GRADUATE STUDENT STAT 840 A1

Vsevolod Ladtchenko 20895137

Problem 6

(a)

$$p(\theta|X,Y) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{\theta^{Y_i}X_i^{Y_i}e^{-\theta X_i}}{Y_i!}$$
$$\propto \theta^{\alpha-1+\sum_{i=1}^n Y_i}e^{-\theta(\beta+\sum_{i=1}^n X_i)}$$

We see this is the kernel of the Gamma distribution with parameters $\alpha + \sum_{i=1}^n Y_i$ and $\beta + \sum_{i=1}^n X_i$.

(b)

For noreact, the posterior will be Gamma(a1 + 2285, b1 + 1037) and for react will be Gamma(a2 + 256, b2 + 95)

```
react = read.table('/home/chad/Desktop/skool/840/cancer_react.txt',header=T)
noreact = read.table('/home/chad/Desktop/skool/840/cancer_noreact.txt',header=T)
react_sum_x = sum(react[1])
react_sum_y = sum(react[2])
noreact_sum_x = sum(noreact[1])
noreact_sum_y = sum(noreact[2])
c(react_sum_x, react_sum_y, noreact_sum_x, noreact_sum_y)
```

```
## [1]   95  256 1037 2285
```

(c)

Opinion 1

Since we started with the same prior, we see that the noreact data gave us a narrower posterior, while the smaller amount of react data gave us a wider posterior, indicating more uncertainty. Additionally, the centers of the distributions have gone separate ways suggesting the expected cancer rate for noreact is lower than the rate for react.

```
a1 = 220
a2 = 220
b1 = 100
b2 = 100

# bayes estimator for theta1
# posterior theta1 ~ Gam(a1 + noreact_sum_y, b1 + noreact_sum_x)
#   its mean is:
(a1 + noreact_sum_y) / (b1 + noreact_sum_x)
```

```
## [1] 2.203166
```

```r
# 95% credible interval for theta1
qgamma(.025, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.117726
```

```r
qgamma(.975, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.290273
```

```r
# posterior theta2 ~ Gam(a2 + react_sum_y, b2 + react_sum_x)
#   its mean is (a2 + react_sum_y) / (b2 + react_sum_x)
(a2 + react_sum_y) / (b2 + react_sum_x)
```

```
## [1] 2.441026
```

```r
# 95% credible interval for theta2
qgamma(.025, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 2.226633
```

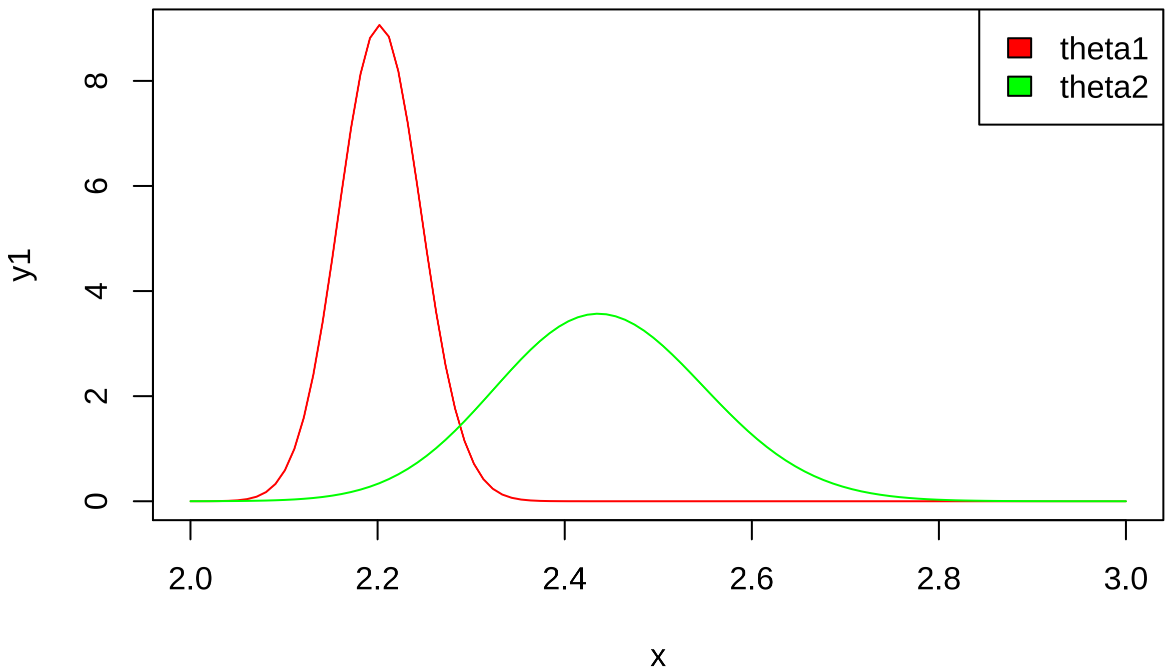
```r
qgamma(.975, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 2.665131
```

```r
# probability theta2 > theta1: generate pairs
n = 1000000
theta1 = rgamma(n, a1 + noreact_sum_y, b1 + noreact_sum_x)
theta2 = rgamma(n, a2 + react_sum_y, b2 + react_sum_x)
sum(theta2 > theta1) / n
```

```
## [1] 0.978194
```

```r
x = seq(from = 2, to = 3, length.out = 100)
y1 = dgamma(x, a1 + noreact_sum_y, b1 + noreact_sum_x)
y2 = dgamma(x, a2 + react_sum_y, b2 + react_sum_x)
plot(x, y1, type='l', col='red', ylim=c(0,9))
lines(x, y2, col='green')
legend(x = 'topright',
       legend=c('theta1', 'theta2'),
       fill = c('red','green'))
```



Opinion 2

The posterior for theta1 is the same. the posterior for theta2 is now wider, with a higher center. For theta2, we have in effect removed the influence of our prior beliefs about noreact towns. This way, react towns are more free to be influenced by the data. In opinion 1 we included our beliefs about noreact towns together with the react data which has influenced the posterior to be closer to the noreact posterior. In opinion 2, the react posterior is farther away and wider.

```
a1 = 220
b1 = 100
a2 = 2.2
b2 = 1

# bayes estimator for theta1
# posterior theta1 ~ Gam(a1 + noreact_sum_y, b1 + noreact_sum_x)
#   its mean is:
(a1 + noreact_sum_y) / (b1 + noreact_sum_x)
```

```
## [1] 2.203166
```

```
# 95% credible interval for theta1
qgamma(.025, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.117726
```

```
qgamma(.975, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.290273
```

```r
# posterior theta2 ~ Gam(a2 + react_sum_y, b2 + react_sum_x)
#   its mean is (a2 + react_sum_y) / (b2 + react_sum_x)
(a2 + react_sum_y) / (b2 + react_sum_x)
```

```
## [1] 2.689583
```

```r
# 95% credible interval for theta2
qgamma(.025, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 2.371497
```

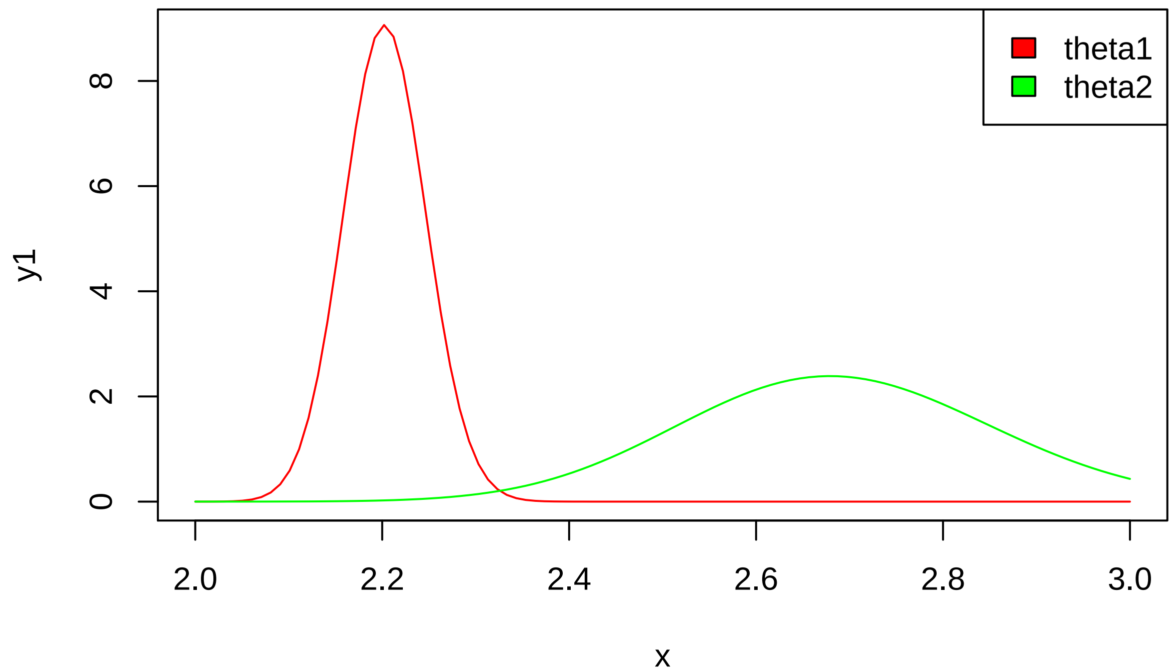
```r
qgamma(.975, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 3.027397
```

```r
# probability theta2 > theta1: generate pairs
n = 1000000
theta1 = rgamma(n, a1 + noreact_sum_y, b1 + noreact_sum_x)
theta2 = rgamma(n, a2 + react_sum_y, b2 + react_sum_x)
sum(theta2 > theta1) / n
```

```
## [1] 0.99842
```

```r
x = seq(from = 2, to = 3, length.out = 100)
y1 = dgamma(x, a1 + noreact_sum_y, b1 + noreact_sum_x)
y2 = dgamma(x, a2 + react_sum_y, b2 + react_sum_x)
plot(x, y1, type='l', col='red', ylim=c(0,9))
lines(x, y2, col='green')
legend(x = 'topright',
       legend=c('theta1', 'theta2'),
       fill = c('red','green'))
```



Opinion 3

Now we relaxed the prior for noreact. there is barely any difference since there is a lot more data for noreact, especially considering how much of a difference there was in making the same change for react data in opinions 1 and 2. This demonstrates that the prior has less influence as we gather more data. so even with a poor choice of prior, eventually the data corrects it. On the other hand, the prior has a lot of power when we have little data, as is seen in the react posterior.

```
a1 = 2.2
a2 = 2.2
b1 = 1
b2 = 1

# bayes estimator for theta1
# posterior theta1 ~ Gam(a1 + noreact_sum_y, b1 + noreact_sum_x)
#   its mean is:
(a1 + noreact_sum_y) / (b1 + noreact_sum_x)
```

```
## [1] 2.203468
```

```
# 95% credible interval for theta1
qgamma(.025, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.114081
```

```
qgamma(.975, a1 + noreact_sum_y, b1 + noreact_sum_x)
```

```
## [1] 2.29468
```

```r
# posterior theta2 ~ Gam(a2 + react_sum_y, b2 + react_sum_x)
#   its mean is (a2 + react_sum_y) / (b2 + react_sum_x)
(a2 + react_sum_y) / (b2 + react_sum_x)
```

```
## [1] 2.689583
```

```r
# 95% credible interval for theta2
qgamma(.025, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 2.371497
```

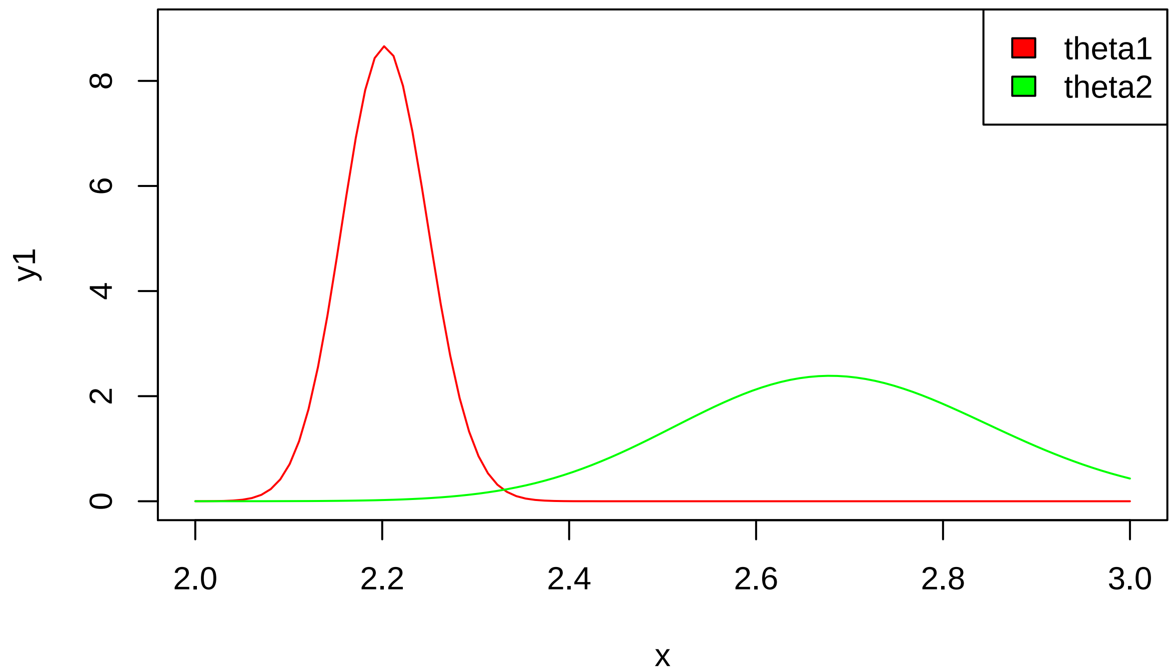
```r
qgamma(.975, a2 + react_sum_y, b2 + react_sum_x)
```

```
## [1] 3.027397
```

```r
# probability theta2 > theta1: generate pairs
n = 1000000
theta1 = rgamma(n, a1 + noreact_sum_y, b1 + noreact_sum_x)
theta2 = rgamma(n, a2 + react_sum_y, b2 + react_sum_x)
sum(theta2 > theta1) / n
```

```
## [1] 0.998405
```

```r
x = seq(from = 2, to = 3, length.out = 100)
y1 = dgamma(x, a1 + noreact_sum_y, b1 + noreact_sum_x)
y2 = dgamma(x, a2 + react_sum_y, b2 + react_sum_x)
plot(x, y1, type='l', col='red', ylim=c(0,9))
lines(x, y2, col='green')
legend(x = 'topright',
       legend=c('theta1', 'theta2'),
       fill = c('red','green'))
```



(d)

We can think of each person as a small-county, and their cancer rate is theta. So the rate for the entire county will be theta*Xi. This is because a sum of Poisson variables is Poisson. So unless the population size has a direct affect on reactor construction, it is reasonable to assume population size gives no information about the cancer rate. If this were not reasoanble, and we had to assume a relationship between the population size and the cancer rate, then each county would have its own cancer rate theta_i. we would need to perform separate analyses for each county, and there would be no common theta we could infer.

(e)

We encoded our a priori beliefs about theta1 and theta2 as two different distributions, parametrized by (a1,b1,a2,b2). They are independent because they share no parameters or information.

Why: Suppose we know about some underlying structure of two parameters. For example we have a prior distribution for the radius of a circle, and a prior for the circumference. we know the circumference is $2pi$r so its prior should simply be a scaling of the radius's prior.

How: One way to accomplish this is to do a transformation of the distribution of the first prior.

Another reason is: if we somehow know the cancer rate for react is twice that of noreact, we can express that in the prior's mean, because that will allow us to make a stronger prior considering we have less data. So if we have a "strong" belief about theta1, this will give us a strong belief about theta2 even when we have less data for theta2. once again, this requires knowing the underlying relationship between the rates.