732A91 - Lab 1

Joris van Doorn || Weng Hang Wong

17 April 2020

1. Bernoulli . . . again.

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

a.

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

The likelihood is given as:

$$y_1, ..., y_n|\theta \sim Bern(\theta) = \theta^s(1-\theta)^f$$

Where s is the number of succes (s = 5) and f = n - s is the number of failures (n = 20; f = 20 - 5 = 15).

We assume a beta prior:

$$Beta(\alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}$$

Where $\alpha_0 = \beta_0 = 2$.

We find the posterior using Bayes Theorem:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Combining the previous formulas gives the following expression for the posterior:

$$p(\theta|y) \propto \theta^s(1-\theta)^f\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} = \theta^{s+\alpha_0 - 1}(1-\theta)^{f+\beta_0 - 1}$$

Filling in the information we know about s, f, α_0 and β_0, we get:

$$\theta|y_1, ..., y_20 \sim Beta(s + \alpha_0, f + \beta_0) = Beta(7, 17)$$

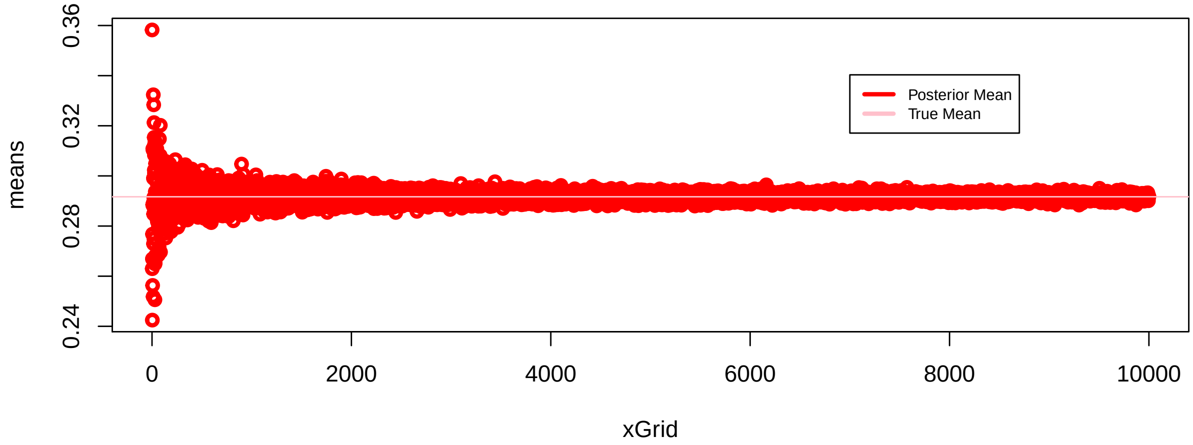
Verification will be done by comparing with the true mean and variance, which for a Beta distribution is given by:

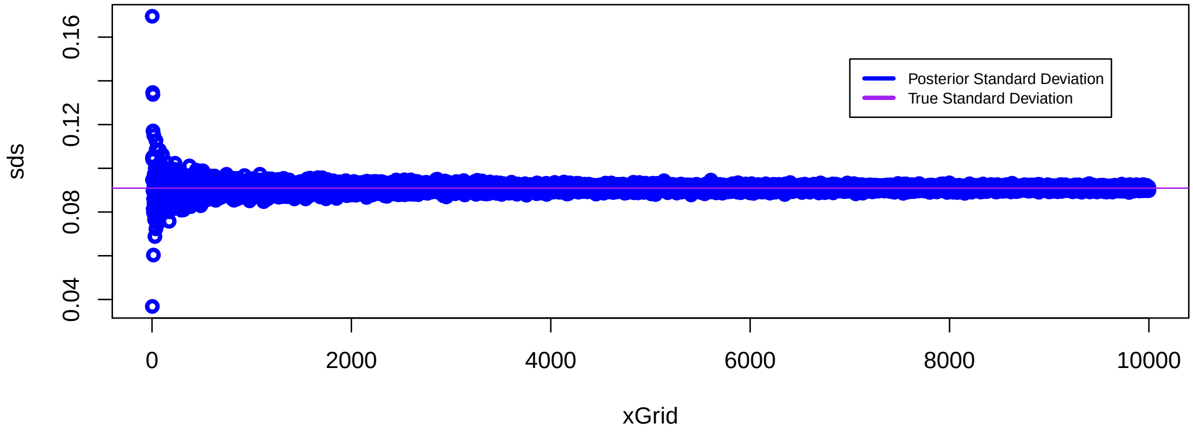
$$\mu = \frac{\alpha}{\alpha + \beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

```
## Number of Draws:  10000
## Posterior Mean:  0.291771
## Posterior Standard Deviation:  0.08982719
```

```
##
## True Mean:  0.2916667
## True Standard Deviation:  0.09090593
```





b.

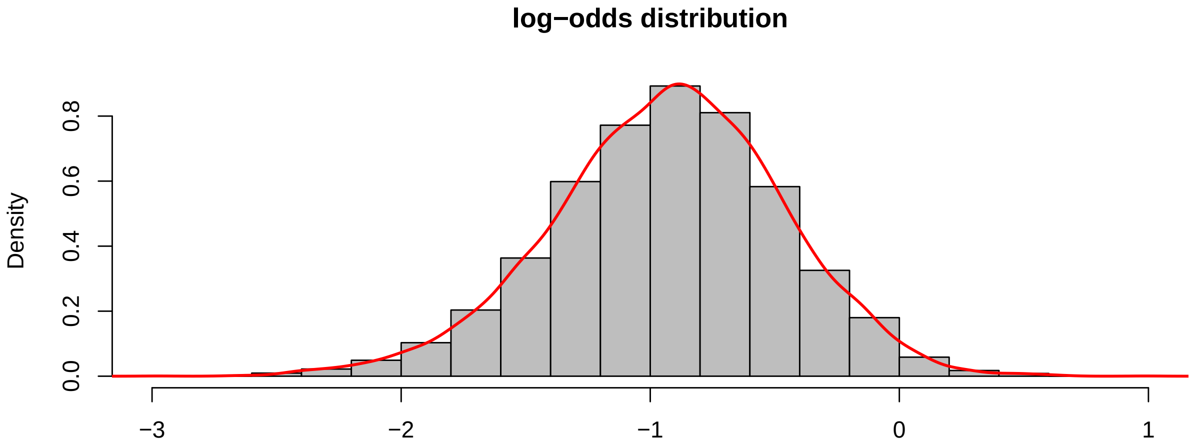
Use simulation (nDraws = 10000) to compute the posterior probability $Pr(\theta > 0.3|y)$ *and compare with the exact value*

```
##
## The probability of theta being larger than 0.3 is:  0.4392
```

```
##
## The true probability of theta being larger than 0.3 is:  0.4399472
```

c.

Compute the posterior distribution of the log-odds $\phi = log(\frac{\theta}{1-\theta})$ *by simulation (nDraws = 10000).*



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, 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 $log(N(\mu, \sigma^2))$ *has density function:*

$$p(y|\mu, \sigma^2) = \frac{1}{y * \sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(log(y)-\mu)^2}$$

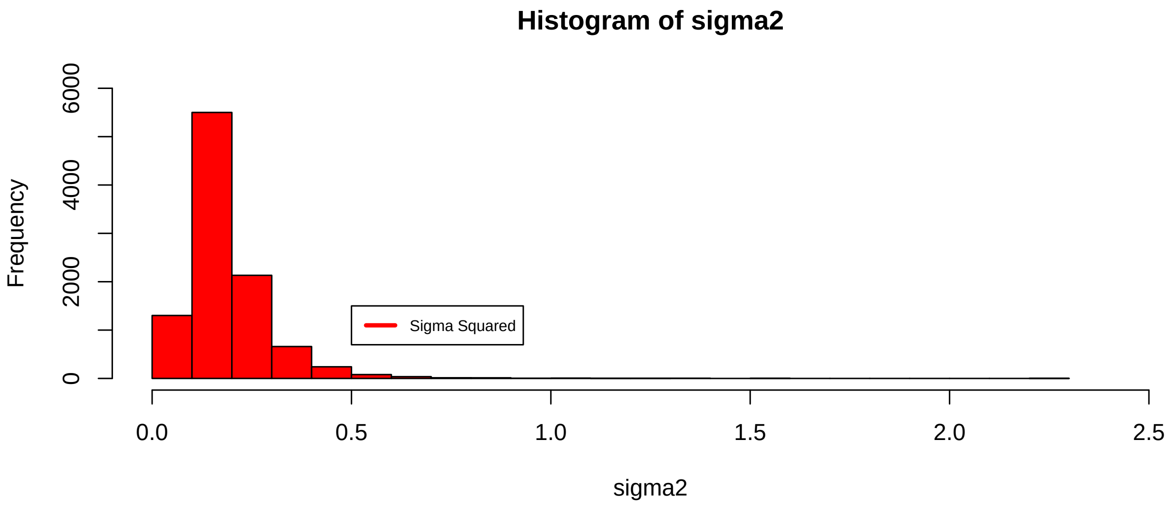
For y>0, $\mu > 0 and \sigma^2 > 0$. *The log-normal distribution is related to the normal distribution as follows: if y* $logN(\mu, \sigma^2)$ *then log(y)* $N(\mu, \sigma^2)$. *Let* $y_1, ..., y_n|\mu, \sigma^2 \, ^{iid} log(N(\mu, \sigma^2))$, *where* $\mu = 3.7$ *is assumed to be known but* σ^2 *is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

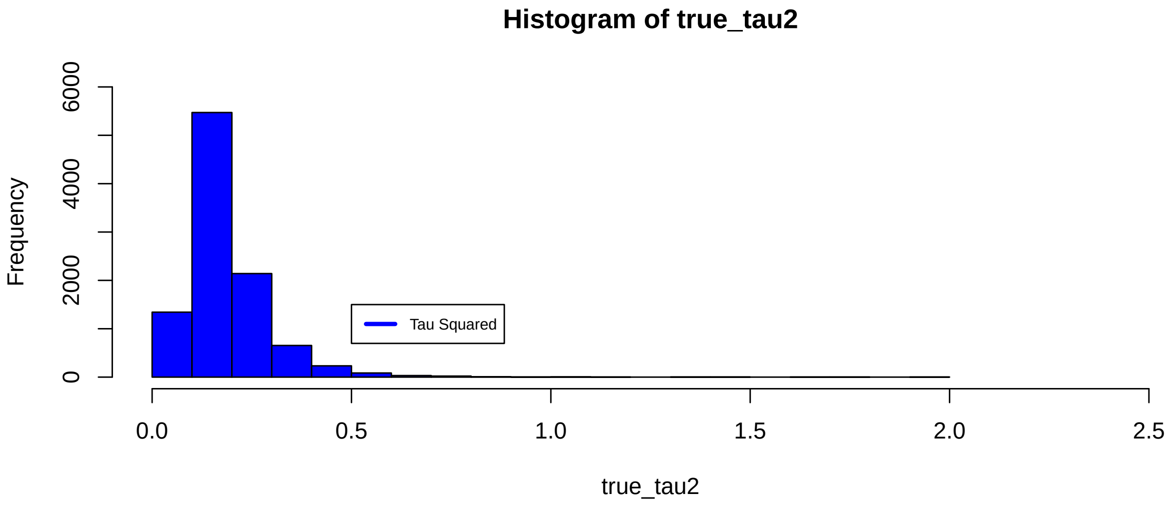
$$\tau^2 = \frac{\sum_{i=1}^{n}(log(y_i) - \mu)^2}{n}$$

a.

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.*

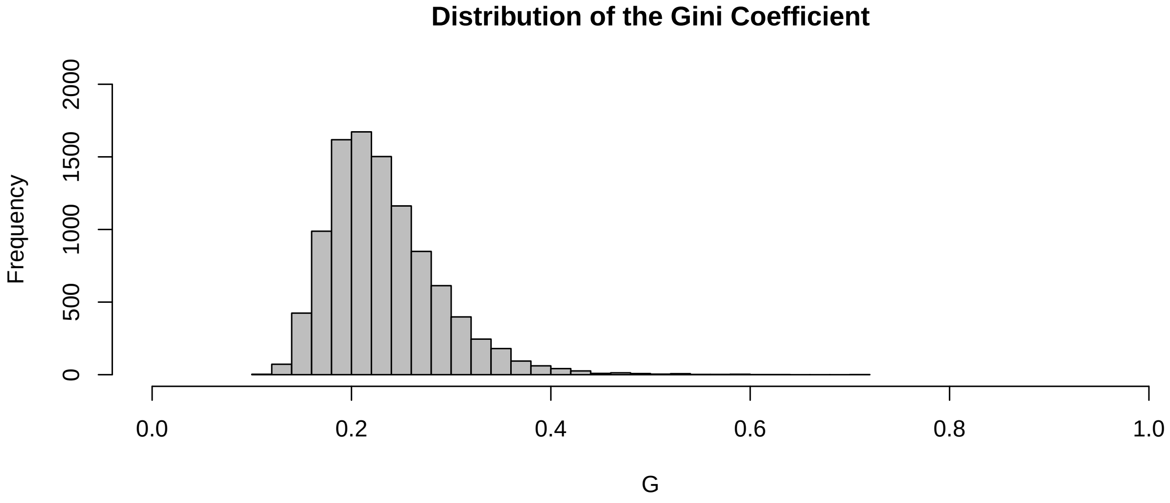
We worked with the log of the data, assuming that is approximately normally distributed with $N(\mu, \sigma^2)$





b.

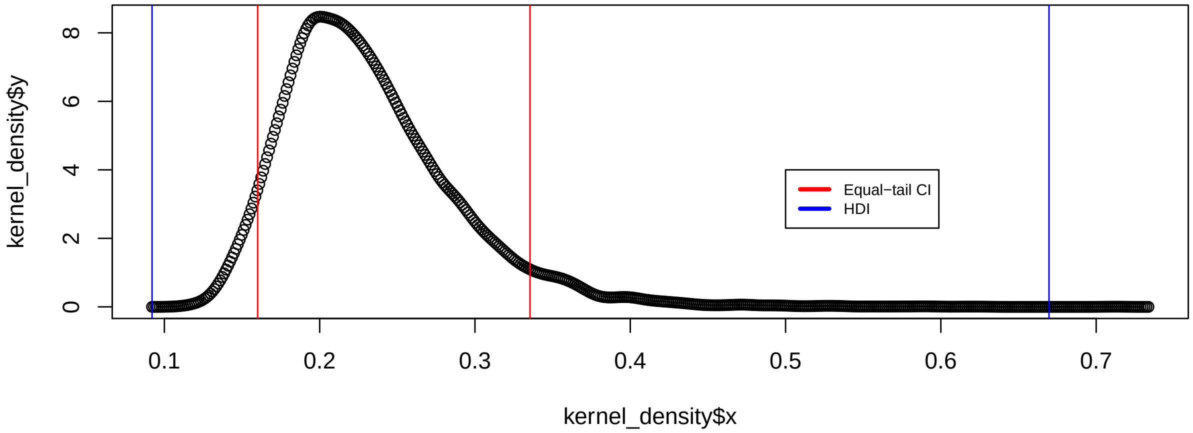
The most common measure of income inequality is the Gini coefficient, G, where $0 \leq G \leq 1$. *G = 0 means a completely equal income distribution, whereas G = 1 means complete income inequality. See Wikipedia for more information. It can be shown that* $G = 2\phi(\frac{\sigma}{\sqrt{2}} - 1)$ *when incomes follow a* $logN(\mu, \sigma^2)$ *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.*



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. 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.

```
## The equal-tail credible interval for the Gini coefficient is: [ 0.1600229 ,  0.3354106 ]
##
## The Highest Posterior Density interval for the Gini coefficient is: [ 0.09207026 ,  0.6696239 ]
```



We estimated both the equal tail credible interval and the Highest Posterior Density interval (HPD) for G. As can be seen from the plot, the HPD is much wider than the equal tail credible interval. This interval is likely much wider because it contains the θ values with the highest pdf, whereas the equal tail CI exactly cuts off 5% on each side.

3. 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:*

$$p(y|\mu, \kappa) = \frac{e^{\kappa * cos(y-\mu)}}{2\pi I_0(\kappa)}, -\pi \leq y \leq \pi$$

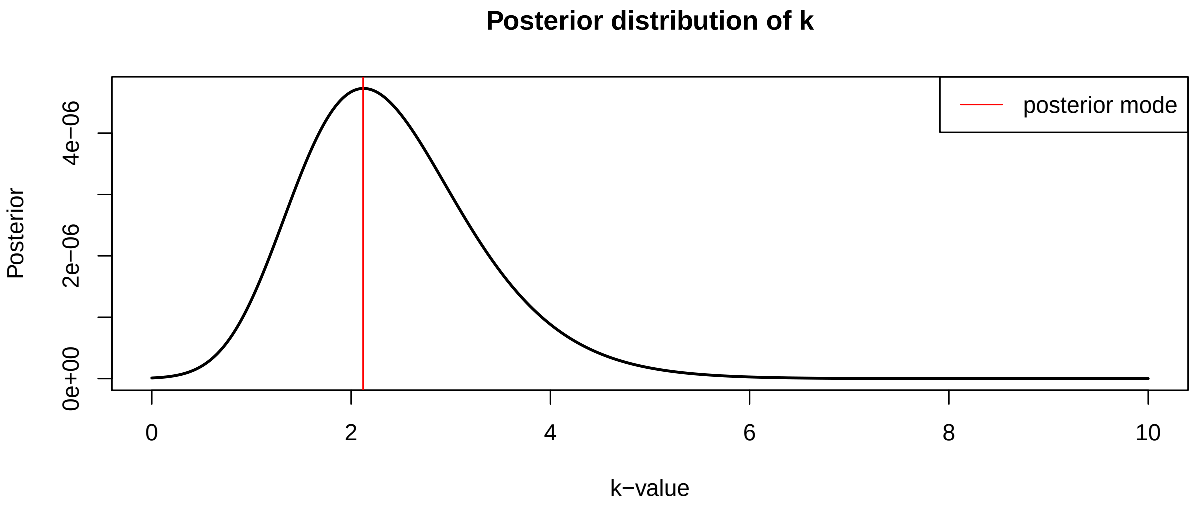
where $I_0(\kappa)$ *is the modified Bessel function of the first kind of order zero. The parameter* $\mu(-\pi \leq y \leq \pi)$ *is the mean direction and* $\kappa > 0$ *is called the concentration parameter. Large* κ *gives a small variance around* μ*, and vice versa. Assume that* μ *is known to be 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* $1/\lambda$*.*

a.

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

Again, using Bayes Theorem we find:

$$p(\mu, \kappa|y) \propto \frac{exp[\kappa \cdot (\Sigma_{i=1}^{n} \cos(y_i - \mu) - 1)]}{(2\pi I_0(\kappa))^n}$$



```
## The posterior mode of k is at:  2.12
## With value:  4.727615e-06
```

Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width=9, fig.height = 4.1)
library(tidyverse)
library(dplyr)
library(knitr)
library(LaplacesDemon)
library(bayestestR)
library(HDInterval)
set.seed(12345)
# ---------------------------------------------------------
# 1a
f = 17
s = 7
nDraws = 10000

# Sampler from the posterior
set.seed(12345)
beta_sampler <- function(nDraws, alpha, beta){
  draws <- rbeta(nDraws,alpha,beta)
  return(list('Mean'=mean(draws),'sd'=sd(draws)))
}


# Creating lists of mean and std
means <- c()
sds <- c()
for(i in 1:nDraws){
  means[i] <- beta_sampler(i, s, f)$Mean
  sds[i] <- beta_sampler(i, s, f)$sd
}

true_mean = s/(s+f)
true_sd =  sqrt((s*f) / (((s+f)^2)*(s+f+1)))

cat("Number of Draws: ", nDraws , "\nPosterior Mean: ", means[nDraws], "\nPosterior Standard Deviation:
cat("\nTrue Mean: ", true_mean, "\nTrue Standard Deviation: ", true_sd)

xGrid <- 1:nDraws

plot(xGrid, means, type = 'p', lwd = 3, col = "red")
abline(h=true_mean, col="pink")
legend(x = max(xGrid)*0.70, y = 0.95*max(means), legend = c("Posterior Mean", "True Mean"), col = c("re

plot(xGrid, sds, type = 'p', lwd = 3, col = "blue")
abline(h=true_sd, col="purple")
legend(x = max(xGrid)*0.70, y = 0.15, legend = c("Posterior Standard Deviation","True Standard Deviatio
# ---------------------------------------------------------
# 1b
set.seed(12345)
posterior_sample <- rbeta(nDraws, s, f)
```

```
prob_03 <- posterior_sample[posterior_sample > 0.3]
prob <- length(prob_03)/nDraws

true_prob <- pbeta(q=0.3, s, f, lower.tail = FALSE)

cat("\nThe probability of theta being larger than 0.3 is: ", prob)
cat("\nThe true probability of theta being larger than 0.3 is: ", true_prob)
# ---------------------------------------------------------
# 1c
log_odds <- log(posterior_sample/(1-posterior_sample))
hist(log_odds, freq = F, main="log-odds distribution", xlab="",col="grey")
lines(density(log_odds), col="red", lwd=2)
# ---------------------------------------------------------
# 2a
y <- c(44,25,45,52,30,63,19,50,34,67)
n <- length(y)
nDraws <- 10000
mu <- 3.7

tau2 <- sum((log(y)-mu)^2) / n

set.seed(12345)
sigma2 <- c()
for(i in 1:nDraws){
  sample <- rchisq(n=1, df=n)
  sigma2[i] <- (tau2*n)/sample
}

true_tau2 <- rinvchisq(nDraws, n, tau2)

hist(sigma2,breaks=25,xlim=c(0,2.5),ylim=c(0,6000),col="red")
legend(x = 0.5, y = 1500, legend = "Sigma Squared", col = "red", lwd = c(3,3), cex = 0.7)

hist(true_tau2,breaks=25,xlim=c(0,2.5),ylim=c(0,6000),col="blue")
legend(x = 0.5, y = 1500, legend = "Tau Squared", col = "blue", lwd = c(3,3), cex = 0.7)
# ---------------------------------------------------------
# 2b
# Calculating gini coefficient G
G <- 2*pnorm(sqrt(sigma2/2))-1

hist(G,breaks=25,xlim=c(0,1),ylim=c(0,2000),col="grey",main="Distribution of the Gini Coefficient")
# ---------------------------------------------------------
# 2c
et_cred_int <- ci(G, ci = 0.90, method = "ETI")
cat("The equal-tail credible interval for the Gini coefficient is: [", et_cred_int$CI_low, ", ", et_cred

kernel_density <- density(G)
HDI <- hdi(kernel_density$x, credMass = 0.90)

hdi_low <- as.numeric(HDI[1])
hdi_high <- as.numeric(HDI[2])

cat("\nThe Highest Posterior Density interval for the Gini coefficient is: [", hdi_low, ", ", hdi_high,
plot(kernel_density$y~kernel_density$x)
abline(v=c(hdi_low,hdi_high), col=c("blue","blue"))
abline(v=c(et_cred_int$CI_low, et_cred_int$CI_high), col=c("red", "red"))
legend(x = 0.5, y = 4, legend = c("Equal-tail CI", "HDI"), col = c("red","blue"), lwd = c(3,3), cex = 0
y <- c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)
n <- length(y)
mu <- 2.39
kGrid <- seq(0,10,0.01)
posterior <- exp(kGrid*(sum(cos(y-mu))-1))/((2*pi*besselI(kGrid,0))^n)

plot(kGrid,posterior, type="l",main="Posterior distribution of k", xlab="k-value",ylab="Posterior", lwd=
abline(v=kGrid[which.max(posterior)], col="red")
legend("topright",legend="posterior mode", lty=1,col="red")

cat("The posterior mode of k is at: ", kGrid[which.max(posterior)], "\nWith value: ", max(posterior))
```