Homework 1

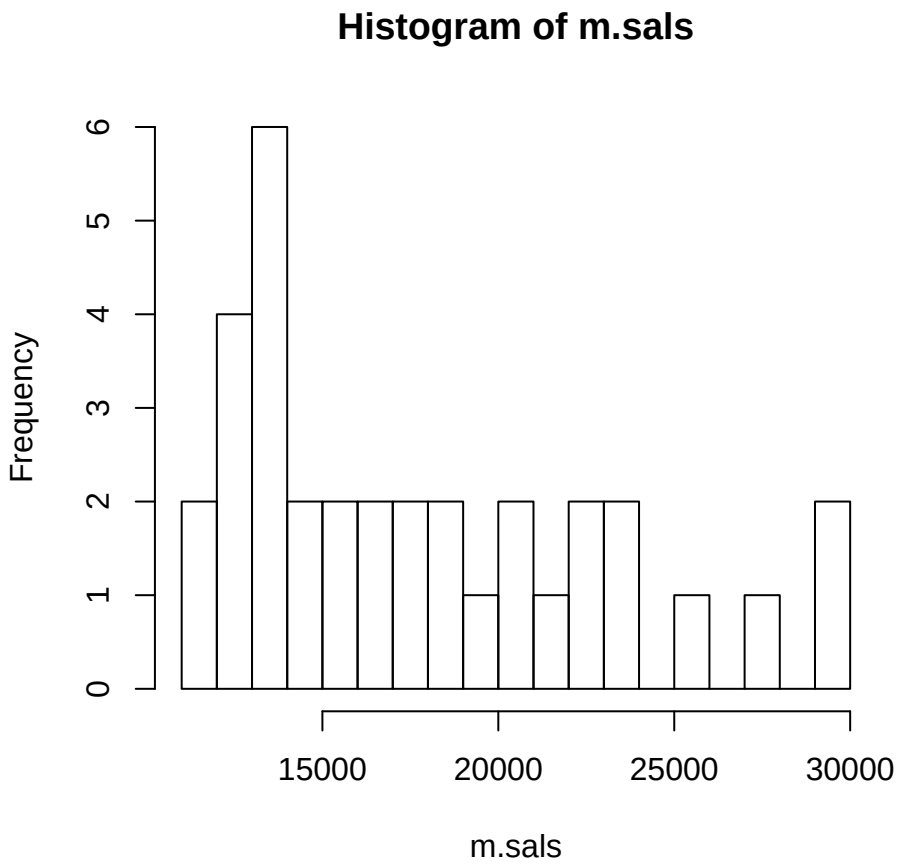
Steve Bronder
Statistical Inference

November 18, 2014

Exercise 1. Normal-normal model

Problem a. Check the sample size with length(m.sals). Plot a histogram of this data and commend on whether you feel it is appropriate to use a normal model for these data

```
sal.dat <- read.csv("./hw_5_data.csv",header=TRUE)

#fix rand num generator
set.seed(1234)

m.sals <- sal.dat[,6][sal.dat[,3]==1]

length(m.sals)

## [1] 34

hist(m.sals,breaks=20)
```



The data appears heavily skewed left. As such I would not recommend a normal model for this data.

Problem b. If we assume the normal model to hold for this data, and that the standard deviation of this model is $\sigma = 1.2$, what is the posterior distribution of μ, the unkown mean of the model? Use a conjugate prior for μ with mean 20 and variance 49.

To answer this question we will find the number of observations, the mean and standard deviation of the original data, and then compute the posterior mean of mu with the given mean and variance for the conjugate.

```
#number of observations
n <- length(m.sals)

# mean of data
mean.d <- mean(m.sals)

#standard deviation of data
sigma <- 1.2
```

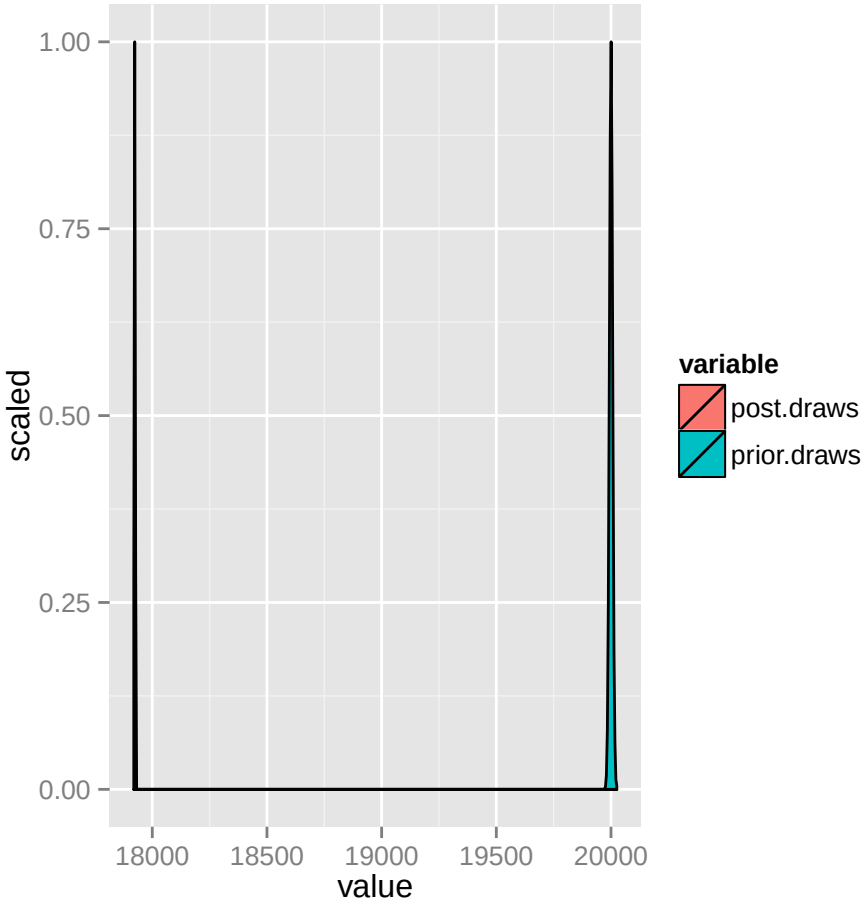
```r
# Mean of mu is gamma
gamma <- 20000

# variance of mu is tau
tau <- 49

#mean of mu
mu.m <- (tau * n * mean.d + sigma^2 * gamma)/(tau * n + sigma^2)

mu.m

## [1] 17923.53
```

In the next section we will find the posterior by simulating from a normal distribution given the updated mean.

Problem c. In a single plot, draw the posterior and prior distributions for this problem. Comment briefly on how the data have "updated" the prior distribution.

First, we will draw random samples from the prior and posterior distributions of the data.Then use **reshape2** to melt the data into long format so both distributions can be easily graphed with the **ggplot2** package.

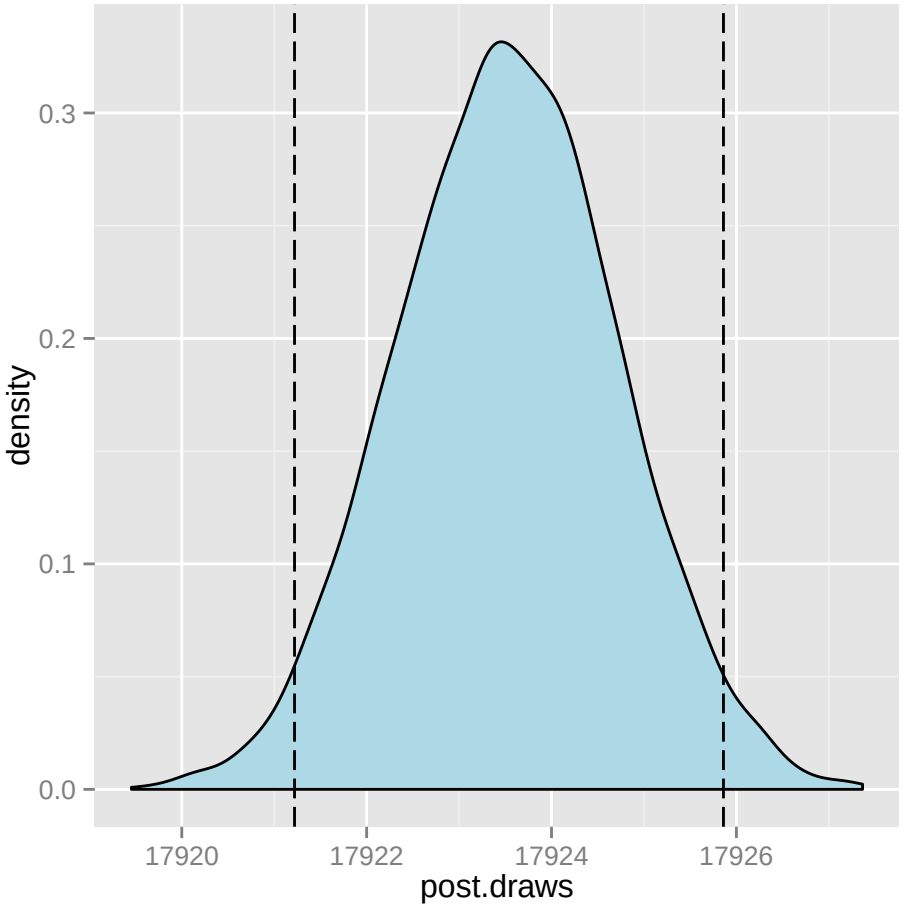
```r
library(ggplot2)
library(reshape2)

# posterior draws
post.draws <- rnorm(5000,mean=mu.m,sd=1.2)

#Original draws
prior.draws <- rnorm(5000,mean=gamma,sd=sqrt(tau))

#Bind both draws together
post.prior.draws <- as.data.frame(cbind(post.draws,prior.draws))

# melt draws
pp.d.melt <- melt(post.prior.draws)

## No id variables; using all as measure variables

#plot densities
den.draws <- ggplot(pp.d.melt,aes(x=value,group=variable)) +
 geom_density(aes(fill=variable,y=..scaled..))

den.draws
```



The posterior draw has updated the prior by shifting the posterior downwards and widening the distribution. This is due to the low value we gave on μ's prior and the uncertainty we placed into μ's variance.

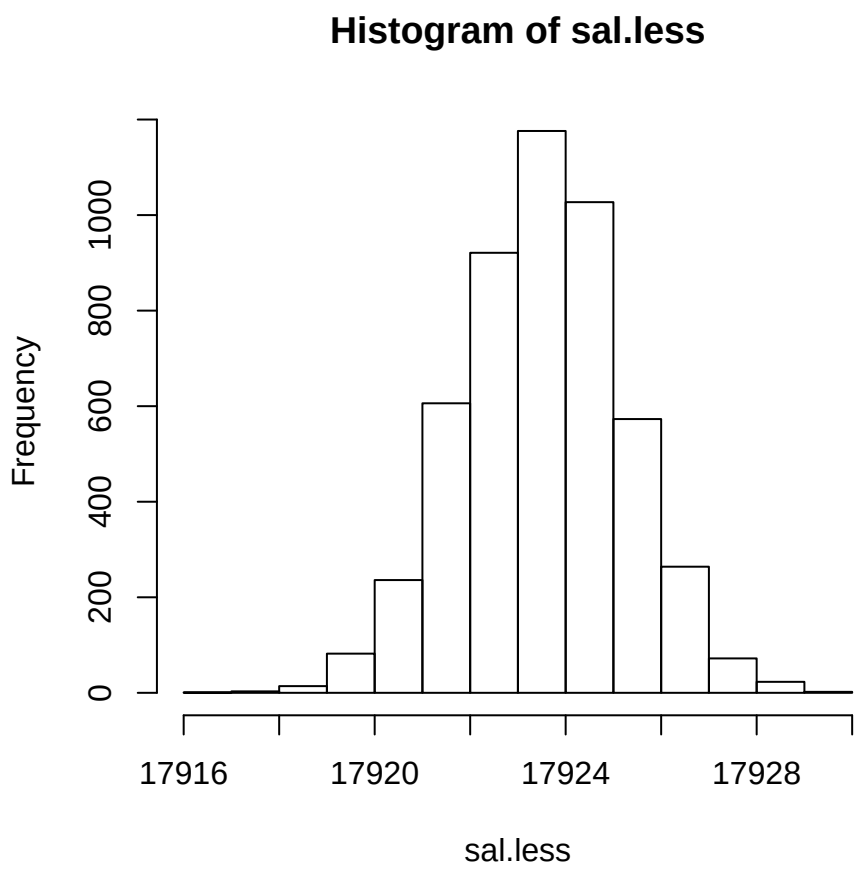
Problem d. Use the posterior distribution of μ to find a 95% credible interval for μ.

```
#95 percent credible interval
quantile(post.draws,c(.025,.975))

##     2.5%    97.5%
## 17921.22 17925.86

#graph of posterior with 95 percent credible interval
post.dens <- ggplot(as.data.frame(post.draws),aes(x=post.draws)) +
  geom_density(fill="lightblue") +
  geom_vline(xintercept = c(17921.22, 17925.86),linetype="longdash" )

post.dens
```



Problem e. Simulate 5000 draws from $\tilde{y}$ from the posterior predictive distribution and plot them in a histogram. Use these draws to find the probability that the next male sampled will have a salary less than \$15,000

```
#probability of salary less than 15000
sal.less <- rnorm(5000,post.draws,sd=1.2)

#histogram
hist(sal.less)
```



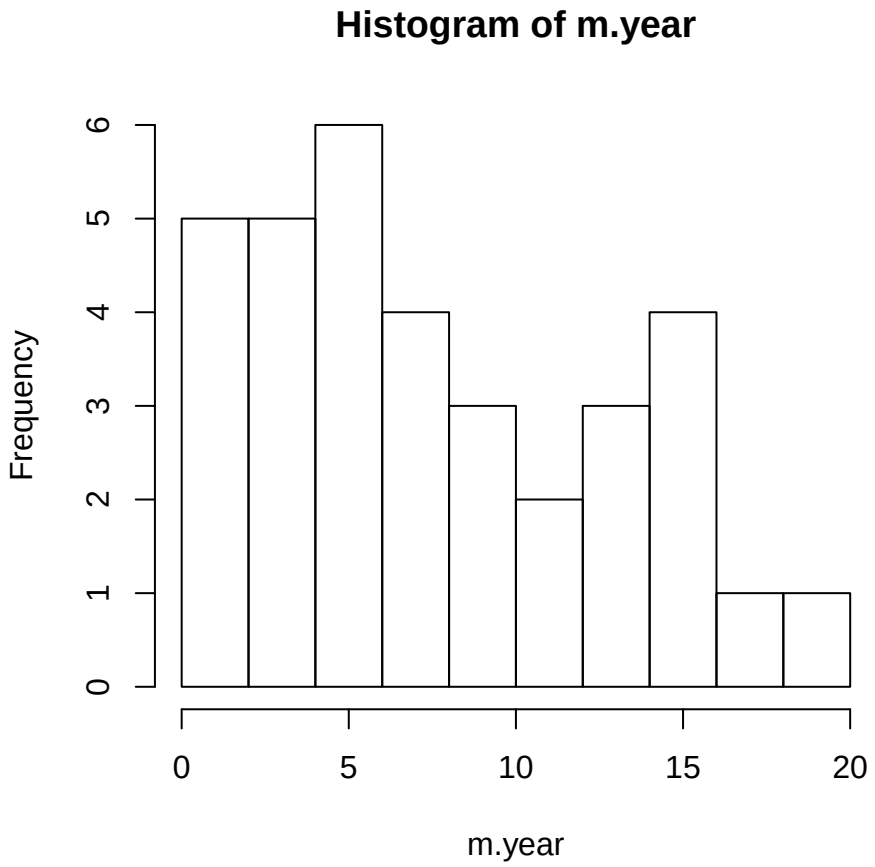

```
#probability of salary less than 15000
ecdf(sal.less)(15000)

## [1] 0

# zero percent chance?
```

Exercise 2. Poisson-gamma model

Problem a. Check the sample size with length(m.sals). Plot a histogram of this data and commend on whether you feel it is appropriate to use a Poisson model for these data

```
m.year <- sal.dat[,5][sal.dat[,3]==1]

#Check length
length(m.year)

## [1] 34
```

```r
# histogram of data
hist(m.year)
```




```r
# A poisson may be reasonable and I would recommend it
```

The data appears heavily skewed left. As such I would not recommend a poisson model for this data.

Problem b. We assume the Poisson model to hold for this data and λ is unkown. Assume the prior on λ is conjugate with mean 3 and variance 30. Find lambda

To answer this question we will use the gamma prior to find the value of lambda.

```r
# mean of gamma
gam.mean <- 3

# var of gamma

var.gamma <- 30
```

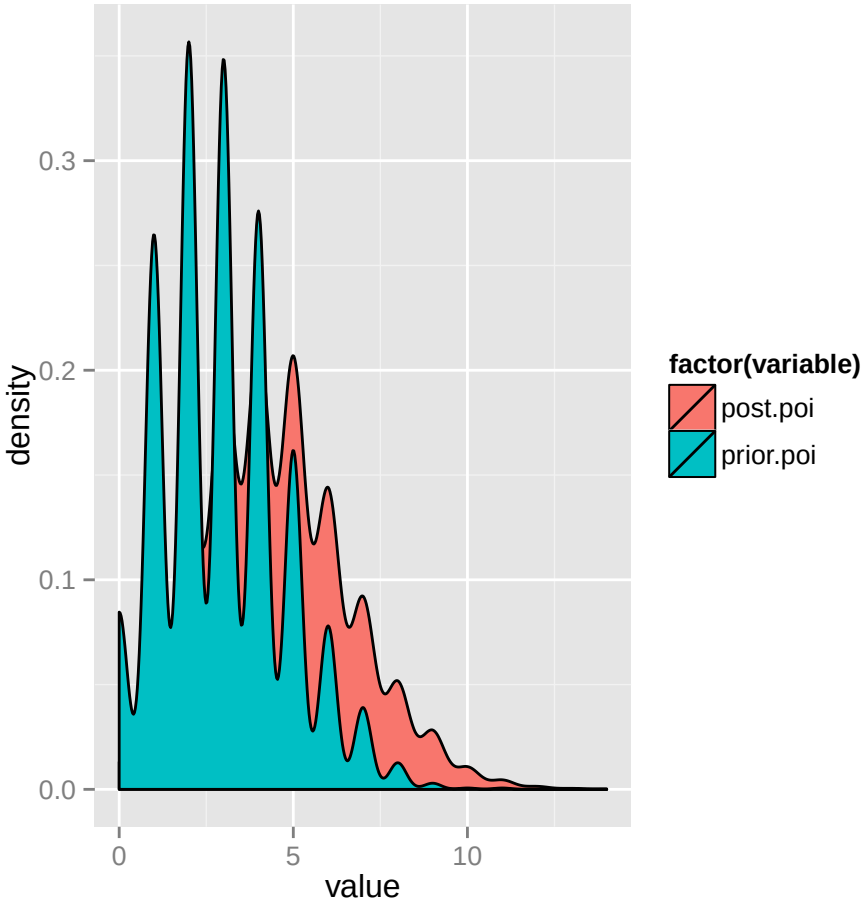
```
# posterior mean of lambda

lam <- (n/(n+var.gamma)) * (sum(m.year)/n) + (var.gamma/(n+var.gamma))*(gam.mean/var.gam

lam

## [1] 4.46875
```

In the next section we will find the posterior by simulating from a Poisson distribution given the updated λ.

Problem c. In a single plot, draw the posterior and prior distributions for this problem. Comment briefly on how the data have "updated" the prior distribution.

First, we will draw random samples from the prior and posterior distributions of the data.Then use **reshape2** to melt the data into long format so both distributions can be easily graphed with the **ggplot2** package.

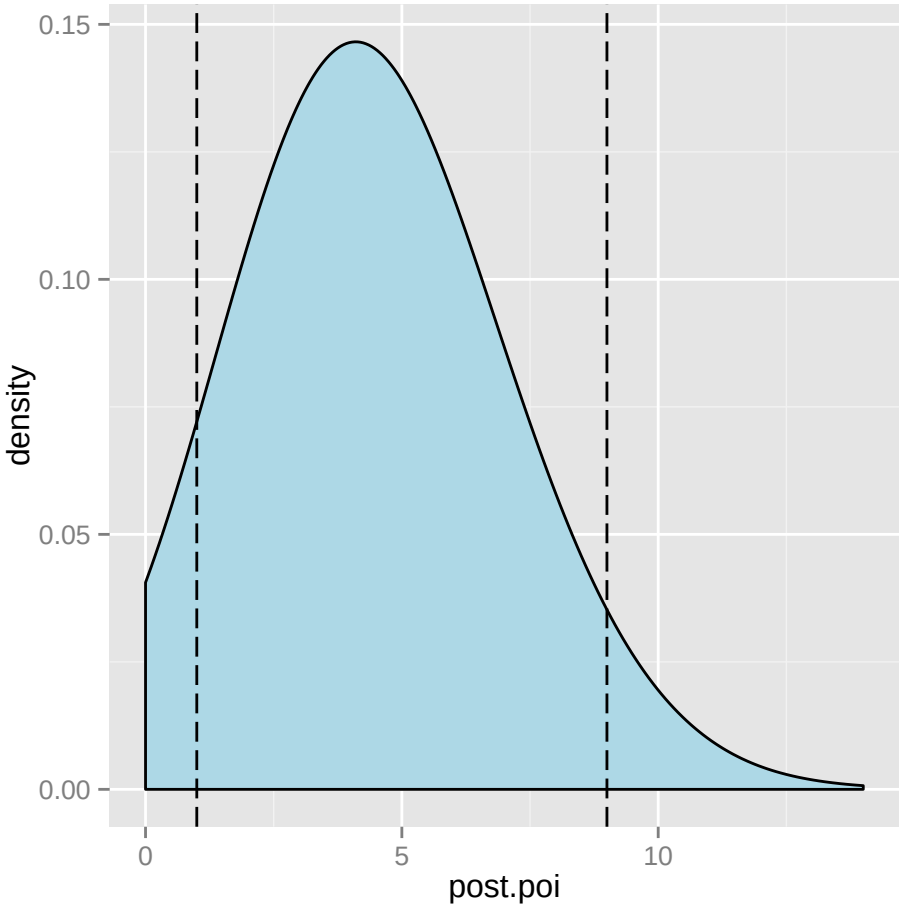
```
#posterior for poisson

post.poi <- rpois(5000,lam)

#prior for poisson

prior.poi <- rpois(5000,gam.mean)

#Bind both draws together

post.orig.draws <- as.data.frame(cbind(post.poi,prior.poi))

# melt draws
pp.d.melt <- melt(post.orig.draws)

## No id variables; using all as measure variables

#plot densities
poi.draws <- ggplot(pp.d.melt,aes(x=value,group=variable)) +
 geom_density(aes(fill=factor(variable),adjust=10))

poi.draws
```



The posterior draw has updated the prior by shifting the posterior updward and widening the distribution. This is due to the high value we gave on λ's prior and the uncertainty we placed into λ's variance.

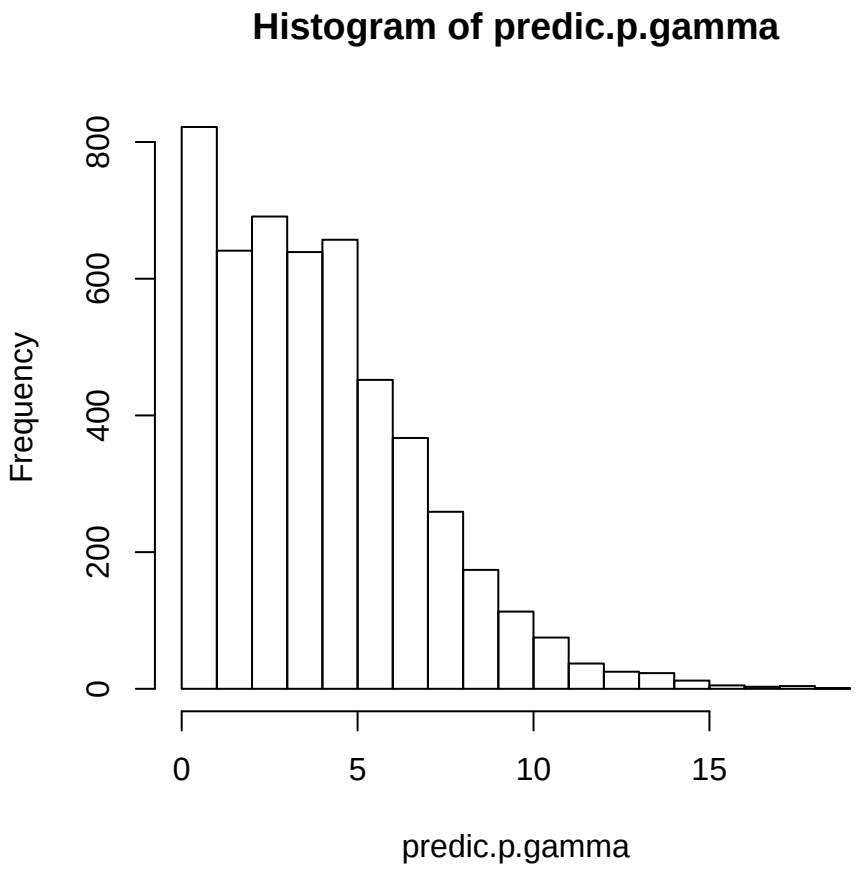
Problem d. Use the posterior distribution of μ to find a 95% credible interval for λ.

```
#95 percent credible interval
quantile(post.poi,c(.025,.975))

##  2.5% 97.5%
##     1     9

#graph of posterior with 95 percent credible interval
post.dens <- ggplot(as.data.frame(post.poi),aes(x=post.poi)) +
  geom_density(fill="lightblue",adjust=5) +
  geom_vline(xintercept = c(1,9),linetype="longdash" )

post.dens
```



Problem e. Simulate 5000 draws from $\tilde{y}$ from the posterior predictive distribution and plot them in a histogram. Use these draws to find the probability that the next male sampled will have a spend more than 10 years at his job.

```
#predictive prior
predic.p.gamma <- rpois(5000,post.poi)


hist(predic.p.gamma)
```




```
# Prob that next male will have more than 10 years on job
1-ecdf(predic.p.gamma)(10)

## [1] 0.037

# three percent chance that someone has spent more than 10 years at the job
```

With less than a 3% chance someone is at their job for more than 10 years we can not say this is a likely circumstance.