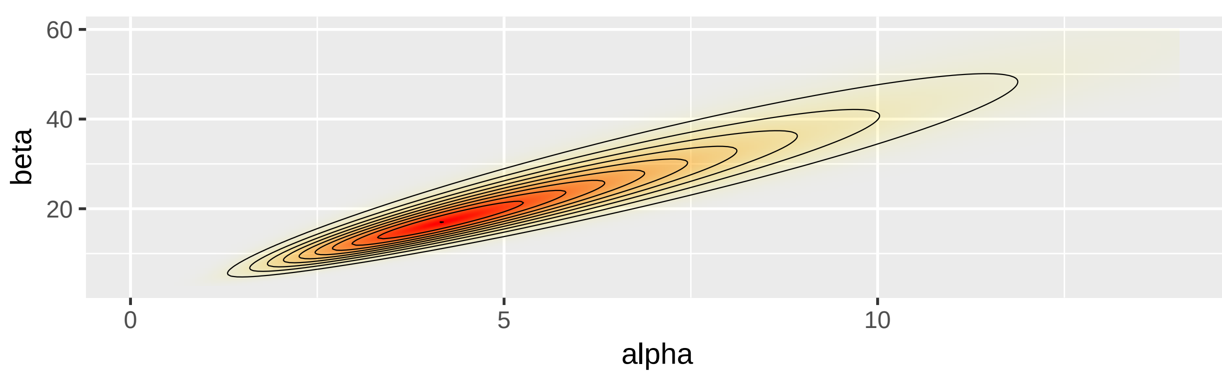
```
y=ynum
n=ynum+yden

x <- seq(0.0001, 0.9999, length.out = 1000)

bdens <- function(n, y, x)
  dbeta(x, y+1, n-y+1)

df_sep <- mapply(bdens, n, y, MoreArgs = list(x = x)) %>%
  as.data.frame() %>% cbind(x) %>% gather(ind, p, -x)
```

```r
labs1 <- paste('posterior of', c('theta_j', 'theta_71'))

A <- seq(0.1, 14, length.out = 200) ## alpha
B <- seq(3, 60, length.out = 200) ## beta
# make vectors that contain all pairwise combinations of A and B
cA <- rep(A, each = length(B))
cB <- rep(B, length(A))
# Use logarithms for numerical accuracy!
lpfun <- function(a, b, y, n)
  sum(lgamma(a+b)-lgamma(a)-lgamma(b)+lgamma(a+y)+lgamma(b+n-y)-lgamma(a+b+n))

lp <- mapply(lpfun, cA, cB, MoreArgs = list(y, n))
df_marg <- data.frame(x = cA, y = cB, p = exp(lp - max(lp)))

# Subtract maximum value to avoid over/underflow in exponentation
title1 <- 'Contour of likelihood for alpha beta'
# create a plot of the marginal posterior density
postdensityalphabeta = 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 = range(cA), ylim = range(cB)) +
  labs(x = 'alpha', y = 'beta', title = title1) +
  scale_fill_gradient(low = 'yellow', high = 'red', guide = F) +
  scale_alpha(range = c(0, 1), guide = F)
```

The following is the simulations from the joint posterior distribution of the parameters and hyperparameters: `samplestheta` and `samplesalphabeta`. The scatter plot and contour plot shows that the region is proper.

```r
nsamp <- 100
samp_indices <- sample(length(df_marg$p), size = nsamp,
                       replace = T, prob = df_marg$p/sum(df_marg$p))

samp_A <- cA[samp_indices[1:nsamp]]
samp_B <- cB[samp_indices[1:nsamp]]

nsamp2 <- 2000
samp_indices2 <- sample(length(df_marg$p), size = nsamp2,
                        replace = T, prob = df_marg$p/sum(df_marg$p))
samplesalphabeta = data.frame(alpha = cA[samp_indices2[1:nsamp2]],
                              beta = cB[samp_indices2[1:nsamp2]])
scatteralphabeta = ggplot(samplesalphabeta, aes(x=alpha, y=beta)) + geom_point()
grid.arrange(scatteralphabeta, postdensityalphabeta)
```





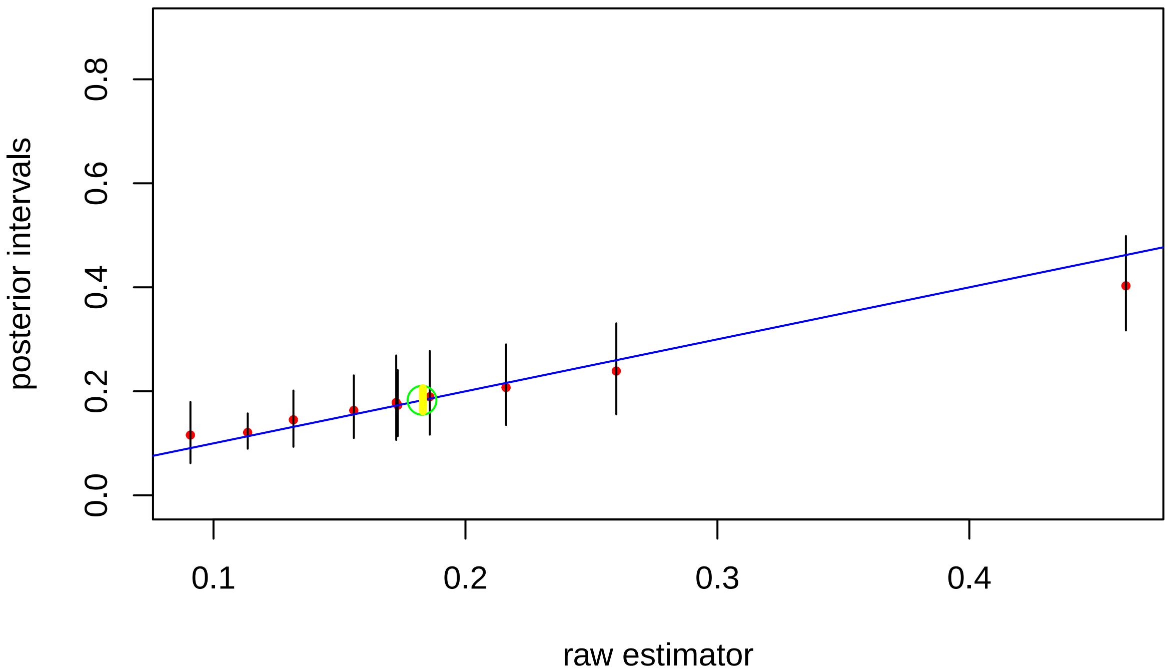

```
nsamp <- 1000
samp_indices <- sample(length(df_marg$p), size = nsamp,
                       replace = T, prob = df_marg$p/sum(df_marg$p))

samp_A <- cA[samp_indices[1:nsamp]]
samp_B <- cB[samp_indices[1:nsamp]]
samplestheta <- matrix(0, nsamp, length(y))
for(j in 1:length(y)){
  samplestheta[, j] = sapply(1:nsamp,
                             function(k) rbeta(1, samp_A[k]+y[j], samp_B[k]+n[j]-y[j]))
}
postintervals_sample <- apply(samplestheta, 2,
                              function(x) c(median(x), c(quantile(x, c(0.025, 0.975)))))
```

(c)

Compare the posterior distributions of the parameters to the raw proportions. The plot shows that raw proportions are near the median of the posterior distribution. Posterior intervals cover the blue line, which indicates that the model is good.

```
rawest <- jitter(y / n)
plot(rawest, postintervals_sample[1, ], pch = 19, col = 'red',
     cex = 0.5, ylim = c(-0.01, 0.9),
     ylab = 'posterior intervals', xlab = 'raw estimator')
for(k in 1:length(y)){
  lines(cbind(rep(rawest[k], 2), postintervals_sample[2:3, k]))
}
```

```
lines(seq(0, 0.9, by = 0.01), seq(0, 0.9, by = 0.01), col = 'blue')
phatpool = sum(y)/sum(n)
points(phatpool, phatpool, cex = 2, col = 'green')
lines(cbind(rep(qbeta(0.5, sum(y)+1, sum(n)-sum(y)+1), 2),
            c(qbeta(0.025, sum(y)+1, sum(n)-sum(y)+1),
              qbeta(0.975, sum(y)+1, sum(n)-sum(y)+1))), col = 'yellow', lwd = 4)
```



(d)

Drawing samples from posterior distribution, a 95% posterior interval for the average underlying proportion is as following.

```
quantile(rowMeans(samplestheta), c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.1705977 0.2189826
```

(e)

Drawing samples from posterior distribution, a 95% posterior interval for the number of those vehicles is as following. The chance of the real number is in the interval is 95%.

```
100*quantile(as.vector(samplestheta),c(0.025,0.975))
```

```
##      2.5%     97.5%
##  9.063624 43.494291
```

(f)

The beta distribution for the θ_j's is reasonable. First, the simulations of θ_j are meaningful because they are proportion in (0,1). Second, the beta distribution is a conjugate prior for binomial distribution, which means the hyperparameters can be intepreted as prior information. Third, the model is good because of plots above.