# STAT 574B Homework 1

Nathan Hattersley

September 13, 2017

## Problem 1

### part a

In this part we want an exact 95% confidence interval on the trial of 25 patients that did not have Lyme disease. The R code shown below extracts the "confidence interval" property of the "binom.test" object to give us a 95% confidence interval of $\theta \in (0.549, 0.906)$:

```
binom.test(19, 25)$conf.int

## [1] 0.5487120 0.9064356
## attr(,"conf.level")
## [1] 0.95
```

Compared with the large-sample confidence interval $\theta \in (0.593, 0.927)$, this interval is wider, expressing greater uncertainty about the true value of $\theta$. This is likely because the sample is so small (25) that the large-sample's z-score used to scale the sample variance about the mean, 1.96, is not a totally accurate representation of the $\alpha/2$ tail probability on either side.

### part b

This is the same analysis of part a, except now we are considering the hypothethical wherein 21/29 patients have received the true negative diagnosis. This translates to a confidence interval of $\theta \in (0.528, 0.873)$:

```
binom.test(21, 29)$conf.int

## [1] 0.5276155 0.8726599
## attr(,"conf.level")
## [1] 0.95
```

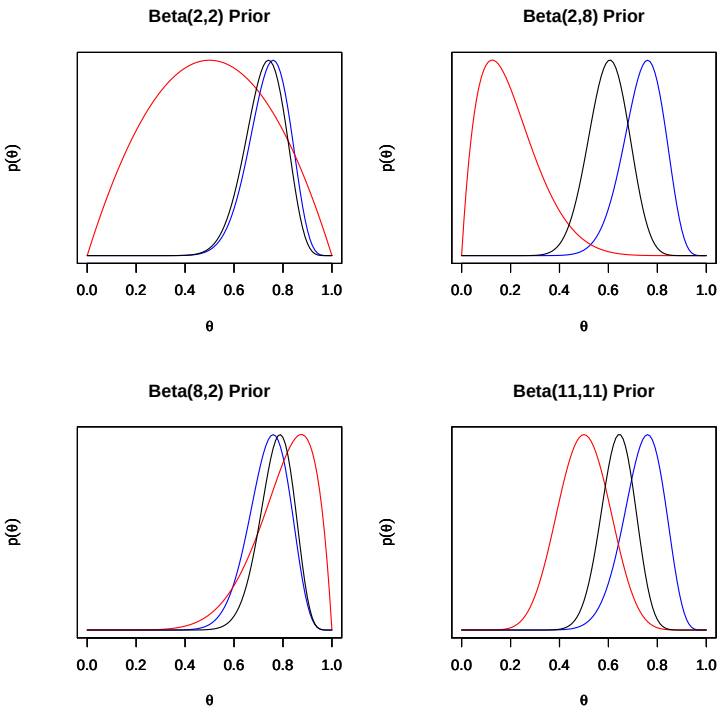
This interval is slightly tighter than that calculated in part a, and its midpoint is slightly lower. The constriction of the interval arises from the increased sample size and hence lower sample variance. The midpoint moved because, marginally, we are tacking on 4 trials at a 50% success rate, a rate lower than our observed sample rate. This would plainly lower the sample proportion of successes.

### part c

We know that the mean of the Beta distribution is $\frac{\alpha}{\alpha+\beta}$, so for priors (1),(2), and (4), whose expected values are below the sample proportion of 0.76, we expect the posterior expected value of $\theta|Y_1, \ldots, Y_{25}$ to be below the sample proportion of successes, $\bar{Y}$. For (3), with a prior expected value of 0.8, we expect the opposite, that the posterior expectation will be above 0.76. The lower the variance of the prior distribution, the weaker the data's effect on the posterior will be, so for prior (4), the posterior expectation will be closer to 0.5 than

will prior (1), although they have the same prior expectation. I've only included the code to produce one graph, as the rest are trivially similar.
**Note:** I graphed the likelihood in blue, prior in red, and posterior in black for each example. I also did not pay attention to y-axis scale so that the densities all looked "pretty."

```
par(mfrow = c(2, 2))
likelihood <- function(s) {
    return(s^19 * (1 - s)^6)
}
## Plot the likelihood in blue
curve(likelihood(x), from = 0, to = 1, col = "blue", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)
## Plot the prior in red
curve(dbeta(x, 2, 2), from = 0, to = 1, col = "red", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)
## Plot the posterior in black
curve(dbeta(x, 21, 8), from = 0, to = 1, ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
title("Beta(2,2) Prior")
```




**part d**

We know the formula for the posterior of a Beta given a binomial sampling of $y$ successes in $n$ trials,

$$\theta|Y_1, \ldots, Y_{25} \sim Beta(\alpha + y, \beta + n - y)$$

Hence we calculate that our particular posterior is a $Beta(21, 8)$ distribution. The mean and variance of $\theta \sim Beta(\alpha, \beta)$ are

$$E(\theta|\alpha, \beta) = \frac{\alpha}{\alpha + \beta}$$
$$V(\theta|\alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

For this example we have

$$E(\theta|y_1, \ldots, y_{25}) = 0.724,$$
$$SD(\theta|y_1, \ldots, y_{25}) = \sqrt{V(\theta|y_1, \ldots, y_{25})} = 0.0816$$

A 95% HPD confidence interval is $\theta \in (0.563, 0.877)$, while a quantile-based 95% interval is $\theta \in (0.551, 0.868)$.

```
## HPD Confidence Interval
hpd(qbeta, shape1 = 21, shape2 = 8)

## [1] 0.5629623 0.8767368

## Quantile-based Confidence Interval
qbeta(c(0.025, 0.975), shape1 = 21, shape2 = 8)

## [1] 0.5512845 0.8677635
```

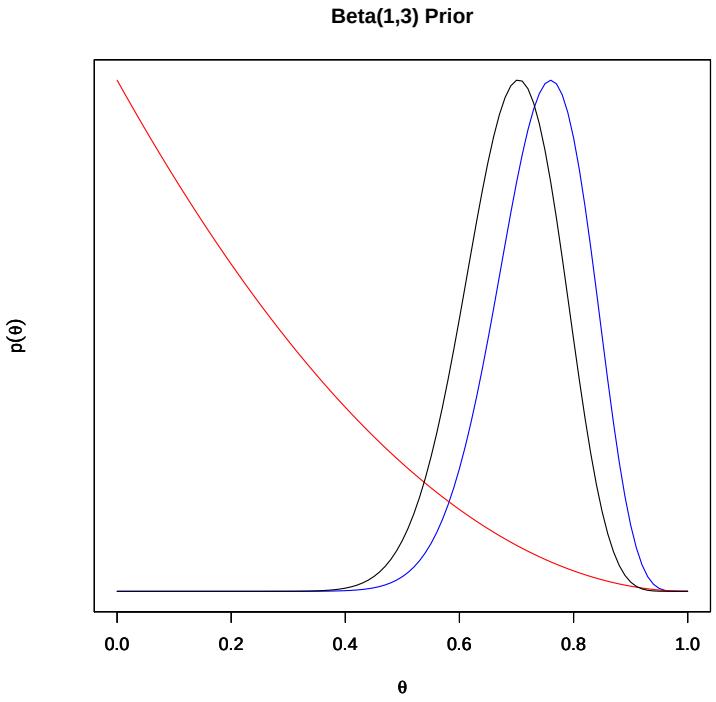
The Bayesian quantile-based confidence interval looks approximately closest to the Agresti-Coull confidence interval. (I did a lot of scrolling up and down and eye-balling, so I wouldn't express 95% confidence in my judgment.) However, it makes intuitive sense that a prior with mean 0.5 would yield a posterior whose confidence intervals are closer to a sample that had four trials at a marginal success rate of 0.5 tacked on. It seems Agresti and Coull inflected a frequentist "a prior-i" prior of some sort on the data.

**part e**

At first blush I considered specifying a prior that matched the historical specificity of the HIV diagnostic measure mentioned in the article. However, I figured that it would be more appropriate to conduct a hypothesis test to see if a posterior confidence band would contain the specificity, .9999, of the HIV test. Hence I found a little difficult the specification of an "informed" prior. I also chose not to use the "idiot's" reference prior of $Beta(1, 1)$ since I was feeling experimental. So I went with a prior whose expected value was less than 0.5, as I wanted to be *really convinced* that the diagnostic was a success. I went with a $Beta(1, 3)$, which gave the pretty figures $E(\theta) = .25$, $V(\theta) = .0375$. Below, I plot the likelihood, prior, and posterior, then I calculate the median and 95% quantile-based confidence interval on the posterior.

```
## Plot the likelihood in blue
curve(likelihood(x), from = 0, to = 1, col = "blue", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)
## Plot the prior in red
curve(dbeta(x, 1, 3), from = 0, to = 1, col = "red", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)
```

```r
## Plot the posterior in black
curve(dbeta(x, 20, 9), from = 0, to = 1, ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
title("Beta(1,3) Prior")
```




```r
## alpha/2 tail, median, and (1-alpha/2) tail, respectively
qbeta(c(0.025, 0.5, 0.975), 20, 9)

## [1] 0.5133317 0.6940678 0.8412240
```

This posterior confidence interval is wider than the quantile-based interval from a $Beta(2, 2)$ prior, and also contains smaller values (both tails of the former are less than the corresponding tails of the latter). It also does not contain the value $\theta = .9$, which is rather condemning in the sense that our specificity is not very high compared to the historical norm.

**part f**

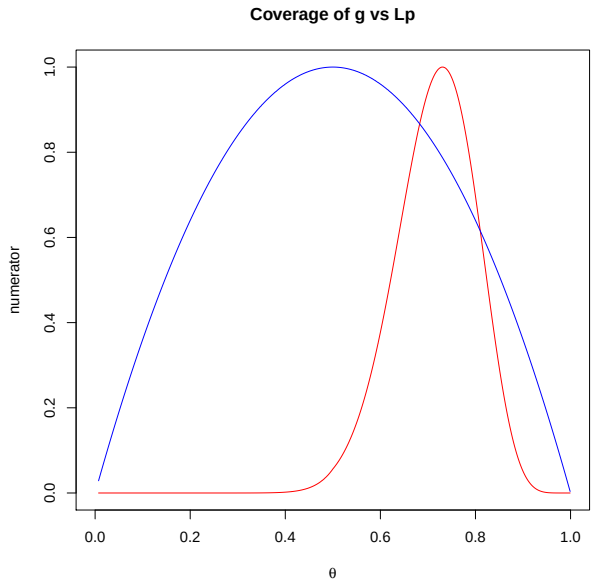
For a nice coverage of the support of possible theta values, I wanted a sample function $g$ that had relatively high variance. Also, to suppress possible inflation of weight values, I wanted a $g$ that somewhat closely approximated the shape of the numerator. It took some toying around, but I settled on the $Beta(2, 2)$, given its high variance and slightly convex shape. I took 10,000 samples from this Beta density and checked the ratio,

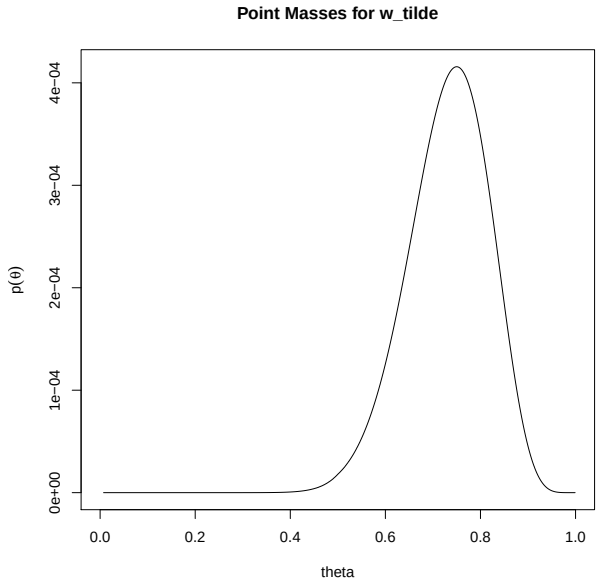
$$\frac{max_\theta \, \tilde{w}(\theta)}{mean_\theta \, \tilde{w}(\theta)} = 4.157$$

and it was not far off from the example Dr. Bedrick gave in class. I also checked the highest "point mass" of the $\tilde{w}(\theta)$,

$$\frac{max_\theta\, \tilde{w}(\theta)}{\sum_{\theta \in \Theta} \tilde{w}(\theta)} = 4.157 \times 10^{-4}$$

and it is absurdly small, giving us reason not to worry about resampling one theta value repeatedly. (It is also 1/10,000th of the max/mean ratio, which I should have figured before doing this analysis.) Below are plots of the numerator (red) and denominator (blue) of the $\tilde{w}(\theta)$ function, which confirm that our $\tilde{w}(\theta)$ values will not be too inflated. I also plot the point mass of $\tilde{w}(\theta)$ in the name of being ever more illustrative:

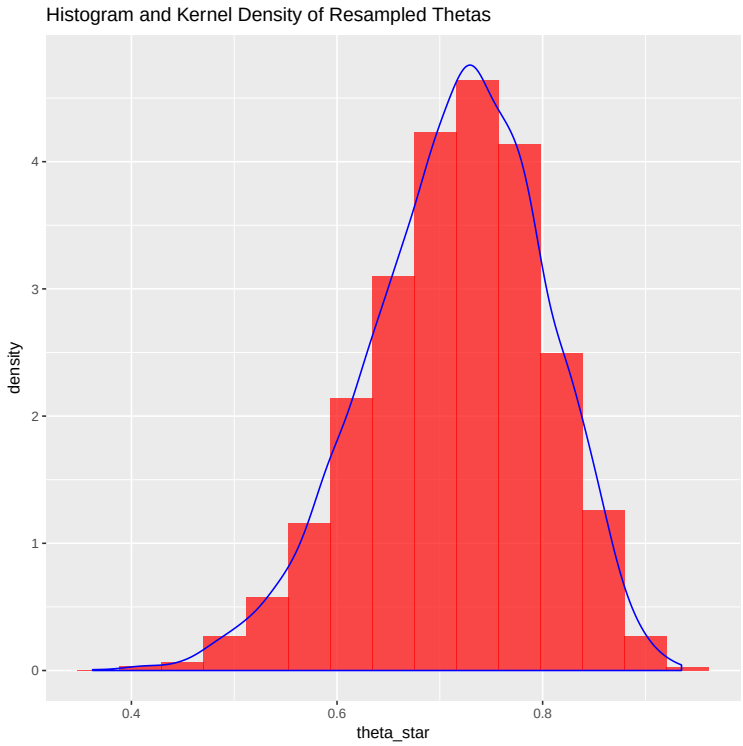


Now that we feel sufficiently good about our point masses that we may begin to resample. I took 5,000 of these and computed the summary stats as one would a simple Monte Carlo sample. I found:

| | |
|---|---|
| mean | 0.715 |
| standard deviation | 0.086 |
| 95% confidence interval | 0.533, 0.864 |

I used ggplot for the histogram and kernel density (since it's pretty and more functional). I will work on transitioning to using more ggplot when I can.




**Code:**

```
## Normalizing likelihood function like in class
L <- function(s) {
    ifelse(s > 0 & s < 1, suppressWarnings(exp(19 * log(s/0.76) +
        6 * log((1 - s)/0.24))), 0)
}
## Triangle prior and plot from the lecture notes
p <- function(s) {
    ifelse(s > 0 & s < 1, ifelse(s > 0.5, 4 * (1 - s), 4 * s),
        0)
}

## Sampling functions (to play with)
g <- function(s) {
    dbeta(s, 2, 2)
```

```
}
sample_g <- function(num) {
    sort(rbeta(num, 2, 2))
}



## Start sample, and normalize weight function
theta <- sample_g(10000)
numerator <- (L(theta) * p(theta))/max(L(theta) * p(theta))
denominator <- (g(theta)/max(g(theta)))
w_tilde <- numerator/denominator


## DIAGNOSTICS: Check graph of num+den, w_tildes for extreme
## values
w_ratio <- max(w_tilde)/mean(w_tilde)
## Numerator in red, denom in blue
plot(theta, numerator, xlim = c(0, 1), xlab = expression(theta),
    main = "Coverage of g vs Lp", type = "l", col = "red")
lines(theta, denominator, xlim = c(0, 1), type = "l", col = "blue")
## Let's normalize so we can treat y-axis values as
## probabilities
plot(theta, w_tilde/sum(w_tilde), type = "l", xlim = c(0, 1),
    ylab = expression(p(theta)), main = "Point Masses for w_tilde")



## Resample using the w_tilde as weights.  Docs for 'sample'
## fxn say we don't have to normalize w_tilde
theta_star <- sample(theta, 5000, replace = T, prob = w_tilde)

## Display the desired summary stats
stats <- list(mean = mean(theta_star), standard_dev = sd(theta_star),
    conf_int = quantile(theta_star, c(0.025, 0.975)))

## Histogram and density function
ggplot(as.data.frame(theta_star), aes(theta_star)) + geom_histogram(aes(y = ..density..),
    bins = 15, fill = "red", alpha = 0.7) + geom_density(color = "blue") +
    labs(title = "Histogram and Kernel Density of Resampled Thetas")
```

## Problem 2

### part a

To specify a prior, I wanted a Gamma distribution whose mean was twenty. Also, I decided "surprise" meant $Pr(\theta > 32|\alpha, \beta) \leq 0.01$. To decide on the Gamma parameters, I made a helper function to vectorize the search process.

```
ddgamma <- function(a) {
    cbind(a, a/20, qgamma(0.99, a, a/20))
}
ddgamma(1:20)
```

Using this function, I found that the first alpha value for which $Pr(\theta > 32) \leq 0.01$ is $\alpha = 20$. Hence I used a Gamma(20,1) prior distribution.

### part b

Using the interpretation of $\alpha$ counts in $\beta$ time periods, this prior gives the weight of seeing 20 fatal accidents in one year. Since the data show 238 fatal accidents in 10 years, our posterior distribution is Gamma(258,11).

### part c

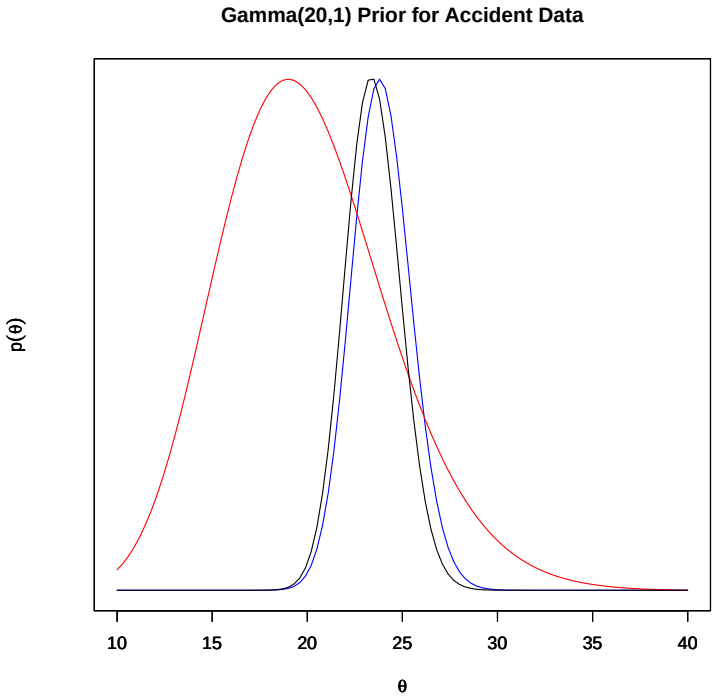
Just like the other multi-plots, I plot the likelihood in blue, the prior in red, and the posterior in black.

```
## Kernel of the likelihood
likelihood <- function(t) {
    exp(-10 * t + 238 * log(t))
}

## Plot the likelihood in blue
curve(likelihood(x), from = 10, to = 40, col = "blue", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)

## Plot the prior in red
curve(dgamma(x, 20, 1), from = 10, to = 40, col = "red", ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
par(new = T)

## Plot the posterior in black
curve(dgamma(x, 258, 11), from = 10, to = 40, ylab = expression(p(theta)),
    xlab = expression(theta), yaxt = "n")
title("Gamma(20,1) Prior for Accident Data")
```




This prior has a relatively small weight in comparison to the data's effect on the posterior. The heuristic would be giving the prior one year's weight, while the data have ten years' weight. On the graphs, it certainly seems like the posterior more closely mirrors the likelihood function than it does the prior.

**part d**

Using the example from the BSM book, we see the predictive distribution for $\tilde{Y}$ is nbinom(size=258,mu=258/11). A 95% predictive interval for this distribution is

```
qnbinom(c(0.025, 0.975), size = 258, mu = 258/11)

## [1] 14 34
```

Now, we want to simulate the predictive distribution. That is, we sample random $\theta$ values from the posterior Gamma distribution, then compute 95% confidence intervals on the resultant Poisson distribution. Using the Monte Carlo method, we can simply take the mean of the 2.5% and 97.5% quantiles to estimate the predictive 95% confidence interval. Here I take 1,000 samples.

```
samp <- rgamma(1000, 258, 11)
c(mean(qpois(0.025, samp)), mean(qpois(0.975, samp)))

## [1] 14.501 33.419
```

There is a slight discrepancy in the upper bound: After multiple trials, I found that every time, our simulated upper bound was closer to 33 than 34. I figured that this has to do with the discrete nature of the negative binomial distribution.

**part e**

If we change the distribution of $Y_i$ to be Poisson($\theta t_i$), then our likelihood function changes:

$$L(\theta|y_1, \dots, y_n, t_1, \dots, t_n) = e^{-\theta \sum_i t_i} \theta^{\sum_i y_i} \prod_i (t_i^{y_i}/y_i!)$$

Discarding everything that doesn't pertain to $\theta$, let's examine the kernel of $L(\theta)p(\theta)$, with p being the prior and $\theta \sim Gamma(\alpha, \beta)$:

$$L(\theta) \propto e^{-\theta \sum_i t_i} \theta^{\sum_i y_i}$$
$$\implies L(\theta)p(\theta) \propto e^{-\theta(\beta + \sum_i t_i)} \theta^{\alpha + \sum_i y_i - 1}$$

Any Bayesian worth his/her salt would recognize that the above is a kernel of a Gamma($\alpha + \sum_i y_i$, $\beta + \sum_i t_i$) random variable. We end our proof there because there is no reason to preoccupy oneself with the nonsense of deriving normalization constants. ■

**part f**

Since we have changed the problem definition a little, it now follows that $\theta$ is more of a rate parameter. If $Y_i$ is the number of fatal accidents and $t_i$ is how many hundred million passenger miles are flown in a given year, then $\theta$ is now approximating how many fatal accidents occur per one hundred million miles flown. In order to come up with a reasonable estimate for $\theta$, I took the expert information from part a) and tried to backtrack. First I modified my dataset:

```
fatalities.df <- within(fatalities.df, {
    miles <- fatalities/rate
})
```

Then I used the mean number of miles flown to estimate the rate of fatal accidents which would produce 20 and 32 accidents in a year:

```
## Mean of our distribution
20/mean(fatalities.df$miles)

## [1] 0.00349903

## 99th percentile of our distribution
32/mean(fatalities.df$miles)

## [1] 0.005598448
```

I then recreated the test function like I did in part a), which fixed the mean and calculated the 99th percentile for a vector of numbers. I'm including a slice of interest, too.

```
ddgamma <- function(s) {
    cbind(s, s/0.00349903, qgamma(0.99, s, s/0.00349903))
}
ddgamma(10:50)
```

The first alpha value for which our 99th percentile was less than the target value was 20, so I chose my prior to be Gamma(20, 5715.87). For perspective, a 95% confidence interval for $\theta$ under this distribution is (2.137e-03, 5.191e-03). Now I'll calculate a 95% confidence interval on the posterior distribution. Magrittr used for a nicer code presentation:

```r
## Alpha and Beta prior parameters
a <- 20
b <- 5715.87

## sum of y and sum of t variables
sy <- sum(fatalities.df$accidents)
st <- sum(fatalities.df$miles)

## Ceci n'est pas une pipe, mais c'est quelque chose que rend
## plus simple le R.
qgamma(c(0.025, 0.975), a + sy, b + st) %>% format(digits = 4,
    scientific = T)

## [1] "3.618e-03" "4.619e-03"
```

Now, since we showed now nice the Monte Carlo method works in part d), I'm going to employ it here. On our mileage scale, $8 \times 10^{11}$ is 8,000 hundreds of millions of miles, so we need to remember to multiply our theta estimates by 8,000 to get the Poisson parameter. I took 10,000 samples for good coverage. I took the mean Poisson parameter as well as the 95% quantile-based interval:

```r
samp <- rgamma(10000, a + sy, b + st)
(out <- c(mean(qpois(0.025, 8000 * samp)), mean(8000 * samp),
    mean(qpois(0.975, 8000 * samp))))

## [1] 22.10510 32.83012 44.50210
```

To conclude, in a year in which passengers travel a total of $8 \times 10^{11}$ miles, we expect there to be about 33 fatal accidents in a year, with 22 to 45 a not totally unreasonable range.