STAT 8700 Homework 3

Brian Detweiler

Friday, September 16, 2016

1. Suppose we have a population described by a Normal Distribution with known variance $\sigma^2 = 1600$ and unknown mean μ. 4 observations are collected from the population and the corresponding values were: 940, 1040, 910, and 990.

```r
y.bar <- mean(940, 1040, 910, 990)
y.bar
```

```
## [1] 940
```

(a) If we choose to use a Normal(1000, 200^2) prior for θ, find the posterior distribution for θ by hand.

First, we'll derive the posterior for the single data point case, then for the general case.

Likelihood for a single data point

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$$

Normal Prior, s.t. $\theta \sim N(\mu_0, \tau_0^2)$

$$p(\theta) \propto e^{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}}$$

Posterior for single observation

$$p(\theta) \propto e^{\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)}$$

$$\theta|y \sim N(\mu_1, \tau_1^2), \text{ s.t. } \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \text{ and} \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Now we are set up to extend this model to multiple observations. We will assume these four observations are *i.i.d.*, such that $y = (y_1, y_2, y_3, y_4)$.

Posterior density for multiple observations

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) \\
&= p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\
&\propto e^{\left(-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right)} \prod_{i=1}^{n} e^{\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right)} \\
&\propto e^{\left(-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\right)}
\end{aligned}
$$

After simplifying algebraicly, we find that the posterior depends only on y by the sample mean, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, which means $\overline{y}$ is a sufficient statistic. Now, since $\overline{y}|\theta, \sigma^2$, we can treat $\overline{y}$ as a single observation and we get

$$
p(\theta|y_1, y_2, y_3, y_4) = p(\theta|\overline{y}) = N(\theta|\mu_n, \tau_n^2), \text{ where } \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \text{ and } \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}
$$

Substituting in our values, we have

$$
\begin{aligned}
n &= 4 \\
\boxed{\overline{y} = 940} \\
\mu &= \theta \\
\sigma^2 &= 1600 \\
\tau_0^2 &= 200^2 \\
\frac{1}{\tau_4^2} &= \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \\
&= \frac{1}{200^2} + \frac{4}{1600} \\
&= \frac{1}{200^2} + \frac{1}{400} \\
&= 0.002525 \\
\mu_0 &= 1000 \\
\mu_4 &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \\
&= \frac{\frac{1}{200^2}1000 + \frac{4}{1600}940}{\frac{1}{200^2} + \frac{4}{1600}} \\
&= \frac{\frac{1}{40} + 2.35}{\frac{1}{400}} \\
&= \boxed{950} \\
p(\theta|y_1, y_2, y_3, y_4) &= p(\theta|\overline{y}) = N(\theta|\mu_4, \tau_4^2) \\
&= N(\theta|950, 396.03960396)
\end{aligned}
$$

■

(b) Find, by hand, a 95% credible interval for θ.

A 95% CI for θ is given by evaluating $p(y|\theta)$ at $y = 0.025$ and $y = 0.975$, with $\nu = 4$ degrees of freedom.

$$\begin{aligned}
p(0.025;\theta) &= \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} y^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{1}{2y}}, y > 0 \\
&= \frac{1}{2^{\frac{4}{2}}\Gamma(\frac{4}{2})} (0.025)^{-\left(\frac{4}{2}+1\right)} e^{-\frac{1}{2(0.025)}} \\
&= \frac{1}{4}(0.025)^{-3} e^{-\frac{1}{0.05}} \\
&\approx 0.000032978457959 \\
p(0.975;\theta) &= \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} y^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{1}{2y}}, y > 0 \\
&= \frac{1}{2^{\frac{4}{2}}\Gamma(\frac{4}{2})} (0.975)^{-\left(\frac{4}{2}+1\right)} e^{-\frac{1}{2(0.975)}} \\
&= \frac{1}{4} 1.07891232152 e^{-\frac{1}{1.95}} \\
&\approx 0.161514323478
\end{aligned}$$

This gives us a 95% Credible Interval of (0.000032978457959, 0.161514323478).

∎

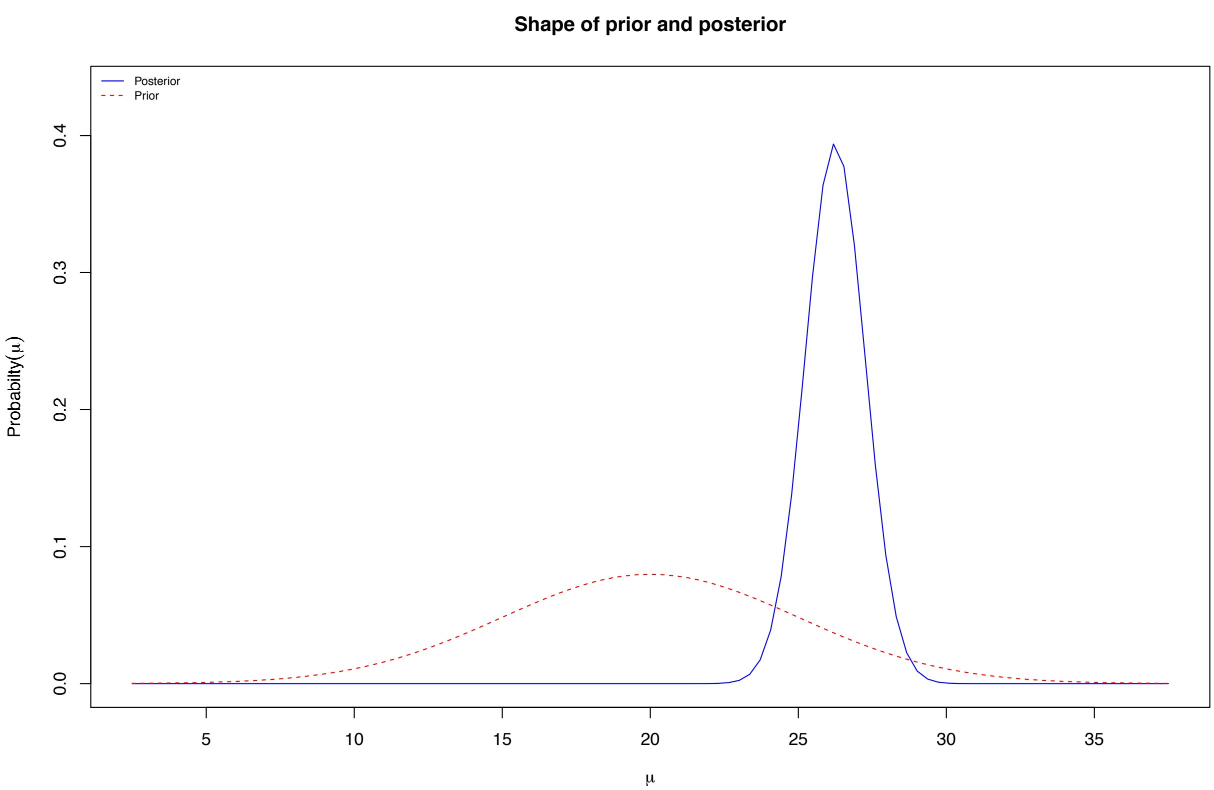
2. The `normnp` function in the `Bolstad` package computes the posterior for the mean with a Normal prior. The function requires 4 inputs (in order): a vector containing the data, the prior mean, the prior standard deviation, and the population standard deviation. Suppose we consider a Normal population with a variance of 16, and we collect 15 observations from this population with values: 26.8, 26.3, 28.3, 28.5, 26.3, 31.9, 28.5, 27.2, 20.9, 27.5, 28.0, 18.6, 22.3, 25.0, 31.5.

```
library(Bolstad)
var <- 16
obs <- c(26.8, 26.3, 28.3, 28.5, 26.3, 31.9, 28.5, 27.2, 20.9, 27.5, 28.0, 18.6, 22.3, 25.0, 31.5)
pop.st.dev <- sqrt(16)
```

(a) If we choose a $Normal(20, 25)$ prior, Use R to find the posterior distribution for the population mean.

```
prior.mu <- 20
prior.st.dev <- sqrt(25)

posterior <- normnp(obs, prior.mu, prior.st.dev, pop.st.dev)
```

```
## Known standard deviation :4
## Posterior mean           : 26.2404092
## Posterior std. deviation : 1.0114435
```



```
##
## Prob.    Quantile
## ------   ----------
## 0.005    23.6351035
## 0.010    23.8874398
## 0.025    24.2580164
## 0.050    24.5767327
## 0.500    26.2404092
## 0.950    27.9040857
## 0.975    28.2228020
## 0.990    28.5933786
## 0.995    28.8457149
```

(b) What are the posterior mean and variance?

The posterior mean is 26.2404092, and variance is 1.0230179.

(c) Find a 95% credible interval for the population mean.

A 95% credible interval for the population mean is found at the 0.025 and 0.975 quantiles, (24.2580164, 28.222802).

■

3. Suppose $y|\theta \sim Poisson(\theta)$, find the Jeffreys' prior density for θ. Find α and β for which the $Gamma(\alpha, \beta)$ density is a close match to the Jeffreys' prior.

Jeffrey's prior is given by $J(\theta) = \sqrt{I(\theta)}$, where $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln p(y|\theta)\right]$.

The Poisson distribution we are interested in, is $p(y_n|\theta) = \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \prod_{i=1}^n \frac{1}{y_i!}$.

So working through this by parts, we start with the natural log,

$$\begin{aligned}
\ln \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \prod_{i=1}^n \frac{1}{y!} &= \ln \frac{1}{y!} - \theta + y \ln \theta \\
&= \sum_{i=1}^n y_i \ln \theta - n\theta - \ln \sum_{i=1}^n y_i!
\end{aligned}$$

Taking the first derivative with respect to θ, we get

$$\begin{aligned}
\frac{\partial}{\partial\theta} \ln p(y_n|\theta) &= \frac{\partial}{\partial\theta} \sum_{i=1}^n y_i \ln \theta - n\theta - \ln \sum_{i=1}^n y_i! \\
&= \sum_{i=1}^n \frac{y_i}{\theta} - n - 0
\end{aligned}$$

Taking the second derivative with respect to θ, we get

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2} \ln p(y_n|\theta) &= \frac{\partial}{\partial\theta} \sum_{i=1}^n \frac{y_i}{\theta} \\
&= -\sum_{i=1}^n y_i \frac{1}{\theta^2}
\end{aligned}$$

Taking expectations,

$$\begin{aligned}
-E\left[-\frac{y}{\theta^2} \middle| \theta \right] &= \frac{n\theta}{\theta^2} \\
&= \frac{n}{\theta}
\end{aligned}$$

Finally, taking the square root to get the Jeffrey's prior, $J(I)$, we have

$$\begin{aligned}
\sqrt{I(\theta)} &= \sqrt{\frac{n}{\theta}} \\
&\propto \sqrt{\frac{1}{\theta}} \\
&= \boxed{\theta^{\frac{1}{2}}}
\end{aligned}$$

This comes closest to $\lim_{\beta \to 0} Gamma(\frac{1}{2}, \boxed{\beta})$, though it is not a proper distribution.

■

4. Suppose we have multiple independent observations $y_1, y_2, \ldots, y_n$ from a $Poisson(\theta)$ distribution.

(a) Consider the conjugate Gamma prior. What values of the hyperparameters would lead to a flat (improper) prior distribution for θ?

With a Gamma prior, we have

$$p(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}$$

So to get a flat prior out of this, we need the hyperparameters that result in $p(\theta) \propto 1$, so we have

$$\begin{aligned}
p(\theta) &\propto e^{-\beta\theta}\theta^{\alpha-1} \\
&= e^{-0\theta}\theta^{1-1} \\
&= e^{0}\theta^{0} \\
&\propto 1 \\
\theta &\sim Gamma(\alpha = 1, \beta = 0)
\end{aligned}$$

■

(b) Using a general $Gamma(\alpha, \beta)$ prior, derive the posterior distribution for θ. What is the required sufficient statistic needed from the data?

$$\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&\propto e^{-n\theta}\theta^{\sum y_i}e^{-\beta\theta}\theta^{\alpha-1} \\
&= e^{-[\theta(n+\beta)]}\theta^{\sum y_i}\theta^{\alpha-1} \\
&= e^{-\theta(n+\beta)}\theta^{n\overline{y}+\alpha-1}
\end{aligned}$$

So we have $Gamma(\alpha + n\overline{y}, \beta + n)$.

Thus, $n\overline{y}$ is sufficient because it is free of θ.

5. Derive the gamma posterior distribution (equation 2.15) for the Poisson model parameterized in terms of rate and exposure with conjugate prior distribution.

$$
\begin{aligned}
p(\theta|y) &\propto p(y|\theta)p(\theta) \\
&\propto \left[\theta^{\left(\sum_{i=1}^{n} y_i\right)} \boxed{e^{-(x_i)\theta}}\right] \cdot \left[\frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\right] \\
&\propto \left[\theta^{\left(\sum_{i=1}^{n} y_i\right)} e^{-(x_i)\theta}\right] \cdot \left[\theta^{\alpha-1} e^{-\beta\theta}\right] \\
&= \theta^{\left(\alpha+\sum_{i=1}^{n} y_i - 1\right)} e^{-\left(\beta+\sum_{i=1}^{n} x_i\right)\theta}
\end{aligned}
$$

And thus we have the posterior as $\theta|y \sim Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i)$.

■

6. The table at the end of the assignment gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten year period from 1976 to 1985.

```r
years <- c(1976:1985)
fatal.accidents <- c(24, 25, 31, 31, 22, 21, 26, 20, 16, 22)
passenger.deaths <- c(734, 516, 754, 877, 814, 362, 764, 809, 223, 1066)
death.rate <- c(0.19, 0.12, 0.15, 0.16, 0.14, 0.06, 0.13, 0.13, 0.03, 0.15)

airline.deaths <- as.data.frame(cbind(years, fatal.accidents, passenger.deaths, death.rate))

airline.deaths
```

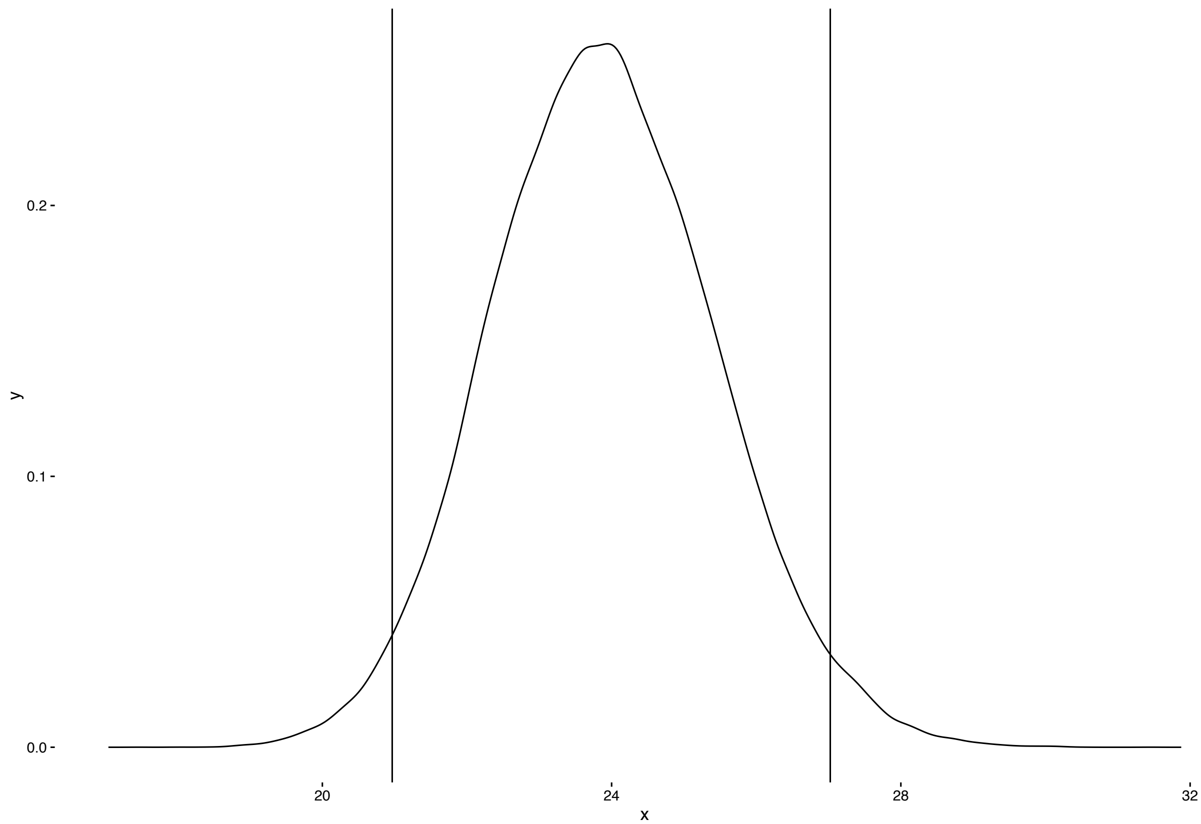
```
##    years fatal.accidents passenger.deaths death.rate
## 1   1976              24              734       0.19
## 2   1977              25              516       0.12
## 3   1978              31              754       0.15
## 4   1979              31              877       0.16
## 5   1980              22              814       0.14
## 6   1981              21              362       0.06
## 7   1982              26              764       0.13
## 8   1983              20              809       0.13
## 9   1984              16              223       0.03
## 10  1985              22             1066       0.15
```

(a) Assume that the number of fatal accidents in each year are independent with a $Poisson(\theta)$ distribution. Using a flat prior for θ, find the posterior distribution for θ based on the the 10 years of provided data. If you have a $Gamma(\alpha, \beta)$ distribution then the function `qgamma(q, shape=a, rate=b)` will return the qth quantile of the $Gamma(\alpha, \beta)$ distribution. Use this to find the 'symmetric' 95% credible interval for θ.

Using a flat prior, we have $\theta \sim Gamma(1, 0)$. So our posterior distribution becomes $\theta|y \sim Gamma(1 + \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} xi)$, where y_i is the number of fatal accidents in the ith year, and x_i is the exposure (in this case, 1 year each).

```r
library(ggplot2)
library(MASS)


theta.given.y <- qgamma(c(0.025, 0.975),
                        shape=(1 + sum(airline.deaths$fatal.accidents)),
                        rate=(length(airline.deaths$years)))


x <- rgamma(100000,
            shape = (1 + sum(airline.deaths$fatal.accidents)),
            rate = length(airline.deaths$years))
den <- density(x)
```

```r
dat <- data.frame(x = den$x, y = den$y)
ggplot(data = dat, aes(x = x, y = y)) +
    geom_line() +
    geom_vline(xintercept = theta.given.y[1]) +
    geom_vline(xintercept = theta.given.y[2]) +
    theme_classic()
```



The symmetric 95% credible interval is $(20.9657674, 27.0236062)$.

∎

(b) Now assume that the number of fatal accidents in each year follow independent Poisson distributions with a constant rate and an exposure in each year proportional to the number of passenger miles flown. Again using a flat prior distribution for θ, determine the posterior distribution based on the data. Give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that 8×10^{11} passenger miles are flown that year.

```
miles.flown <- (airline.deaths$passenger.deaths / airline.deaths$death.rate) * 100000000
airline.deaths$miles.flown <- miles.flown
airline.deaths
```

```
##    years fatal.accidents passenger.deaths death.rate  miles.flown
## 1   1976              24              734       0.19 386315789474
## 2   1977              25              516       0.12 430000000000
## 3   1978              31              754       0.15 502666666667
## 4   1979              31              877       0.16 548125000000
## 5   1980              22              814       0.14 581428571429
## 6   1981              21              362       0.06 603333333333
## 7   1982              26              764       0.13 587692307692
## 8   1983              20              809       0.13 622307692308
## 9   1984              16              223       0.03 743333333333
## 10  1985              22             1066       0.15 710666666667
```

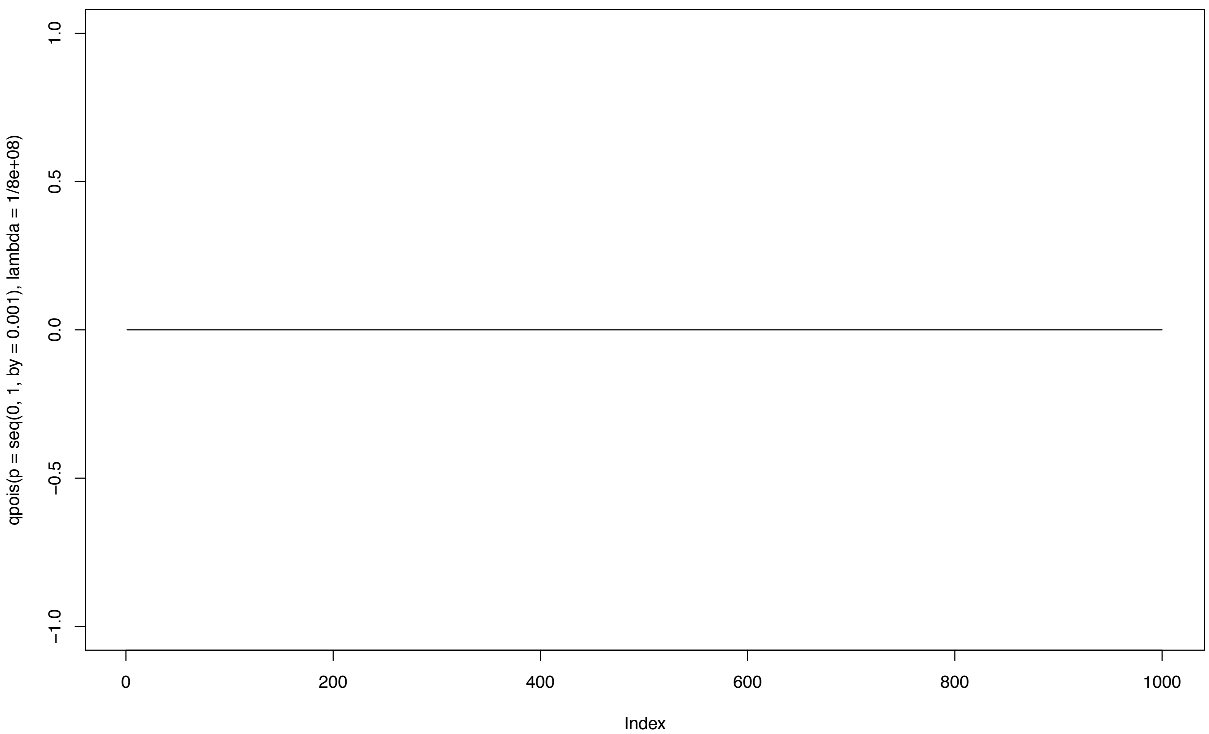
```
theta.given.y <- qgamma(c(0.025, 0.975),
                        shape = sum(1 + airline.deaths$fatal.accidents),
                        rate = airline.deaths$miles.flown[1])

theta.given.y
```

```
## [1] 5.645440e-10 7.242818e-10
```

```
plot(qpois(p=seq(0, 1, by=0.001), lambda = 1/800000000), type="l")
```

If asked to determine the posterior distribution, you should make sure you write what the posterior distribution is.

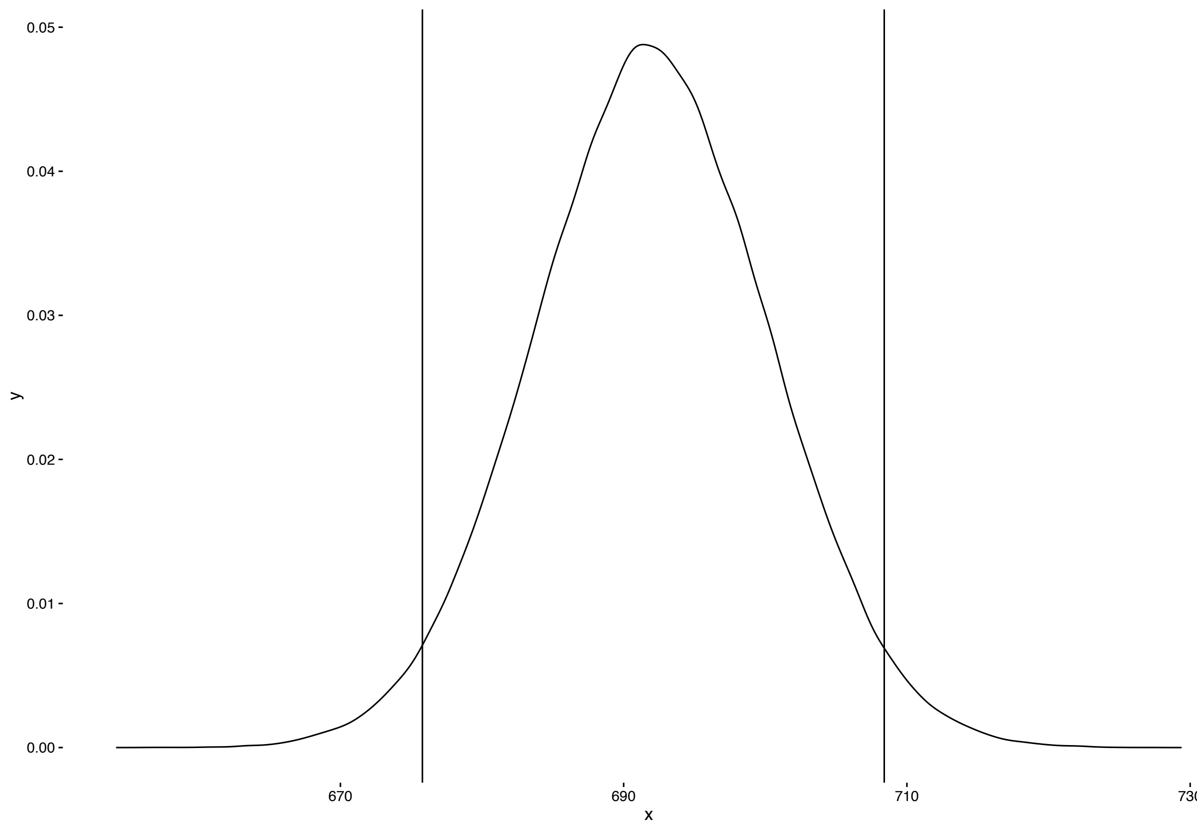
To find the predictive interval, first simulate values of theta from the posterior distribution, then use those to simulate the number of accidents in 1986. Use those simulations to construct the predictive interval.



I'm not seeing this one. The numbers keep ending up too big or too small. Definitely missing something here. :/

(c) Repeat (a) above, replacing 'fatal accidents' with 'passenger deaths.'

```r
theta.given.y <- qgamma(c(0.025, 0.975),
                        shape=(1 + sum(airline.deaths$passenger.deaths)),
                        rate=(length(airline.deaths$years)))

x <- rgamma(100000,
            shape = (1 + sum(airline.deaths$passenger.deaths)),
            rate = length(airline.deaths$years))
den <- density(x)
dat <- data.frame(x = den$x, y = den$y)
ggplot(data = dat, aes(x = x, y = y)) +
    geom_line() +
    geom_vline(xintercept = theta.given.y[1]) +
    geom_vline(xintercept = theta.given.y[2]) +
    theme_classic()
```



The symmetric 95% credible interval is $(675.7906593, 708.3987693)$.

■

(d) Repeat (b) above, replacing ‘fatal accidents’ with ‘passenger deaths.’

Since I couldn’t figure out b), I’m not getting this one either.