732A91 Lab 1

Fanny Karelius (fanka300), Milda Poceviciute (milpo192)

11 april 2018

Question 1: Bernoulli... again

Let $y_1, ..., y_n|\theta \sim Bern(\theta)$, and assume $n = 20$ with $s = 14$ successes. Assume a $Beta(\alpha_0, \beta_0)$ prior, where $\alpha_0 = \beta_0 = 2$.

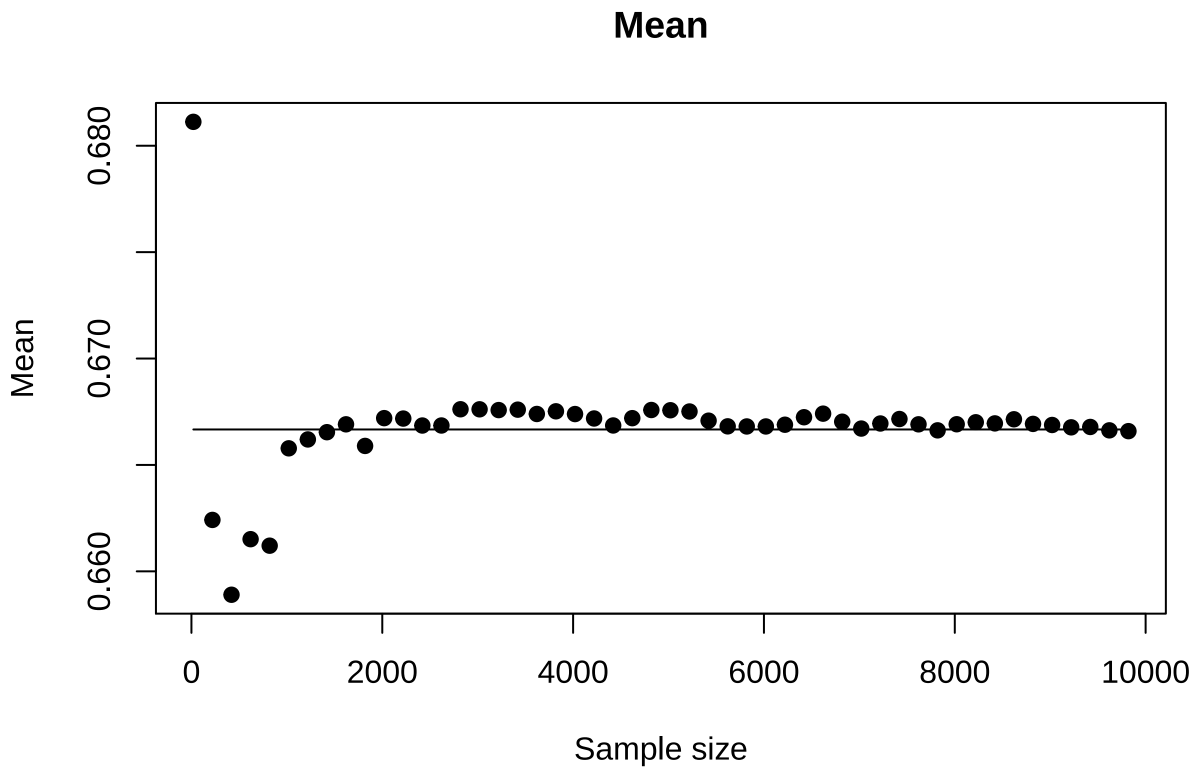
a)

Random numbers were drawn from $\theta|y \sim Beta(\alpha_0 + s, \beta_0 + f)$ and the posterior mean and standard deviation where plotted for different sample sizes.

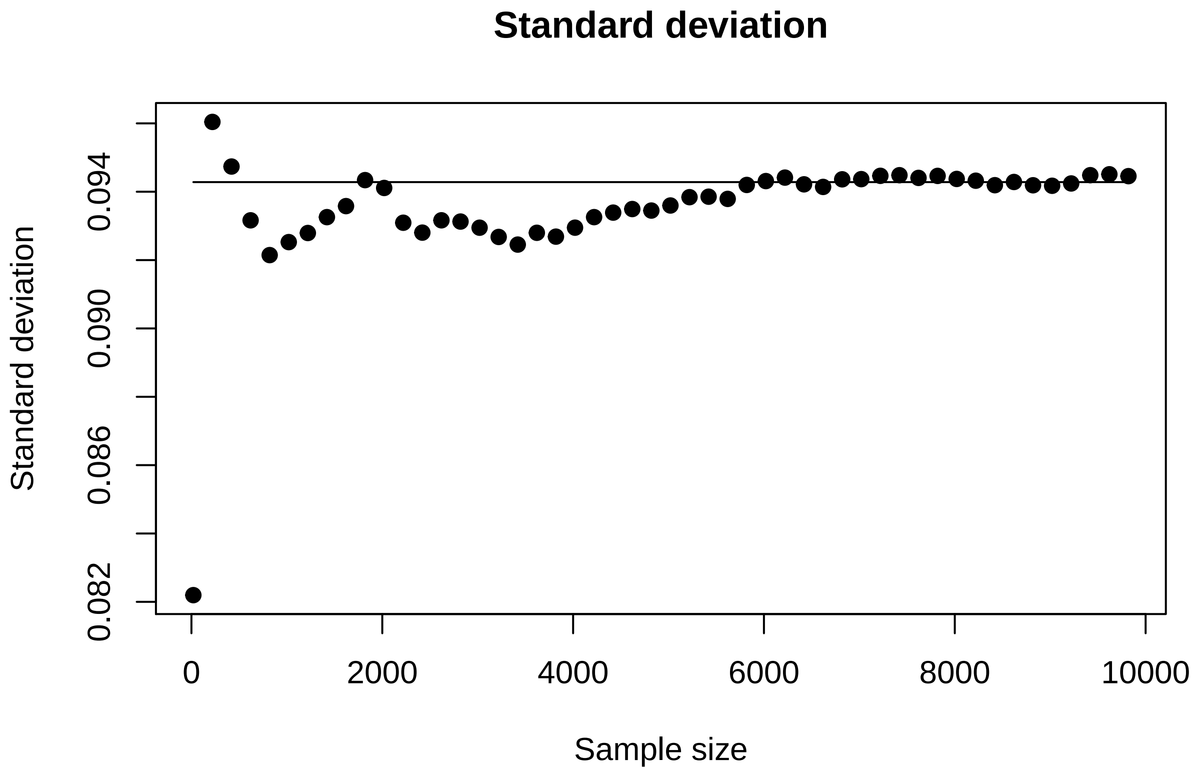
```r
alpha <- 2+14
beta <- 2+6
y1<-rbeta(20, alpha, beta)

t_mean <- alpha/(alpha+beta)
t_sd <- sqrt(alpha*beta/((alpha+beta)^2*(alpha+beta+1)))
samples = seq(20,10000,by=200)
result = data.frame()
for(i in 1:length(samples)){
  set.seed(12345)
  temp <- rbeta(samples[i], alpha, beta)
  result[i,1] <- mean(temp)
  result[i,2] <- sd(temp)
}
colnames(result)<-c("Mean", "Sd")

plot(samples, result$Mean, ylab="Mean", xlab="Sample size", main="Mean", pch=19)
lines(samples, rep(t_mean, length(samples)))
```



```
plot(samples, result$Sd, ylab="Standard deviation", xlab="Sample size", main="Standard deviation", pch=
lines(samples, rep(t_sd, length(samples)))
```



As can be seen by the graphs, the posterior mean and standard deviation converges to the true values (lines) as the sample size increases.

b)

```
set.seed(12345)
y<-rbeta(10000, alpha, beta)
prob <- length(y[y<0.4])/length(y)
true_prob <- pbeta(0.4, alpha, beta)
cat("Computed posterior probability: ", prob)
```

```
## Computed posterior probability:  0.0046
```

```
cat("True posterior probability: ", true_prob)
```

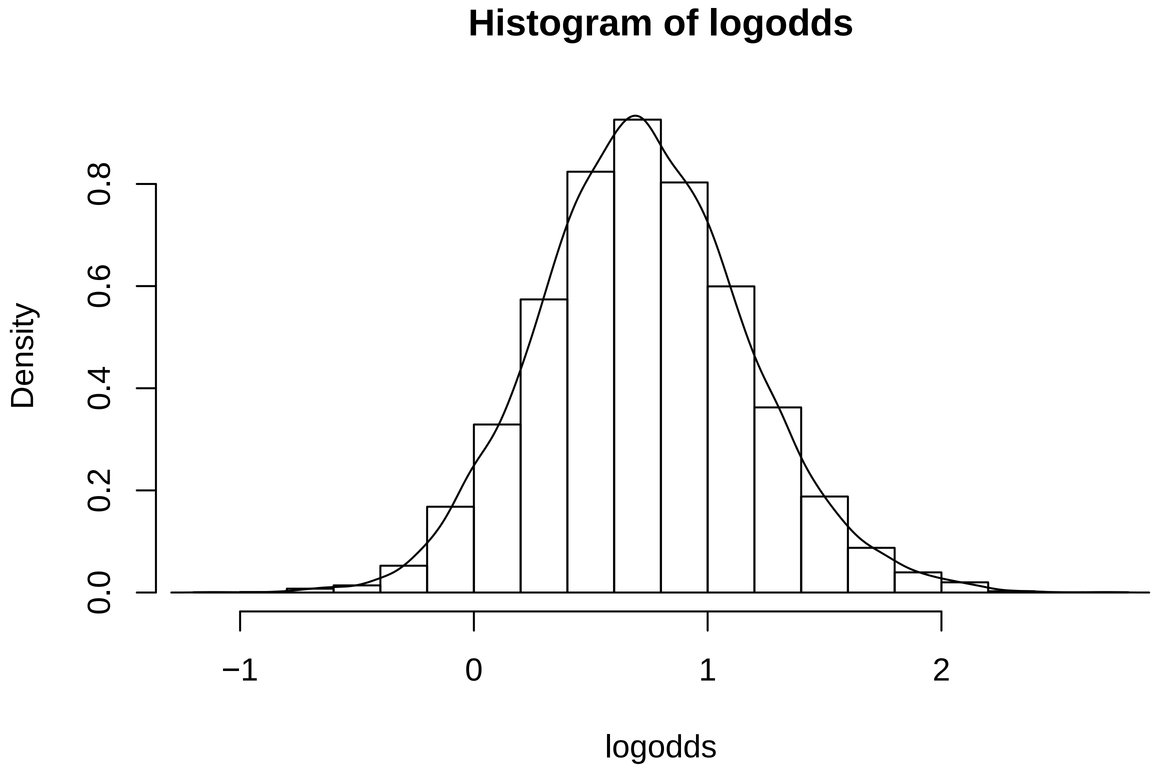
```
## True posterior probability:  0.003972681
```

The computed and true posterior probability values ($Pr(\theta < 0.4|y)$) are quite close to each other for 10000 draws.

c)

The computed posterior distribution for the log-odds function $\phi = log\frac{\theta}{1-\theta}$.

```
logodds <- log(y/(1-y))
hist(logodds, freq = FALSE)
lines(density(logodds))
```



Question 2: Log-normal distribution and the Gini coefficient

The data given follows a Log-normal distribution, the density function is given by: $p(y|\mu,\sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\}$, for $y > 0, \mu > 0, \sigma^2 > 0$. If $y \sim \log N(\mu,\sigma^2)$, then $\log y \sim N(\mu,\sigma^2)$. Let $y_1, ..., y_n|\mu,\sigma^2 \sim \log N(\mu,\sigma^2)$, where $\mu = 3.5$ is assumed to be known. σ^2 is unknown with non-informative prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$. The posterior for σ^2 is the (scaled) $Inv - \chi^2(n,\tau^2)$ distribution, where $\tau^2 = \frac{\sum_{i=1}^n (\log y_i - \mu)^2}{n}$.

a)

10000 draws were simulated from the posterior of σ^2.

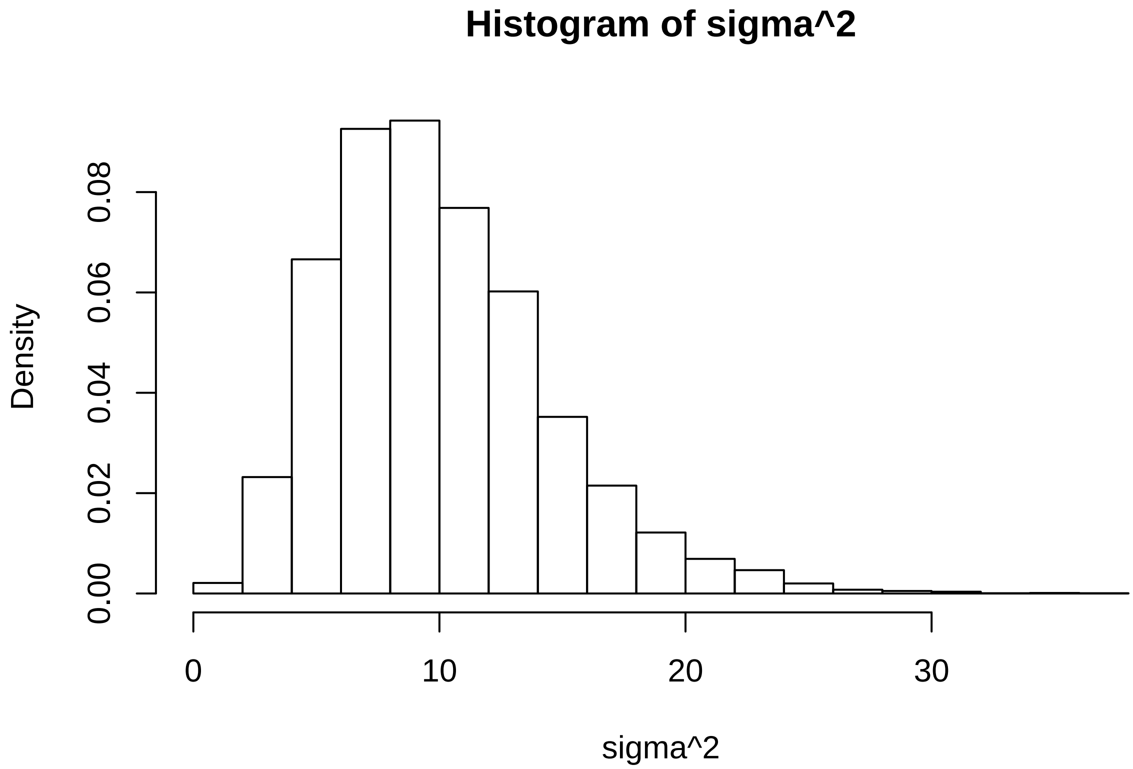
```
x<-c(14, 25, 45, 25, 30, 33, 19, 50, 34, 67)
mu <- 3.5
n <- length(x)
tau2 <- sum((log(x)-mu)^2)/n
set.seed(12345)
sim_sigma2 <- rchisq(10000,n)
```

```
sigma2 <- n*tau2/sim_sigma2

s_mean<-mean(sigma2)
s_sd<-sd(sigma2)
#Theoretical values for scaled inv-chi (wikipedia):
inv_mean<-n*tau2/(n-2)
inv_sd <- sqrt(2*inv_mean^2/(n-4))
mat <- matrix(c(s_mean, inv_mean, s_sd, inv_sd), ncol=2)
colnames(mat)<-c("Mean", "Sd")
rownames(mat)<-c("Simulated", "Theoretical")
mat
```

```
##                  Mean        Sd
## Simulated   0.2467178 0.1439459
## Theoretical 0.2473497 0.1428074
```

```
hist(sim_sigma2, xlab="sigma^2", main="Histogram of sigma^2", freq = FALSE)
```

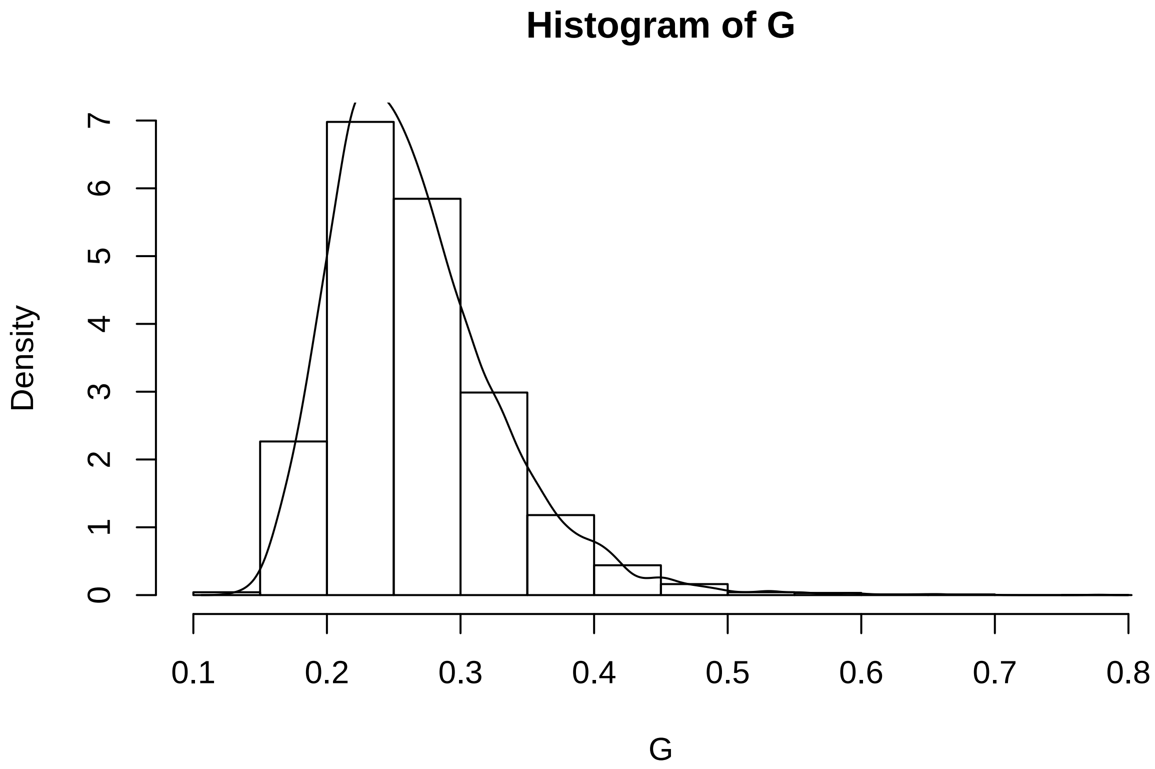


The mean and standard deviation for the simulated values are very close to the theoretical values.

b)

The Gini coefficient is given by $G = 2\Phi(\sigma/\sqrt{2}) - 1$ $(0 \leq G \leq 1)$ when income follows a log-Normal distribution. The posterior distribution of the Gini coefficient G for the data set:

```
set.seed(12345)
G <- 2*pnorm(sqrt(sigma2)/sqrt(2))-1
dens_G<-density(G)
lognorm = rlnorm(10000, mean(G), sd(G))
hist(G, freq=FALSE)
lines(dens_G)
```



c)

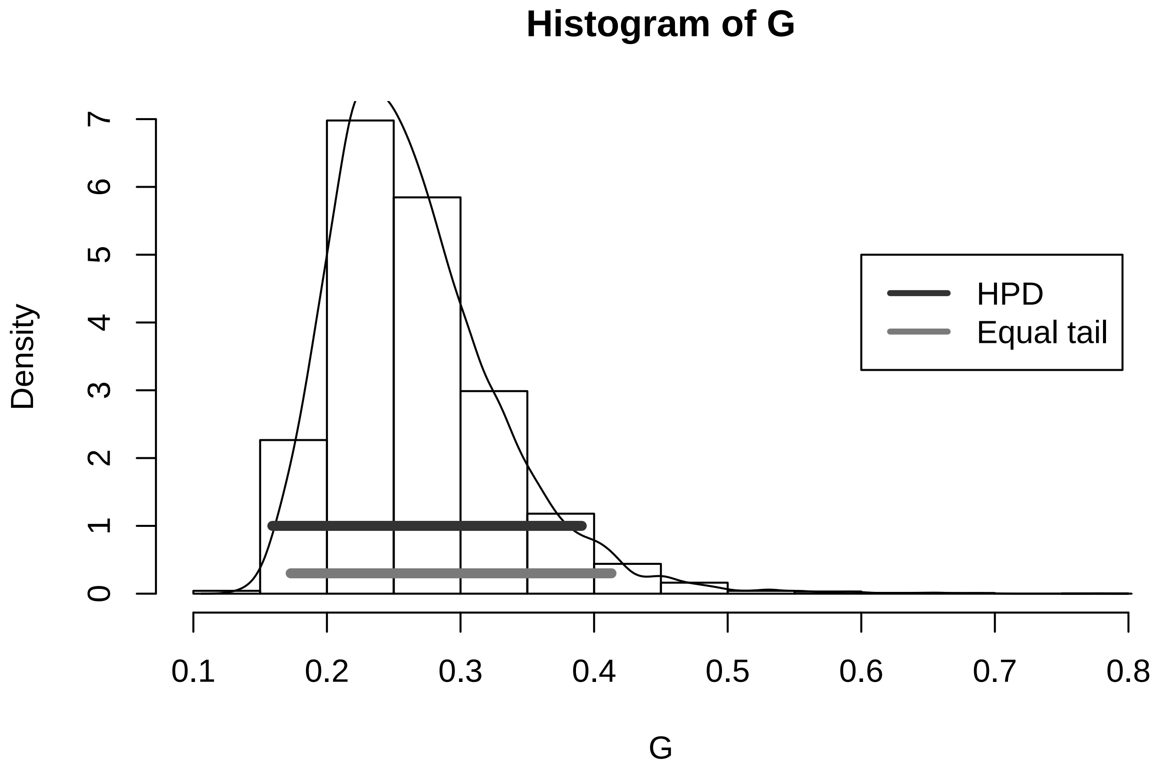
95% equal tail credible interval and a 95% highest posterior density interval for G.

```
perc = 0.025*10000
lowertail <- G[order(G, decreasing = FALSE)[perc+1]]
uppertail <- G[order(G, decreasing = FALSE)[10000-perc-1]]

dataframe <- data.frame(y=dens_G$y,x=dens_G$x)
dataframe <- dataframe[order(dataframe$y,decreasing = TRUE),]
dataframe$dens <- cumsum(dataframe$y)/sum(dataframe$y)

dfx<-dataframe$dens<0.95
low<-which.min(dataframe$x[dfx])
upp<-which.max(dataframe$x[dfx])
x_low <- dataframe$x[low]
x_upp <- dataframe$x[upp]
```

```r
hist(G, freq=FALSE)
lines(density(G), lwd=1)
lines(c(lowertail, uppertail), c(0.3,0.3), col="grey48", lwd=5)
lines(c(x_low, x_upp), c(1,1), col="grey20", lwd=5)
legend(x = 0.6, y=5, c("HPD", "Equal tail"), col=c("grey20", "grey48"), lwd = 3)
```



Fromm the plot above we see that the 95% Highest Posterior Density is placed more to the left of the histogram in comparison to the 95% Equal Tail Credible Interval. That is an expected result, because the density function is right-skewed.

Quetsion 3

The data points are assumed to be independent observations following a von Mises distribution: $p(y|\mu,\kappa) = \frac{exp\{\kappa\cos(y-\mu)\}}{2\pi I_0(\kappa)}$, $-\pi \leq y \leq \pi$ and $I_0(\kappa)$ is the modified Bessel function of the first kind (order zero). $\mu = 2.39$ is assumed to be known and $\kappa > 0$. $\kappa \sim Exp(\lambda = 1)$ a priori.

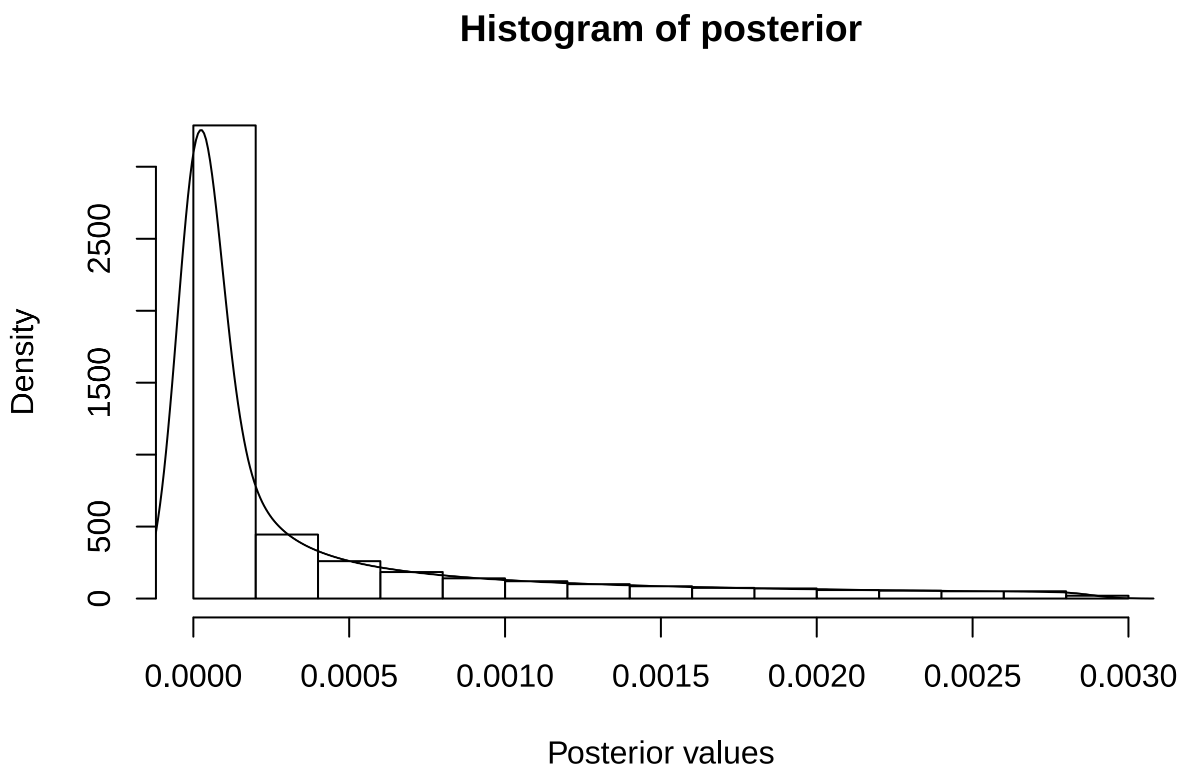
a)

Plot the posterior distribution of κ for the wind direction data over a fine grid of κ values:

```r
z <- c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)
mu2 <- 2.39
lambda <- 1
```

```
posterior <- function(k){
  result <- lambda*exp(k*sum(cos(z-mu2))-lambda*sum(z))/(2*pi*besselI(k,0))
  return(result)
}

ks <- seq(0,1,by=0.001)
post <- posterior(ks)
dens_post <- density(post)
hist(post, freq = FALSE, xlab="Posterior values", main="Histogram of posterior")
lines(dens_post)
```



b)

The (approximate) posterior mode of κ is:

```
post_mode <- max(dens_post$x)
post_mode
```

```
## [1] 0.003080267
```

Appendix

```
alpha <- 2+14
beta <- 2+6
y1<-rbeta(20, alpha, beta)

t_mean <- alpha/(alpha+beta)
t_sd <- sqrt(alpha*beta/((alpha+beta)^2*(alpha+beta+1)))
samples = seq(20,10000,by=200)
result = data.frame()
for(i in 1:length(samples)){
  set.seed(12345)
  temp <- rbeta(samples[i], alpha, beta)
  result[i,1] <- mean(temp)
  result[i,2] <- sd(temp)
}
colnames(result)<-c("Mean", "Sd")

plot(samples, result$Mean, ylab="Mean", xlab="Sample size", main="Mean", pch=19)
lines(samples, rep(t_mean, length(samples)))
plot(samples, result$Sd, ylab="Standard deviation", xlab="Sample size", main="Standard deviation", pch=
lines(samples, rep(t_sd, length(samples)))
set.seed(12345)
y<-rbeta(10000, alpha, beta)
prob <- length(y[y<0.4])/length(y)
true_prob <- pbeta(0.4, alpha, beta)
cat("Computed posterior probability: ", prob)
cat("True posterior probability: ", true_prob)
logodds <- log(y/(1-y))
hist(logodds, freq = FALSE)
lines(density(logodds))
x<-c(14, 25, 45, 25, 30, 33, 19, 50, 34, 67)
mu <- 3.5
n <- length(x)
tau2 <- sum((log(x)-mu)^2)/n
set.seed(12345)
sim_sigma2 <- rchisq(10000,n)

sigma2 <- n*tau2/sim_sigma2

s_mean<-mean(sigma2)
s_sd<-sd(sigma2)
#Theoretical values for scaled inv-chi (wikipedia):
inv_mean<-n*tau2/(n-2)
inv_sd <- sqrt(2*inv_mean^2/(n-4))
mat <- matrix(c(s_mean, inv_mean, s_sd, inv_sd), ncol=2)
colnames(mat)<-c("Mean", "Sd")
rownames(mat)<-c("Simulated", "Theoretical")
mat
hist(sim_sigma2, xlab="sigma^2", main="Histogram of sigma^2", freq = FALSE)
set.seed(12345)
G <- 2*pnorm(sqrt(sigma2)/sqrt(2))-1
dens_G<-density(G)
lognorm = rlnorm(10000, mean(G), sd(G))
hist(G, freq=FALSE)
```

```
lines(dens_G)
perc = 0.025*10000
lowertail <- G[order(G, decreasing = FALSE)[perc+1]]
uppertail <- G[order(G, decreasing = FALSE)[10000-perc-1]]

dataframe <- data.frame(y=dens_G$y,x=dens_G$x)
dataframe <- dataframe[order(dataframe$y,decreasing = TRUE),]
dataframe$dens <- cumsum(dataframe$y)/sum(dataframe$y)

dfx<-dataframe$dens<0.95
low<-which.min(dataframe$x[dfx])
upp<-which.max(dataframe$x[dfx])
x_low <- dataframe$x[low]
x_upp <- dataframe$x[upp]

hist(G, freq=FALSE)
lines(density(G), lwd=1)
lines(c(lowertail, uppertail), c(0.3,0.3), col="grey48", lwd=5)
lines(c(x_low, x_upp), c(1,1), col="grey20", lwd=5)
legend(x = 0.6, y=5, c("HPD", "Equal tail"), col=c("grey20", "grey48"), lwd = 3)

z <- c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)
mu2 <- 2.39
lambda <- 1

posterior <- function(k){
  result <- lambda*exp(k*sum(cos(z-mu2))-lambda*sum(z))/(2*pi*besselI(k,0))
  return(result)
}

ks <- seq(0,1,by=0.001)
post <- posterior(ks)
dens_post <- density(post)
hist(post, freq = FALSE, xlab="Posterior values", main="Histogram of posterior")
lines(dens_post)
post_mode <- max(dens_post$x)
post_mode
```