Lab1

Computer Lab 1

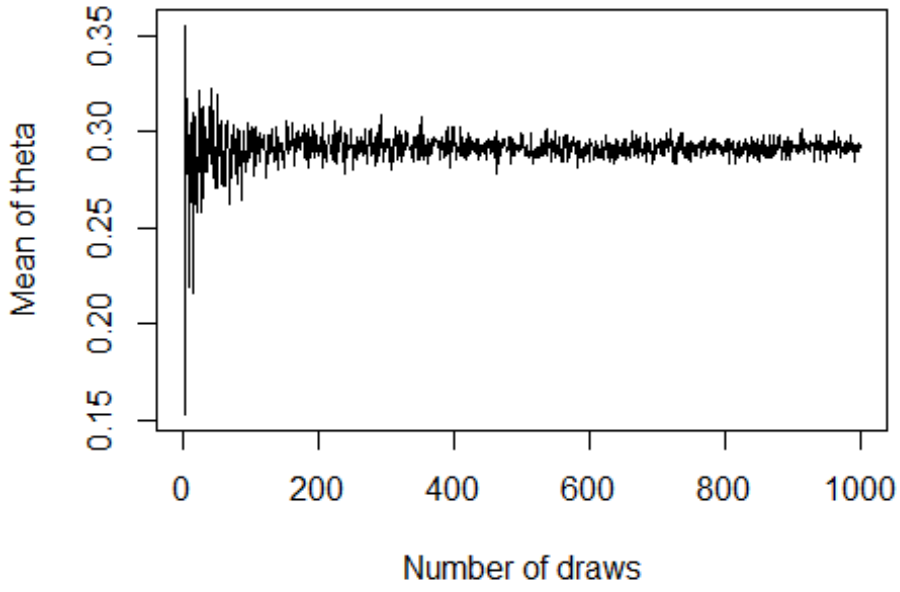
Assignment 1

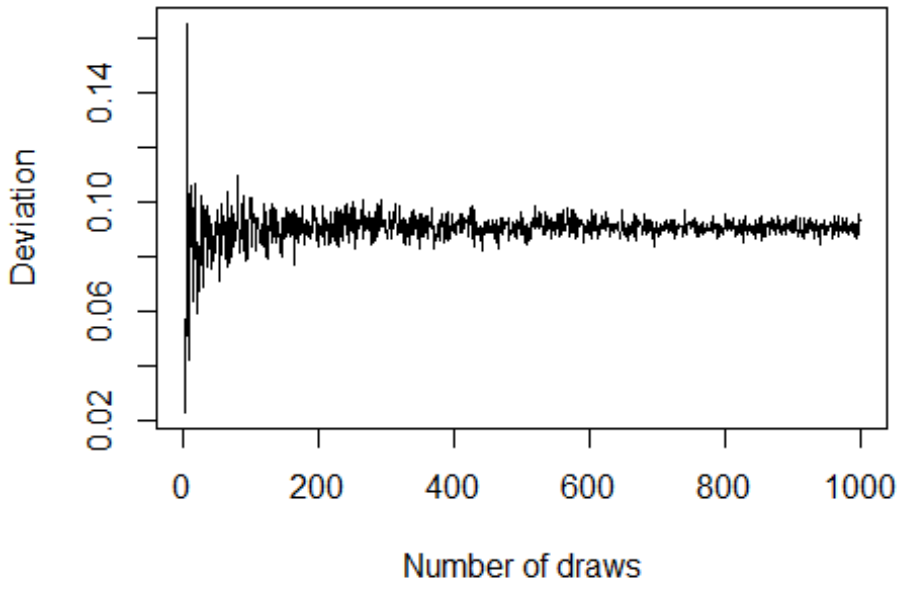
*(1a)

Let $y_1, ..., y_n|\theta \sim \text{Bern}(\theta)$, and assume that you have obtained a sample with $s = 5$ successes in $n = 20$ trials. Assume a $\text{Beta}(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 2$.

(a) Draw random numbers from the posterior $\theta|y \sim \text{Beta}(\alpha_0 + s, \beta_0 + f)$, $y = (y_1, \ldots, y_n)$, and verify graphically that the posterior mean and standard deviation converges to the true values as the number of random draws grows large.

As n gets bigger the mean and standard deviation converges towards its true value. Bigger amount of data (n) results in that the prior has less influence on the posterior. Mean -> 0,29 st -> 0,09





(b) Use simulation (nDraws = 10000) to compute the posterior probability Pr(phi > 0.3|y) and compare with the exact value [Hint: pbeta()].

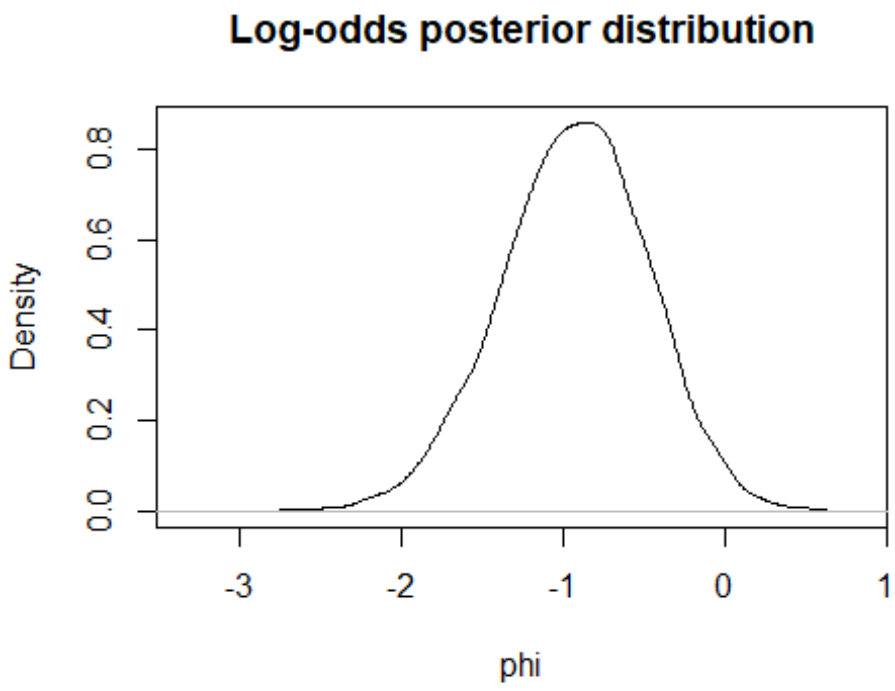
When similating nDraws = 10000 and counting the cases where the estimated probablity is bigger than 0,3 we see that the value is fairly close to the true value given the data y. Depening on the amount of draws we see that the posterior probaliblity for theta>0,3 given the data will be closer and closer to the true value.

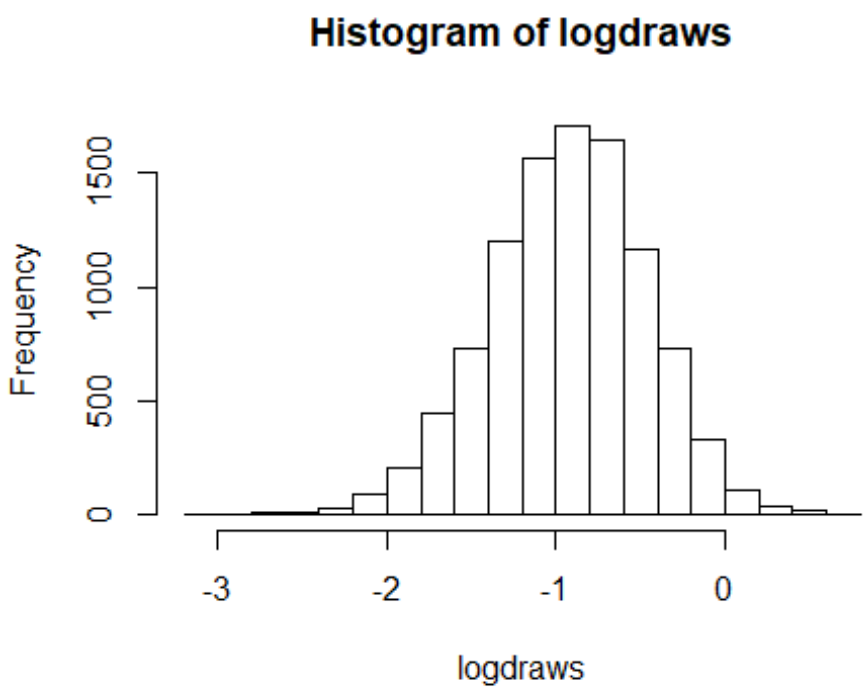
```
## [1] 0.4406
```

```
## [1] 0.4399472
```

c) Compute the posterior distribution of the log-odds phi = log(phu / (1-phi) by simulation (nDraws = 10000). [Hint: hist() and density() might come in handy]

The log-odds posterior distribution can be seen in the plot. It looks like the same distribution as the data of shown in histogram.


Log-odds posterior distribution
Density
0.0
0.2
0.4
0.6
0.8
-3
-2
-1
0
1
phi


Histogram of logdraws
Frequency
0
500
1000
1500
-3
-2
-1
0
logdraws

Assignment 2

Assume that you have asked 10 randomly selected persons about their monthly income (in thousands Swedish Krona) and obtained the following ten observations: 44, 25, 45, 52, 30, 63, 19, 50, 34 and 67. A common model for non-negative continuous variables is the log-normal distribution. The log-normal distribution $\log \mathcal{N}(\mu, \sigma^2)$ has density function

$$p(y|\mu, \sigma^2) = \frac{1}{y \cdot \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log y - \mu)^2\right],$$

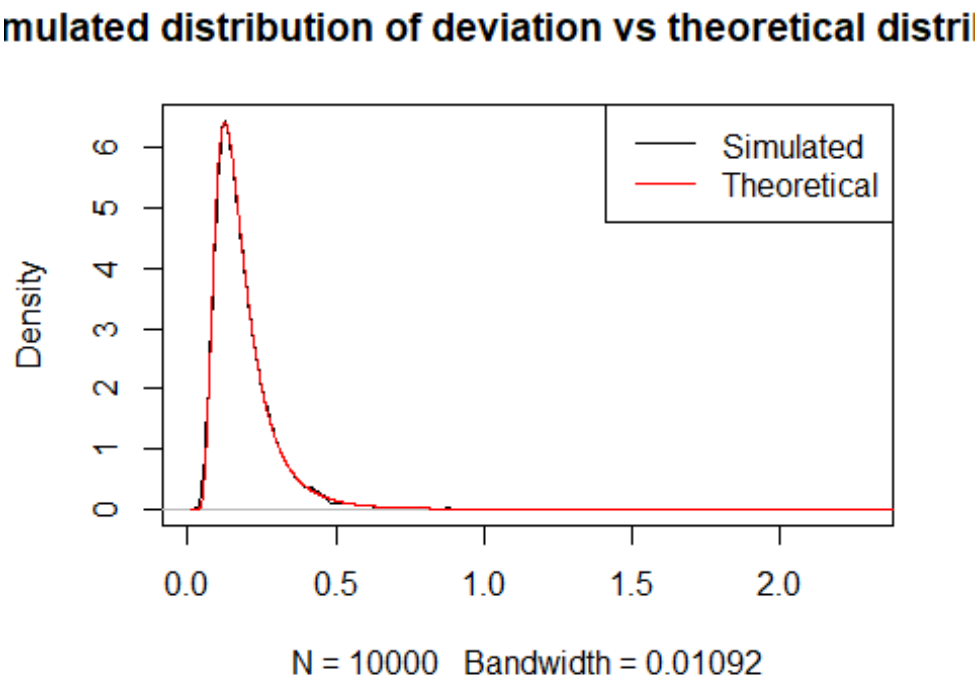
for $y > 0$, $\mu > 0$ and $\sigma^2 > 0$. The log-normal distribution is related to the normal distribution as follows: if $y \sim \log \mathcal{N}(\mu, \sigma^2)$ then $\log y \sim \mathcal{N}(\mu, \sigma^2)$. Let $y_1, ..., y_n|\mu, \sigma^2 \overset{iid}{\sim} \log \mathcal{N}(\mu, \sigma^2)$, where $\mu = 3.7$ is assumed to be known but σ^2 is unknown with non-informative prior $p(\sigma^2) \propto 1/\sigma^2$. The posterior for σ^2 is the $Inv - \chi^2(n, \tau^2)$ distribution, where

$$\tau^2 = \frac{\sum_{i=1}^{n}(\log y_i - \mu)^2}{n}.$$

Simulate $10,000$ draws from the posterior of σ^2 (assuming $\mu = 3.7$) and compare with the theoretical $Inv - \chi^2(n, \tau^2)$ posterior distribution.

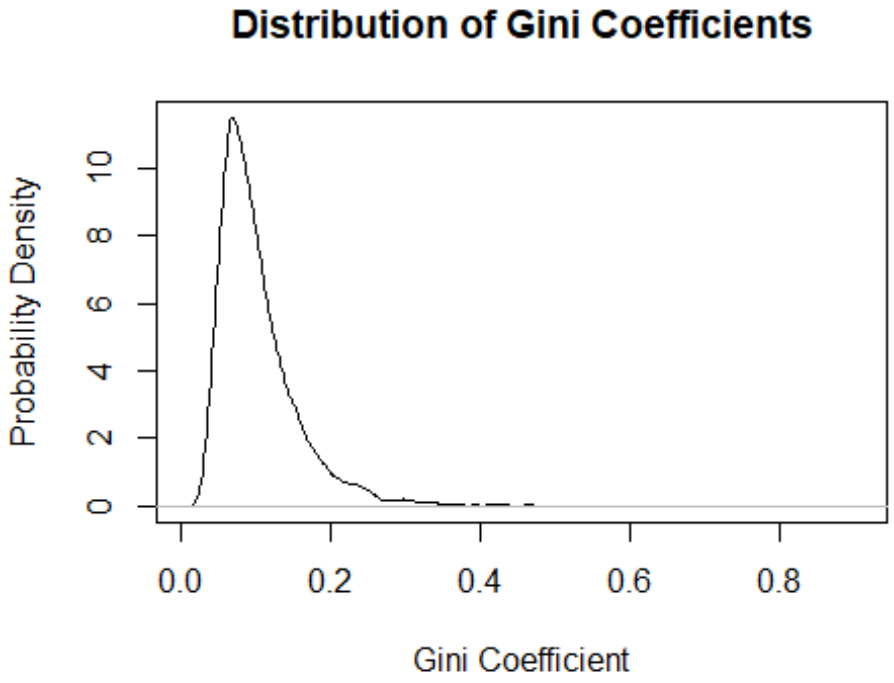
(2a)

We simulated a distribution from the posterior by drawing 10000 values and using density() to get the distribution. We are aware that density() only fits the data and does not yield the true distribution, but since n is so big (10000) we believe that density() will fit the model realisticly. Using density makes it easier to compare distribution than if we would use a histogram. We compared this distribution with the theoretical distribution and found that they were almost identical.



2b) Use the posterior draws in a) to compute the posterior distribution of the Gini coefficient G for the current data set.

The most common measure of income inequality is the Gini coefficient, G, where $0 \leq G \leq 1$. $G = 0$ means a completely equal income distribution, whereas $G = 1$ means complete income inequality. See Wikipedia for more information. It can be shown that $G = 2\Phi\left(\sigma/\sqrt{2}\right) - 1$ when incomes follow a $\log \mathcal{N}(\mu, \sigma^2)$ distribution. $\Phi(z)$ is the cumulative distribution function (CDF) for the standard normal distribution with mean zero and unit variance. Use the posterior draws in a) to compute the posterior distribution of the Gini coefficient G for the current data set.



When plotting the posterior distribution of the Gini Coefficients we can observe that it is very similar to the plot of the posterior distribution sigma^2. It is reasonable that the distribution is similar since the Gini Coeffiecients is just a transfromaion of the sigma^2, since the income has a conjugate prior which is scaled inverse chi-squared.

2c)

Use the posterior draws from b) to compute a 90% equal tail credible interval for G. A 90% equal tail interval (a, b) cuts off 5% percent of the posterior probability mass to the left of a, and 5% to the right of b. Also, do a kernel density estimate of the posterior of G using the density function in R with default settings, and use that kernel density estimate to compute a 90% Highest Posterior Density interval for G. Compare the two intervals

We obtained a credible interval of (0.046,0.21) and a highest posterior density interval of (0.33,0.176). We used the default kernel (Gaussian) with the default bandwidth for the HDI. In symmetric distributions, credible interval and HDI will return same results. With skewed distributions such as this one the HDI will move to the more probable values while credible

interval is limited to the 5 and 95 percentile. Since the HDI covers more probable values we believe that this interval seems to be a more intuitive and meaningful summary of the posterior.

```
## Warning: package 'HDInterval' was built under R version 3.6.3
```

Assignment 3

(3a) Plot the posterior distribution of κ for the wind direction data over a fine grid of κ value

The posterior distribution of k is obtained through the proportinality on likelihood vector of y|k multiplied by probability distribution p(k). Since the prior distribution of k is given the probability vector of k can be calculated. The likelihood vector of y|k is obtained by feeding the given likelihood function with the k values used to create the probability vector of k. By the final mltiplikation likelilhood(y|k)*p(k) we the get the posterior density for all values of k.

p(k|y) proportial to p(y|k)*p(k)

Bayesian inference for the concentration parameter in the von Mises distribution. This exercise is concerned with directional data. The point is to show you that the posterior distribution for somewhat weird models can be obtained by plotting it over a grid of values. The data points are observed wind directions at a given location on ten different days. The data are recorded in degrees:

$$(40, 303, 326, 285, 296, 314, 20, 308, 299, 296),$$

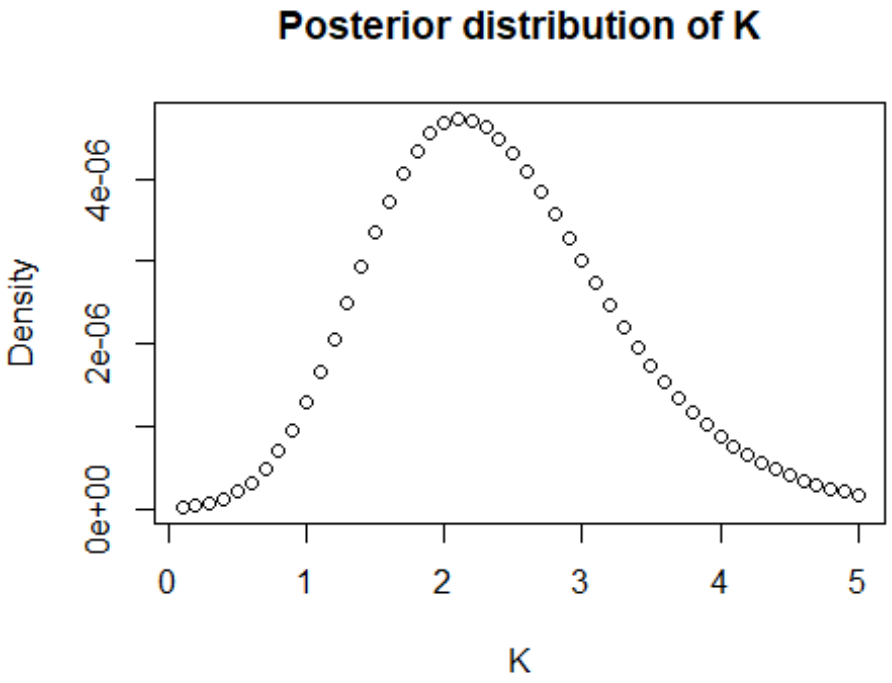
where North is located at zero degrees (see Figure 1 on the next page, where the angles are measured clockwise). To fit with Wikipedias description of probability distributions for circular data we convert the data into radians $-\pi \leq y \leq \pi$. The 10 observations in radians are

$$(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02).$$

Assume that these data points are independent observations following the von Mises distribution

$$p(y|\mu, \kappa) = \frac{\exp\left[\kappa \cdot \cos(y - \mu)\right]}{2\pi I_0(\kappa)}, \ -\pi \leq y \leq \pi,$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind of order zero [see `?besselI` in R]. The parameter μ $(-\pi \leq \mu \leq \pi)$ is the mean direction and $\kappa > 0$ is called the concentration parameter. Large κ gives a small variance around μ, and vice versa. Assume that μ is known to be 2.39. Let $\kappa \sim$ Exponential$(\lambda = 1)$ a priori, where λ is the rate parameter of the exponential distribution (so that the mean is $1/\lambda$).



(b)Find the(approximate) posterior mode of κ from the information in a)

By finding the maximum value of the density and extracting its index we can then get the k value which results in the highest probability

```
# Bernoulli ... again.
# Let y1, ..., yn|?? ??? Bern(??), and assume that you have obtained a sample
with s = 5
# successes in n = 20 trials. Assume a Beta(??0, ??0) prior for ?? and let
??0 = ??0 = 2.

# a)
# Draw random numbers from the posterior ??|y ??? Beta(??0 + s, ??0 + f), y =
(y1, . . . , yn),
# and verify graphically that the posterior mean and standard deviation
# converges to the true values as the number of random draws grows large.

vec = c()
deviation = c()
number = c()

# Prior
a = b = 2
s = 5
n = 20
f = n-s

# Posterior
posterior_alpha = s + a
posterior_beta = b + f

for (i in 1:1000){
  post = rbeta (i, posterior_alpha, posterior_beta)
  mean =mean(post)
  st = sd(post)
  vec = append(vec,mean)
  deviation = append(deviation,st)
  number = append(number, i)
}

# Plot theta means
plot(number, vec, type='l', xlab='Number of draws', ylab='Mean of theta')

# Plot theta deviations
plot(deviation)
plot(number, deviation, type='l', xlab='Number of draws', ylab='deviation')

# b)
# Use simulation (nDraws = 10000) to compute the posterior probability
# Pr(?? > 0.3|y) and compare with the exact value [Hint: pbeta()].
```

```r
draws = rbeta (10000, posterior_alpha, posterior_beta)
bigger_than = ifelse(draws>0.3,1,0)
prob_bigger_than = sum(bigger_than)/length(draws)
prob_bigger_than


theta_bigger_than = pbeta(0.3, posterior_alpha, posterior_beta, ncp = 0,
lower.tail = FALSE, log.p = FALSE)
print(theta_bigger_than)

# c)
# Compute the posterior distribution of the log-odds ?? = log(?? / 1?????)
# by simulation (nDraws = 10000). [Hint: hist() and density() might come in
handy]

logdraws = log(draws/(1-draws))
densityLogDraws = density(logdraws)
plot(densityLogDraws,
     main = "Log-odds posterior distribution",
     ylab = "Density",
     xlab = "phi")

hist(logdraws)

## ASSIGNEMNT 2

# Log-normal distribution and the Gini coefficient.
# Assume that you have asked 10 randomly selected persons about their monthly
income
# (in thousands Swedish Krona) and obtained the following ten observations:
44,
# 25, 45, 25, 30, 33, 19, 50, 34 and 67. A common model for non-negative
continuous
# variables is the log-normal distribution

# a)
# Simulate 10, 000 draws from the posterior of ??2
# (assuming μ = 3.7) and compare with the theoretical
# Inv ??? ??2(n, ?? 2) posterior distribution.

my = 3.7
n = 10

Y = c(44,25,45,52,30,63,19,50,34,67)
#compute sample variance s^2
T2 =  sum((log(Y)-my)^2)/n

#Draw X from chi2(n) (This is a draw from Inv- X^2(n,s^2))
```

```r
set.seed(12345)
XposteriorDraw = rchisq(10000,10)

# get deviation^2 from X ( deviation^2 = df*s^2 /X)
deviationPostDraw = 10*T2/XposteriorDraw
distDeviationPostDraw = density(deviationPostDraw)

library(invgamma)

# function for scaled inverse chi-squared pdf
invscaledchi2 <- function(x, df, tao2) {
  a <- df / 2
  ((n*tao2 / 2)^a)/gamma(a) * x^(-a-1) * exp(-(n*tao2 / 2)/x)
}


# sequence of x-values to illustrate the distribution
xrange = seq(0.01,3.0,0.001)

# values from inverse chisquared mapped on x-range
deviations = invscaledchi2(xrange,10,T2)

# plot simulated distribution with theoretical distribution
plot(distDeviationPostDraw)
lines(xrange, deviations, type="l",col="red")


# b)
# Use
#the posterior draws in a) to compute the posterior distribution of the Gini
#coefficient G for the current data se

#Use draws from posterior to form GiniCoeffient
Gini_coefficients = 2*pnorm(deviationPostDraw/sqrt(2),0,1)-1
Gini_density = density(Gini_coefficients)
plot(Gini_density, main = "Distribution of Gini Coefficients", xlab = "Gini
Coefficient", ylab = "Probability Density")

#c
#Use the posterior draws from b) to compute a 90% equal tail credible
interval
#for G. A 90% equal tail interval (a, b) cuts off 5% percent of the posterior
#probability mass to the left of a, and 5% to the right of b.

GiniQuantiles = quantile(Gini_coefficients, probs=seq(0,1,0.05))
Crediblerange = c(GiniQuantiles[2],GiniQuantiles[20])
# gets range (0.046,0.21)

#Also, do a kernel density estimate of the posterior of G using the density
```

```r
function in R with
#default settings, and use that kernel density estimate to compute a 90% Highest
#Posterior Density interval for G. Compare the two intervals

# using library from CRAN to get highest (posterior) density interval
library(HDInterval)
hdiRange = hdi(Gini_density, credMass=0.9)
# gets range (0.033,0.176)

hist(Gini_coefficients)
plot(Gini_density)


# Assignment 3
# a)
# Plot the posterior distribution of ?? for the wind direction data over a fine grid
# of ?? values


y = c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)
my = 2.39
lambda = 1

k_prior = function(lambda, k){
  lambda*exp(-lambda*k)
}

likelihood_function = function(k, y, my){
  prod(exp(k*cos(y-my))/(2*pi*besselI(k, 0)))
}

sequence = seq(0.1, 5, 0.1)

k_prior_vector = k_prior(lambda, sequence)

likelihood_vector = c()
test = c()
for (k in sequence){
  likelihood_vector = append(likelihood_vector, likelihood_function(k, y, my))
}

k_posterior = likelihood_vector*k_prior_vector
plot(sequence, k_posterior, main = "Posterior distribution of K", xlab = "K", ylab = "Density" )
```

```r
# b)
# Find the approximate  posterior mode of k  from the information in a

k_index = which(k_posterior == max(k_posterior))
k_mode = sequence[k_index]
```