quiz1takehome

Michael Pena

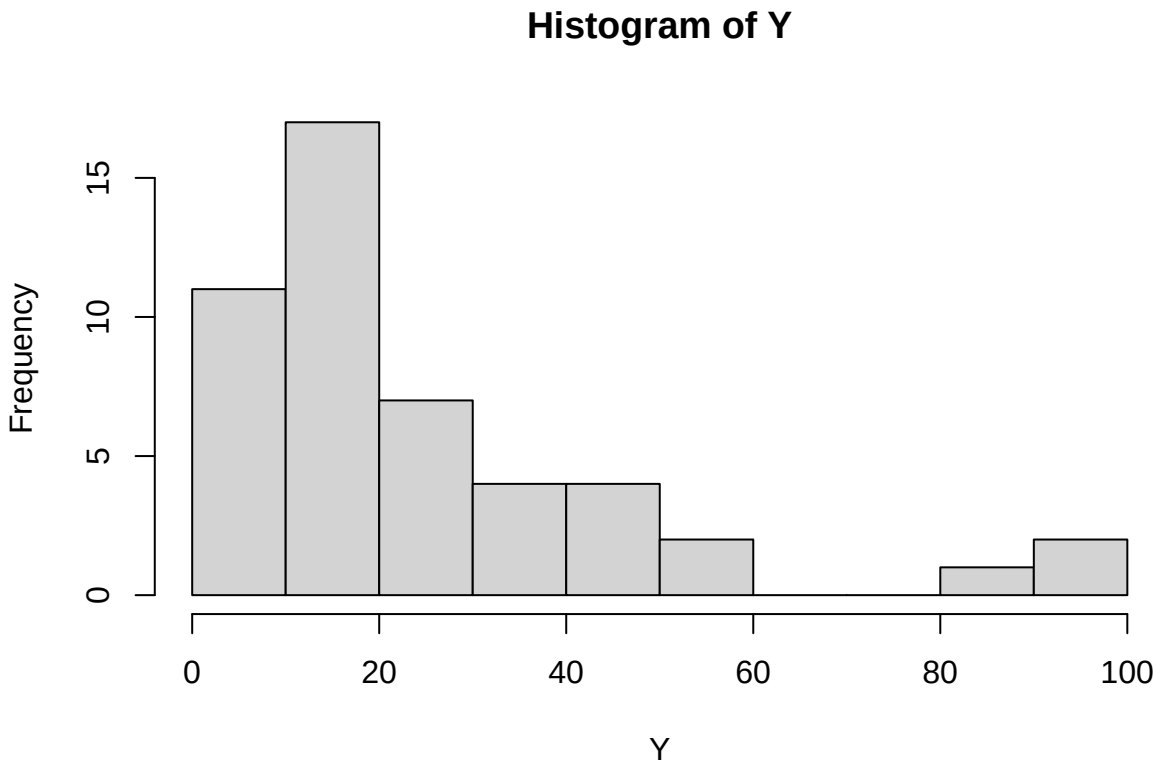
2024-09-20

[65 marks] [RELIABILITY APPLICATION] Consider the Engine.csv data set.The Raleigh distribution was named after Lord Raleigh, a renowned mathematician and physicist who received the Nobel Prize in 1904 for his discovery of Argon and related research. The Raleigh distribution, which is defined by the magnitude of two vectors arising from independent normal distributions centered at zero and having the same variance, is often utilized to model lifetime data. Here, we will apply the Raleigh distribution to a dataset describing the time (in weeks) to a valve seat replacement in 24 diesel engines.

$$y_i \sim f(y_i|\sigma^2) = \frac{y_i}{\sigma^2} e^{-\frac{y_i^2}{2\sigma^2}} \text{ with prior} \sigma^2 \sim p(\sigma^2|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} (\sigma^2)^{-(a_0+1)} e^{-\frac{b_0}{\sigma^2}}$$

problem (a)

```
# load data
Y <- as.matrix(read.csv('Engine.csv',header = T))
# visuals
hist(Y)
```



problem (b)

for the time being let's just pretend $\theta = \sigma^2$

$$L(y_i|\theta) = \prod_i^n \left[\frac{y_i}{\theta} e^{-\frac{y_i^2}{\theta}}\right] = (\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{\theta}} \cdot \prod_i^n y_i$$

```
Y = Y[,2]
n = length(Y)
Ysum = sum(Y^2)
Yprod = prod(Y)
n;Ysum;Yprod
```

```
## [1] 24
```

```
## [1] 47698.02
```

```
## [1] 2.844826e+35
```

$n = 24$ $\sum_i^n y_i^2 = 47698.02$ $\prod_i^n y_i = 2.844826\text{e+}35$

we can integrate over all $\theta \in (0, \infty)$

$$\prod_i^n y_i \int_0^\infty (\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}} \cdot d\theta$$

we can use the kernel of the Inverse Gamma function to find the normalizing constant where $\alpha = 25, \beta = 47698.02/2$

$$\int_0^\infty (\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}} \cdot d\theta = \frac{\Gamma(25)}{(\frac{\sum_i^n y_i^2}{2})^{n+1}}$$

$$\frac{(\prod_i^n y_i)(\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}}}{(\prod_i^n y_i)\int_0^\infty (\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}} \cdot d\theta} = \frac{(\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}}}{\frac{\Gamma(25)}{(\frac{\sum_i^n y_i^2}{2})^{n+1}}} = \frac{(\frac{\sum_i^n y_i^2}{2})^{n+1}}{\Gamma(25)} (\frac{1}{\theta})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\theta}}$$

this leaves the normalized likelihood as such

$$Lnorm(y_i|\sigma^2) = \frac{(\frac{\sum_i^n y_i^2}{2})^{n+1}}{\Gamma(25)} (\frac{1}{\sigma^2})^n \cdot e^{-\frac{\sum_i^n y_i^2}{2\sigma^2}}$$

problem (c)

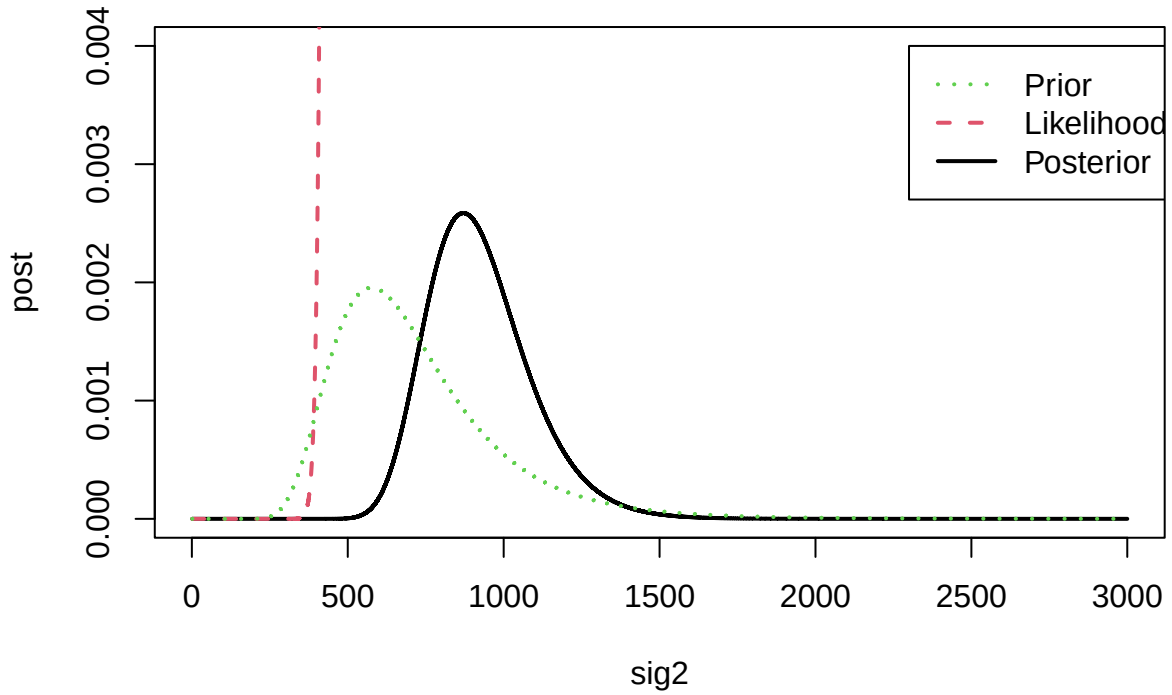
apriori that $6 = \frac{b_0}{a_0+1}$.

Let's choose that $b_0 = 7.2, a_0 = 0.2$

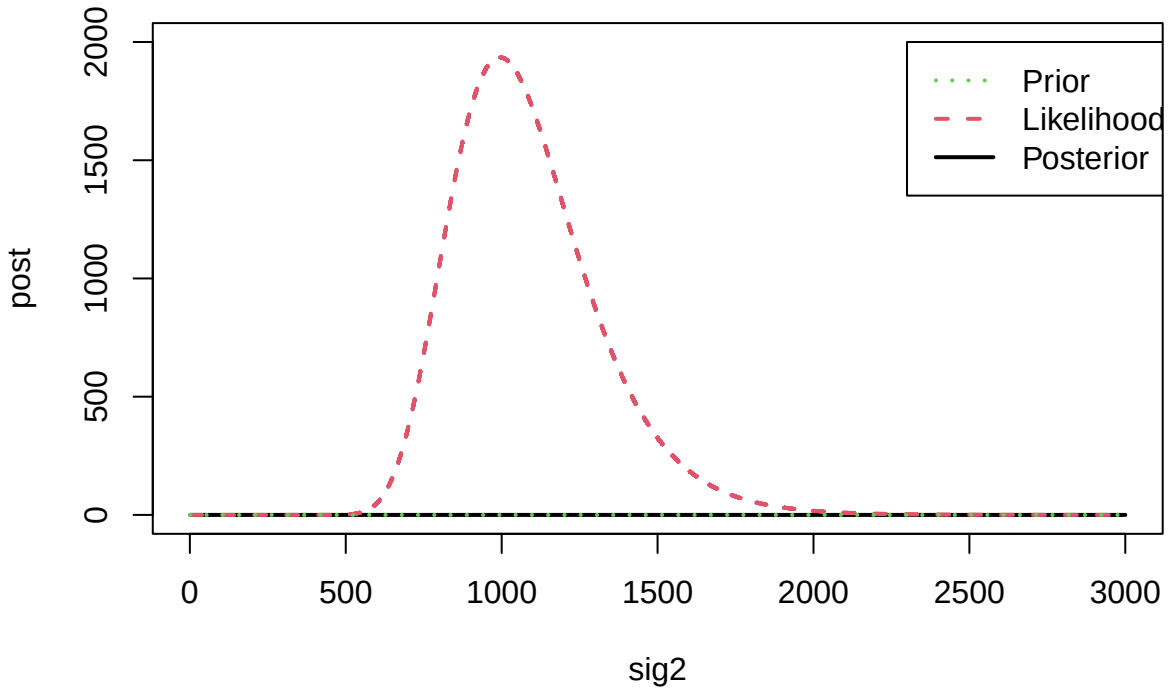
```
a0 = 9
b0 = (a0+1)*24^2

sig2 <- seq(1,3000,0.01)
dinvgamma(sig2,a0,b0) -> pri
like = (Ysum/2)^(n+1)/factorial(n) * (1/sig2)^n * exp(-Ysum/(2*sig2))
post = dinvgamma(sig2,n+a0,b0+0.5*Ysum)
```

```
plot(sig2,post, ylim = c(0,.004), type = 'l', col = 1, lwd = 2)
lines(sig2,like,lty = 2, col=2, lwd = 2)
lines(sig2,pri,lty=3,col=3,lwd=2)
legend(2300, .004, c("Prior", "Likelihood", "Posterior"),
       col = c(3, 2, 1),
       lty = c(3, 2, 1),
       lwd = c(2, 2, 2))
```




```
plot(sig2,post, ylim = c(0,2000), type = 'l', col = 1, lwd = 2)
lines(sig2,like,lty = 2, col=2, lwd = 2)
lines(sig2,pri,lty=3,col=3,lwd=2)
legend(2300, 2000, c("Prior", "Likelihood", "Posterior"),
       col = c(3, 2, 1),
       lty = c(3, 2, 1),
       lwd = c(2, 2, 2))
```

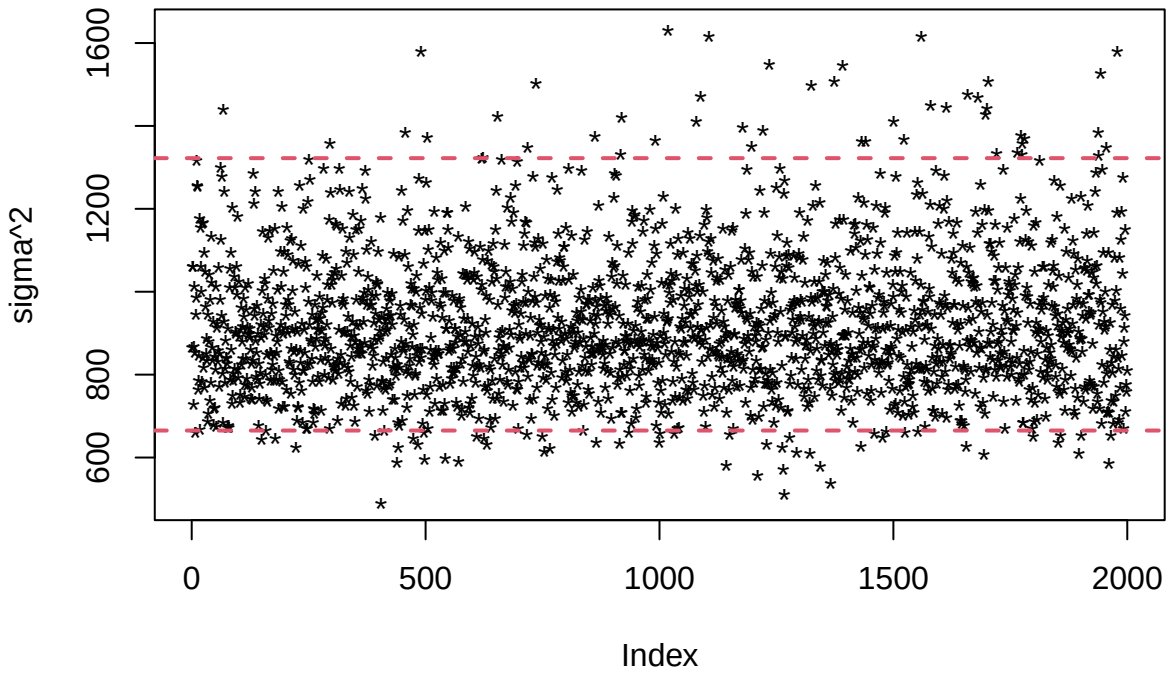


It does seem that the data driven likelihood has a closer mode to the posterior than that of the prior. Perhaps this means that the data still tells us more than the prior as is normal in bayesian data analysis.

I had to zoom out quite a bit but you can see that posterior peaks around 993.7088 which and no where near 24; the posterior mode is closer to the MLE

problem (d)

```
# generation MC samples
theta.post = rinvgamma(2000,n+a0,b0+0.5*Ysum)
plot(theta.post[1:2000], main = "MC Samples", pch = "*",ylab = "sigma^2")
abline(h = quantile(theta.post,.975) ,lty = 2, col=2, lwd = 2)
abline(h = quantile(theta.post,.025) ,lty = 2, col=2, lwd = 2)
```




```
# credibla interval
c(quantile(theta.post,.025),quantile(theta.post,.975))
```

```
##      2.5%     97.5%
##  665.1508 1322.2967
```

problem (e)

predictive prior

$$p(\tilde{y}) = \int_0^\infty p(\tilde{y}|\sigma^2)p(\sigma^2)d\sigma^2 = \int_0^\infty \frac{\tilde{y}}{\sigma^2} e^{-\frac{\tilde{y}^2}{2\sigma^2}} \cdot (\frac{1}{\sigma^2})^{a_0+1} e^{-\frac{b_0}{\sigma^2}} d\sigma^2$$

$$= \tilde{y}\frac{b_0^{a_0}}{\Gamma(a_0)} \int_0^\infty (\frac{1}{\sigma^2})^{a_0+2} e^{-\frac{\frac{1}{2}\tilde{y}^2+b_0}{\sigma^2}} d\sigma^2 = \tilde{y}\frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \frac{\Gamma(a_0+1)}{(\frac{1}{2}\tilde{y}^2+b_0)^{a_0+1}}$$

$$\Rightarrow p(\tilde{y}) = \frac{\tilde{y}a_0 b_0^{a_0}}{(\frac{1}{2}\tilde{y}^2+b_0)^{a_0+1}}$$

predictive posterior

$$p(\tilde{y}|y) = \int_0^\infty p(\tilde{y}|\sigma^2)p(\sigma^2|y)d\sigma^2 = \int_0^\infty \frac{\tilde{y}}{\sigma^2} e^{-\frac{\tilde{y}^2}{2\sigma^2}} \cdot \frac{(\frac{1}{2}\sum_i^n y_i^2)^{n+a_0}}{\Gamma(n+a_0)} (\frac{1}{\sigma^2})^{a_0+n+1} e^{-\frac{\frac{1}{2}\sum_i^n y_i^2}{\sigma^2}} d\sigma^2$$

$$= \frac{\tilde{y}(\frac{1}{2}\sum_i^n y_i^2)^{n+a_0}}{\Gamma(n+a_0)} \int_0^\infty e^{-\frac{\tilde{y}^2+\sum_i^n y_i^2+2b_0}{2\sigma^2}} (\frac{1}{\sigma^2})^{a_0+n+2} d\sigma^2$$

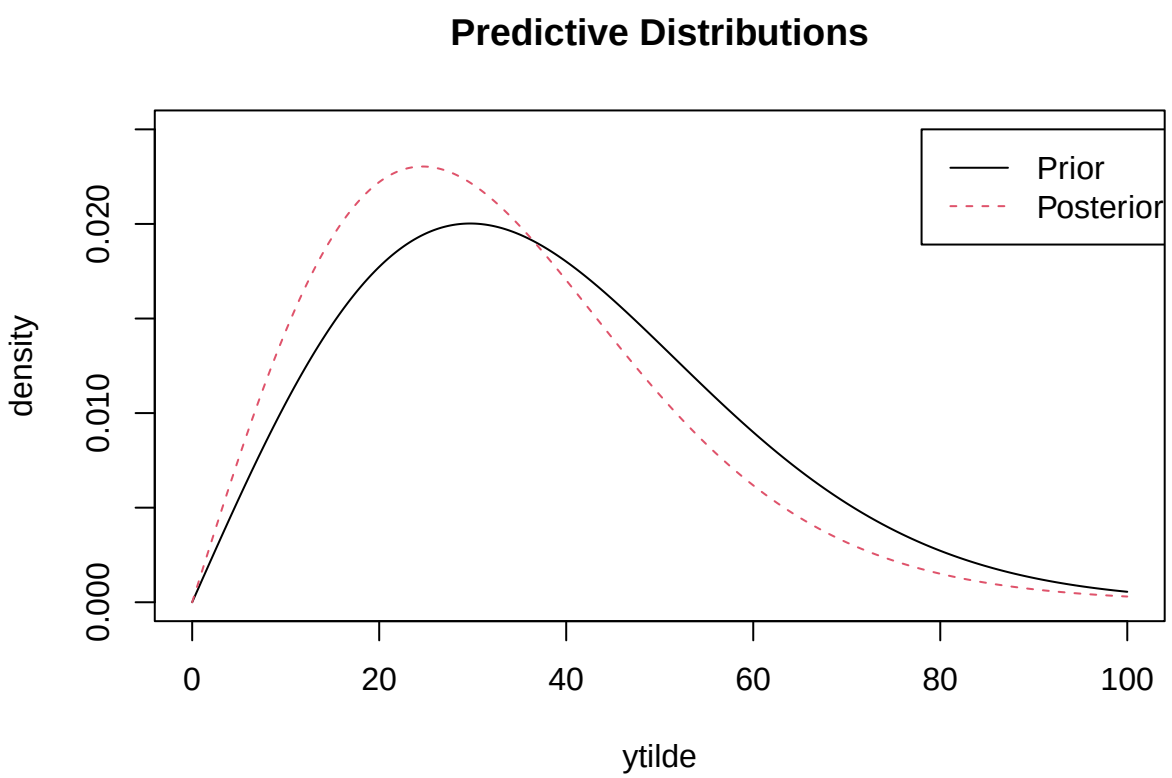
this contains a Inverse Gamma kernel for $\Gamma^{-1}(\frac{\tilde{y}^2+\sum_i^n y_i^2+2b_0}{2}, a_o+n+1)$

$$\Rightarrow \tilde{y}\frac{(\frac{1}{2}\sum_i^n y_i^2+b_o)^{n+a_o}}{\Gamma^{-1}(n+a_o)}\frac{\Gamma(a_o+n+1)}{(\frac{1}{2}\tilde{y}+\frac{1}{2}\sum_i^n y_i^2+b_o)^{a_o+n+1}}=\frac{\tilde{y}^2(a_o+n)(\frac{1}{2}\sum_i^n y_i^2+b_o)^{n+a_o}}{(\frac{1}{2}\tilde{y}^2+\frac{1}{2}\sum_i^n y_i^2+b_o)^{a_o+n+1}}$$

$$\Rightarrow p(\tilde{y}|y)=\frac{\tilde{y}(a_o+n)(\frac{1}{2}\sum_i^n y_i^2+b_o)^{n+a_o}}{(\frac{1}{2}\tilde{y}+\frac{1}{2}\sum_i^n y_i^2+b_o)^{a_o+n+1}}$$

```
# prior predictive
prior_pred <- function(y){
  (y*a0*b0^a0)/(0.5*y^2+b0)^(a0+1)
}
#posterior predictive
post_pred <- function(y){
  (y*(a0+n)*(0.5*Ysum+b0)^(n+a0))/(0.5*y^2+0.5*Ysum+b0)^(a0+n+1)
}
ytilde = seq(0,100,0.01)
```

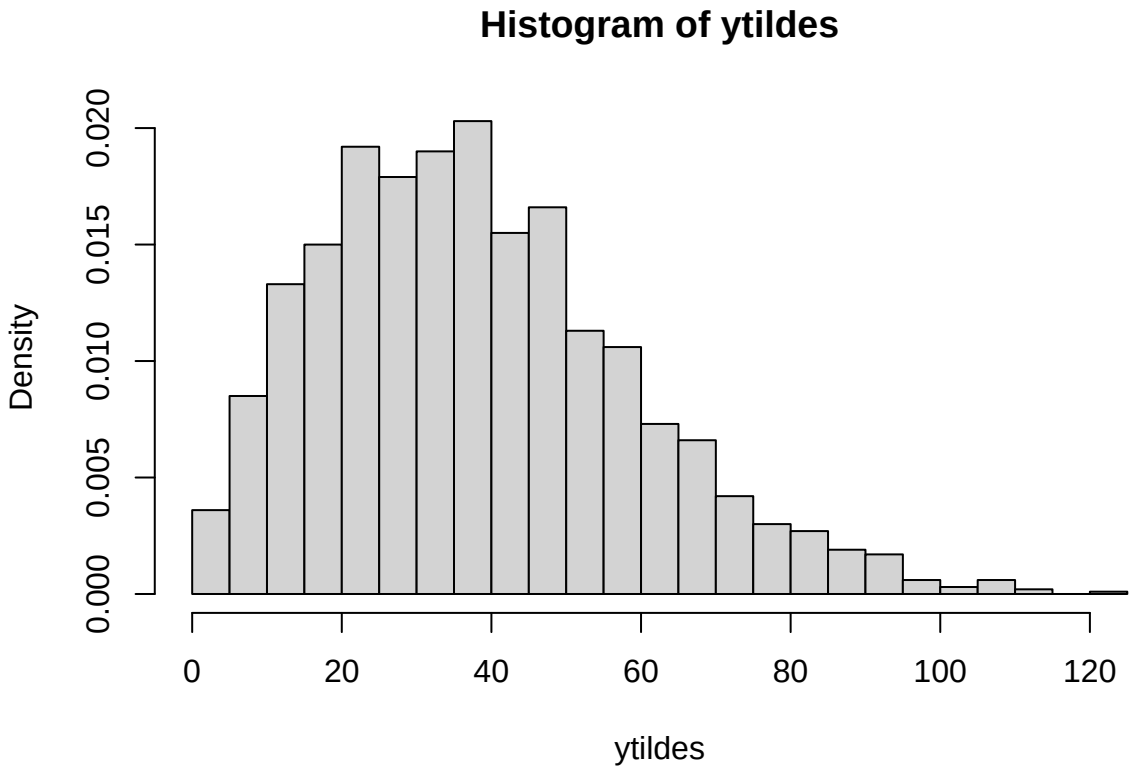
```
plot(ytilde,post_pred(ytilde),ylim= c(0,0.025),type='l',ylab="density",main="Predictive Distributions")
lines(ytilde,prior_pred(ytilde), lty = 2, col =2)
legend(78, .025, c("Prior", "Posterior"),
       col = c(1, 2),
       lty = c(1, 2))
```



The modes of the predictive distributions are closer together. The predictive distributions are closer to each other than the other predictions we have graphed.

problem (f)

```r
# monte carlo
rrayleigh(2000,sqrt(rinvgamma(2000,a0+n,0.5*Ysum+b0))) -> ytildes
hist(ytildes,freq=F,breaks = 40)
```




```r
# credible interval
c(quantile(ytildes,.025),quantile(ytildes,.975))
```

```
##      2.5%     97.5%
##  6.006863 85.324537
```

```r
# prob of picking a ytildes that is 24 or less
sum(ytildes <= 24)/length(ytildes)
```

```
## [1] 0.2785
```