Bayesian Learning

2024-04-10

Computer Lab 1

Hugo Morvan (hugmo418), Xinyuan Zhai (xinzh539) (dropped course but contributed to exercise 1 before dropping).

```r
library(ggplot2)
```

1. Daniel Bernoulli

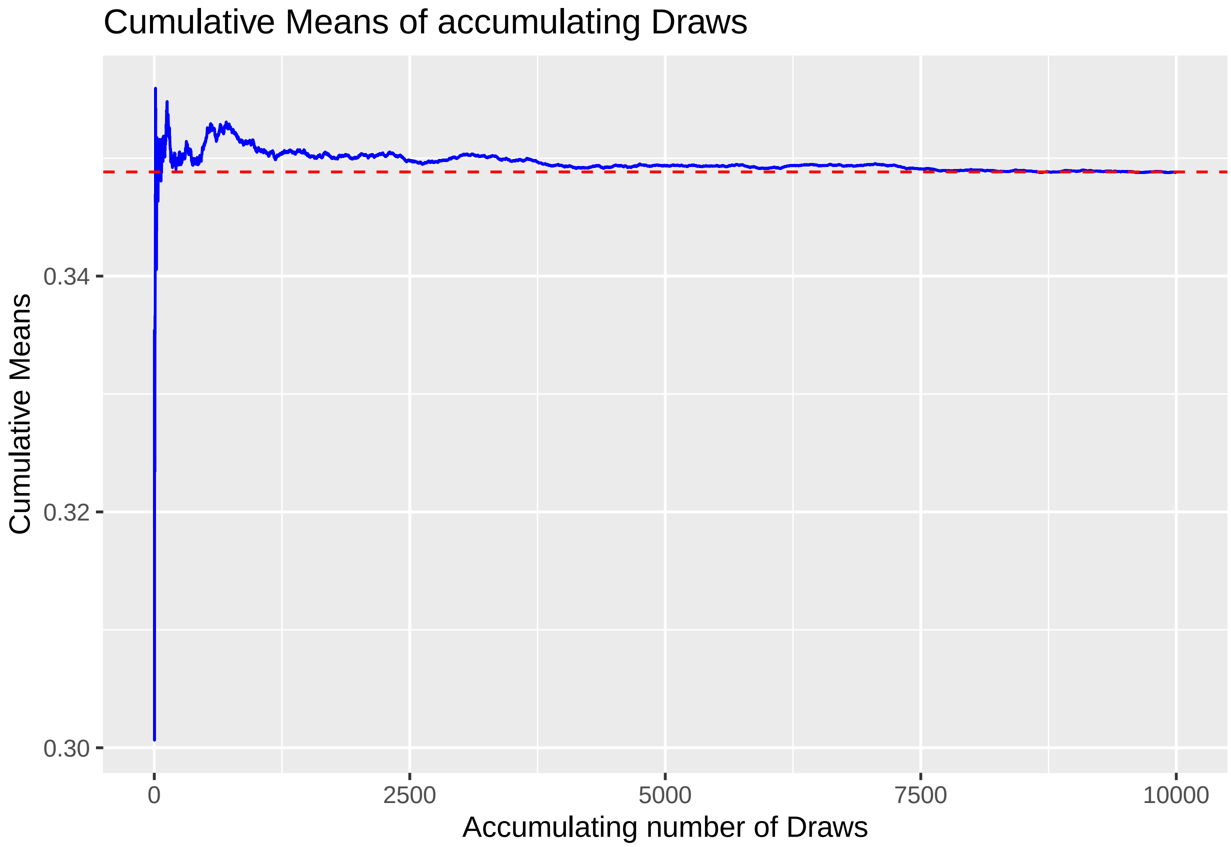
Let $y_1, ..., y_n|\theta \sim Bern(\theta)$, and assume that you have obtained a sample with s = 22 successes in n = 70 trials. Assume a Beta(α_0, β_0) prior for θ and let $\alpha_0 = \beta_0 = 8$.

1.a

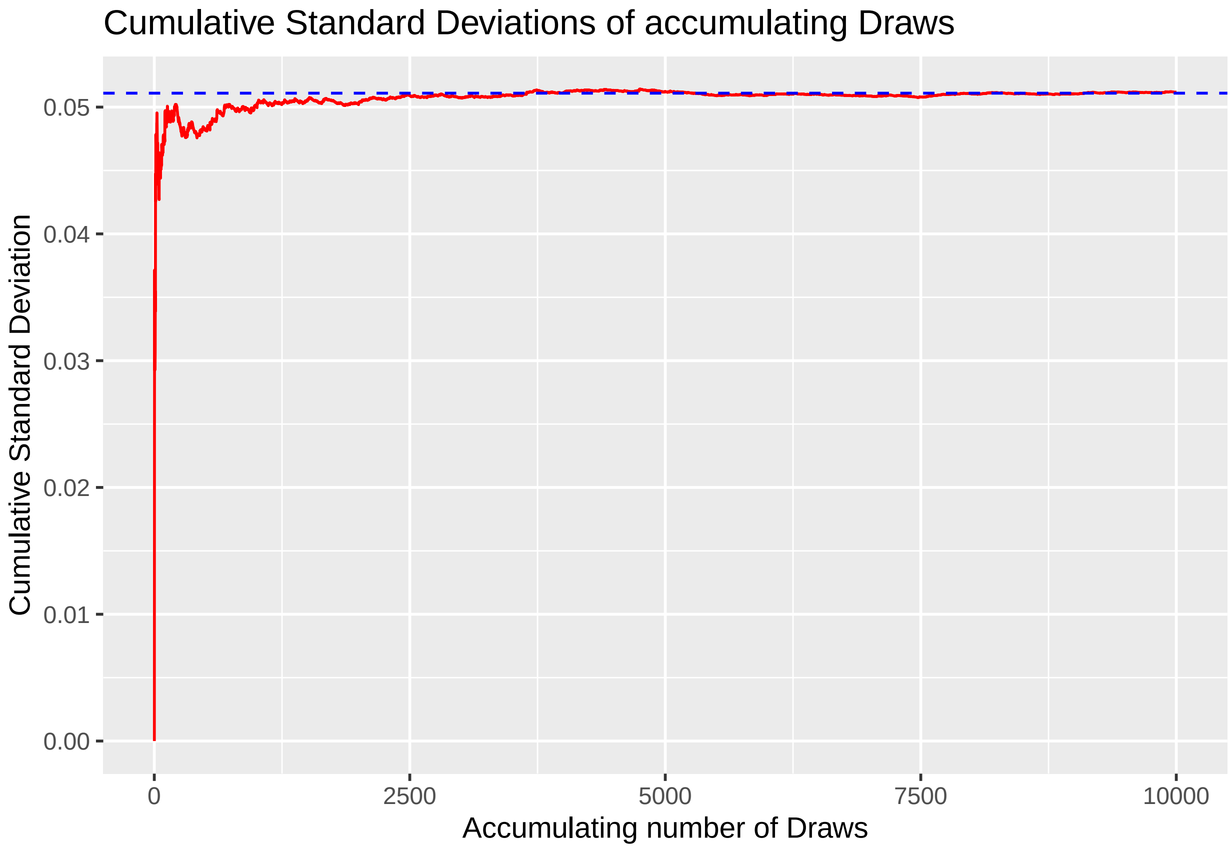
Draw 10000 random values (nDraws = 10000) from the posterior $\theta|y \sim Beta(\alpha_0 + s, \beta_0 + f)$, where $y = (y_1, ..., y_n)$, and verify graphically that the posterior mean $E[\theta|y]$ and standard deviation $SD[\theta|y]$ converges to the true values as the number of random draws grows large. [Hint: use rbeta() to draw random values and make graphs of the sample means and standard deviations of θ as a function of the accumulating number of drawn values].

```r
s <- 22
n <- 70
f <- n-s
nDraws <- 10000
alpha_0 <- 8
beta_0 <- 8

post_alpha <- alpha_0 + s
post_beta <- beta_0 + f

# rbeta drawing
delta <- rbeta(nDraws, post_alpha, post_beta)

# true values of mean and sd
mean_true <- post_alpha/(post_alpha + post_beta)
sd_true <- sqrt((post_alpha*post_beta)/((post_alpha+post_beta+1)*(post_alpha+post_beta)^2))

# sample means based on accumulating number of drawn values
cumu_mean <- cumsum(delta)/(1:nDraws)
cumu_sd <- sqrt(cumsum((delta - cumu_mean)^2) / (1:nDraws))
```

```r
plot_data <- data.frame(
  Draw = 1:nDraws,
  Mean = cumu_mean,
  SD = cumu_sd
)

# Mean plot
cumulativeMeans_plot <- ggplot(plot_data, aes(x = Draw)) +
  geom_line(aes(y = Mean), color = "blue") +
  # true mean line
  geom_hline(yintercept = mean_true, color = "red", linetype = "dashed") +
  ggtitle("Cumulative Means of accumulating Draws") +
  xlab("Accumulating number of Draws") +
  ylab("Cumulative Means")

# SD plot
cumulativeSD_plot <- ggplot(plot_data, aes(x = Draw)) +
  geom_line(aes(y = SD), color = "red") +
  # true standard deviation line
  geom_hline(yintercept = sd_true, color = "blue", linetype = "dashed") +
  ggtitle("Cumulative Standard Deviations of accumulating Draws") +
  xlab("Accumulating number of Draws") +
  ylab("Cumulative Standard Deviation")

cumulativeMeans_plot
```



```
cumulativeSD_plot
```



1.b

Draw 10000 random values from the posterior to compute the posterior probability $Pr(\theta > 0.3|y)$ and compare with the exact value from the Beta posterior. [Hint: use pbeta()]

```
# posterior probability:
p_post <- mean(delta > 0.3)

# beta posterior probability:
p_beta_post <- 1-(pbeta(0.3, post_alpha, post_beta))

p_post
```

```
## [1] 0.83041
```

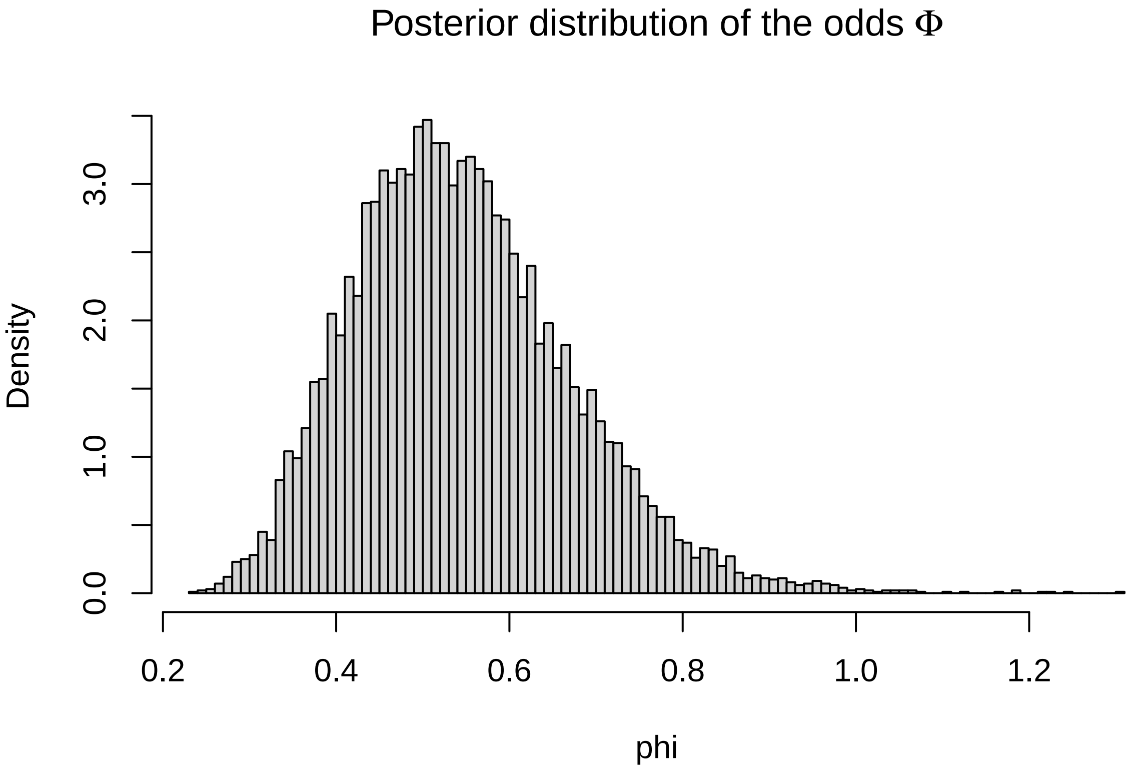
```
p_beta_post
```

```
## [1] 0.8285936
```

1.c

Draw 10000 random values from the posterior of the odds $\phi = \frac{\theta}{1-\theta}$ by using the previous random draws from the Beta posterior for θ and plot the posterior distribution of ϕ. [Hint: hist() and density() can be utilized].

```
phi <- delta/(1-delta)

hist(phi, breaks=100, prob=TRUE, xlab="phi", main=expression(paste("Posterior distribution of the odds
```



2. Log-normal distribution and Gini coefficient

Assume that you have asked 8 randomly selected persons about their monthly income (in thousands Swedish Krona) and obtained the following eight observations: 33, 24, 48, 32, 55, 74, 23, and 17. A common model for non-negative continuous variables is the log-normal distribution. The log-normal distribution $logN(\mu, \sigma^2)$ has density function

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

, where y > 0, $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The log-normal distribution is related to the normal distribution as follows: if $y \sim logN(\mu, \sigma^2)$ then $logy \sim N(\mu, \sigma^2)$. Let $y_1, ..., y_n|\mu,\ \sigma^2 iid \sim logN(\mu, \sigma^2)$, where $\mu = 3.6$ 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 $\tau^2 = \frac{\sum_{i=1}^{n}(\log y_i - \mu)^2}{n}$.

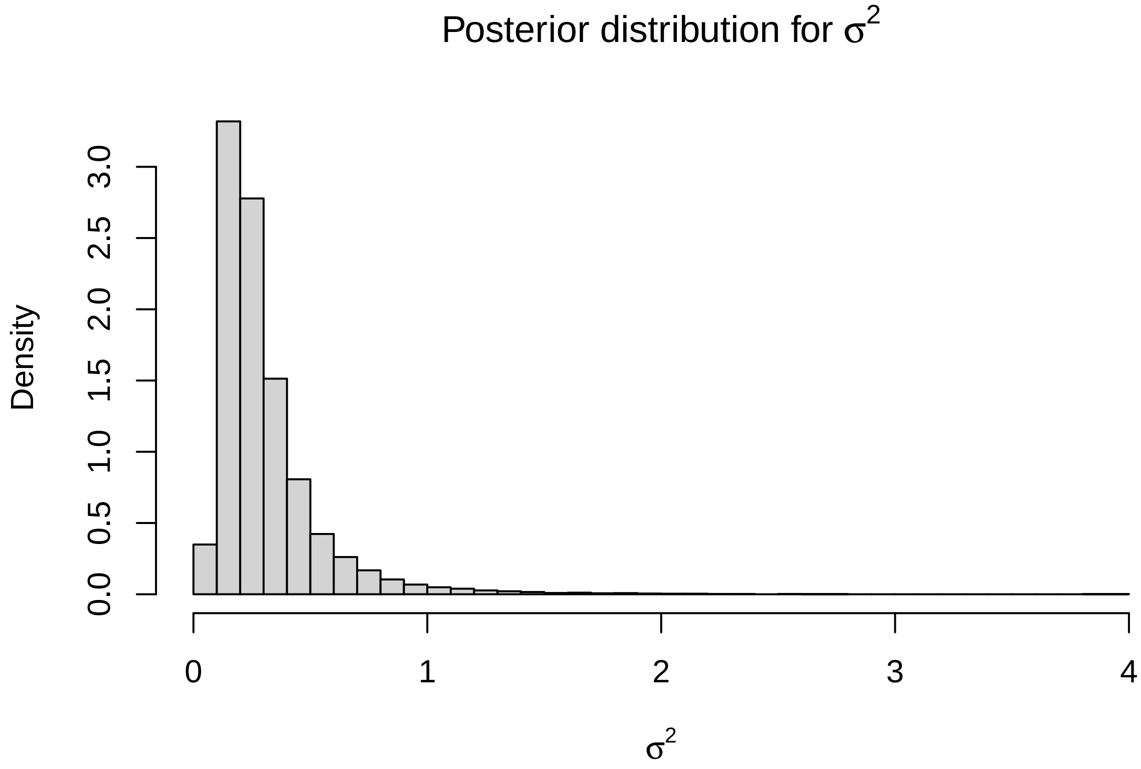
2.a

Draw 10000 random values from the posterior of σ^2 by assuming $\mu = 3.6$ and plot the posterior distribution.

To simulate from the posterior distribution, we use the steps indicated in lecture 3:

1. Draw $X \sim \chi^2(n-1)$
2. Compute $\sigma^2 = \frac{(n-1)s^2}{X}$ (a draw from $Inv - \chi^2(n-1, s^2)$)
3. Draw a θ from $N(\bar{x}, \frac{\sigma^2}{n})$ conditional on the previous σ^2.
4. Repeat step 1-3 many times.

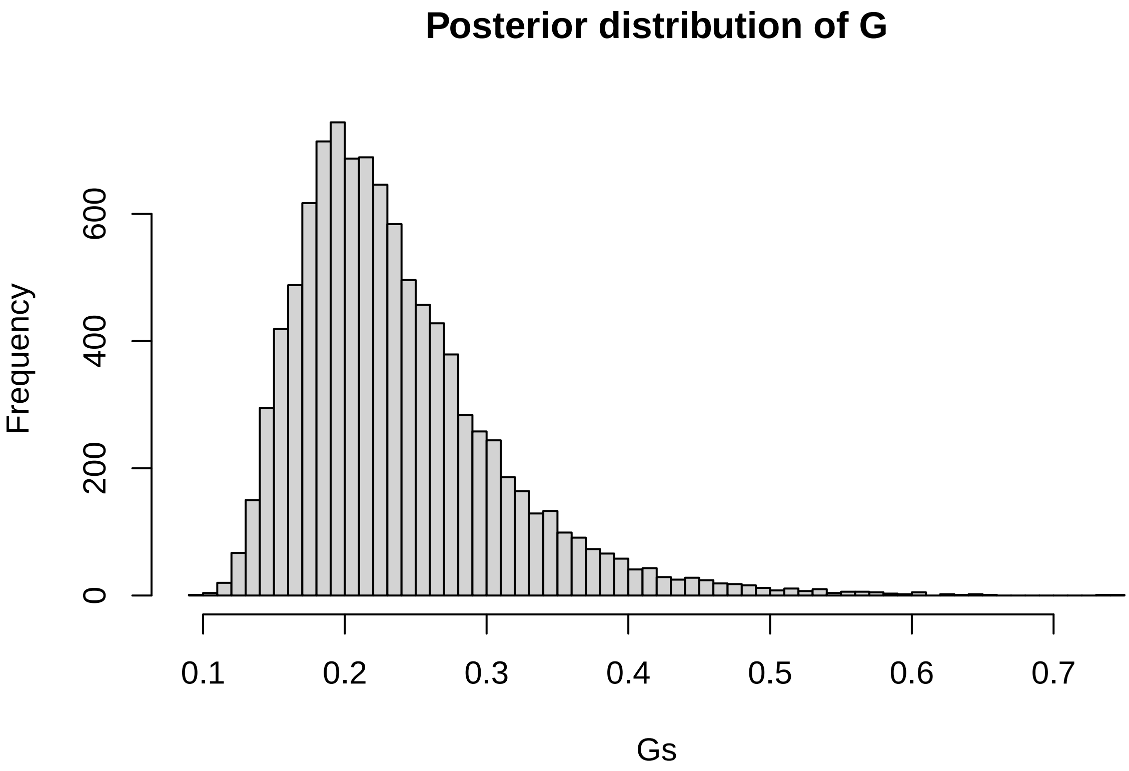
```
Y = c(33, 24, 48, 32, 55, 74, 23, 17)
n = length(Y)
df = n-1

mu = 3.6
tau_sq = (sum((log(Y)-mu)^2))/n #s^2 in the lecture

sigsq_draws = (df*tau_sq) / rchisq(10000, df)

hist(sigsq_draws, prob=TRUE, xlab = expression(sigma^2),
     main=expression(paste("Posterior distribution for ", sigma^2)),
     breaks = 50)
```



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 e.g. Wikipedia for more information about the Gini coefficient). It can be shown that $G = 2\Phi(\sigma/\sqrt{2}) - 1$ when incomes follow a $logN(\mu, \sigma^2)$ distribution. $\Phi(z)$ is the cumulative distribution function function (CDF) for the standard normal distribution with mean zero and unit variance. Use the posterior draws in a) to compute the posterior distributiuon of the Gini coefficient G for the current data set.

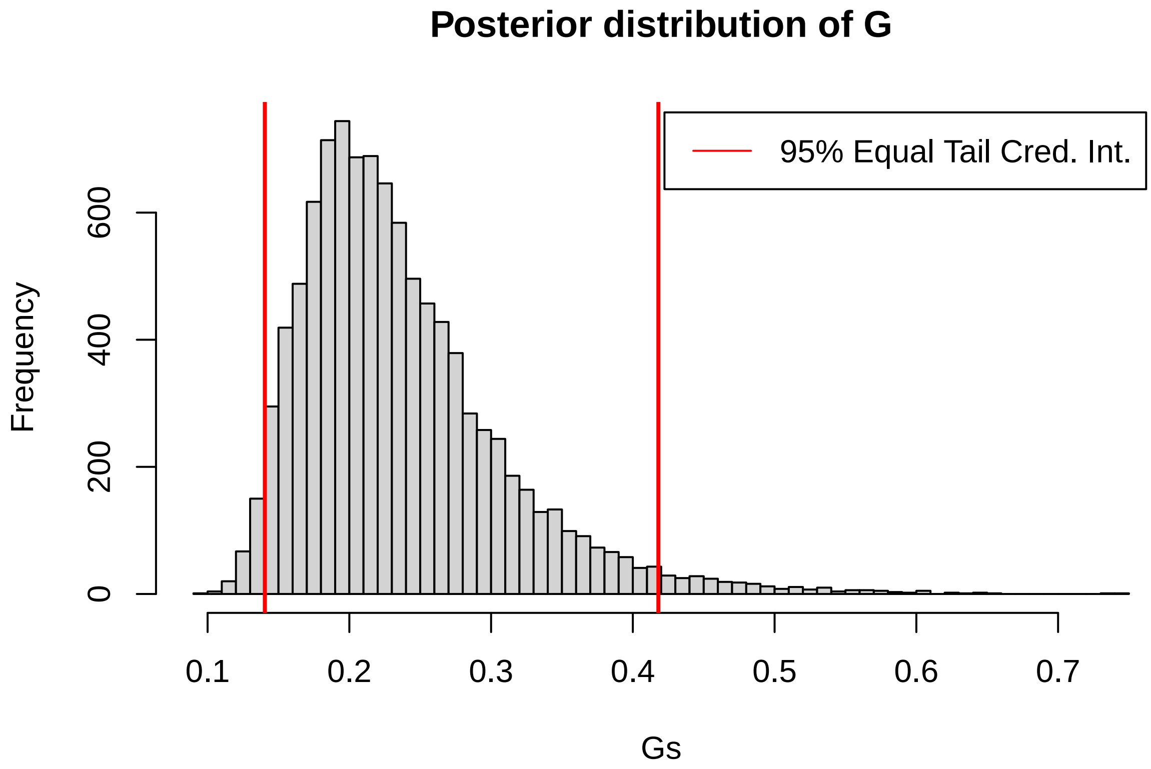
```
Gs = 2*pnorm((sqrt(sigsq_draws)/sqrt(2))) - 1

hist(Gs, main = "Posterior distribution of G", breaks = 50 )
```



2.c

Use the posterior draws from b) to compute a 95% equal tail credible interval for G. A 95% equal tail credible interval (a, b) cuts off 2.5% percent of the posterior probability mass to the left of a, and 2.5% to the right of b.

```
sortedGs = sort(Gs)
G250 = sortedGs[250]
G9750 = sortedGs[9750]

hist(Gs, main = "Posterior distribution of G", breaks = 50 )
abline(v = G250, col = "red", lwd=2)
abline(v = G9750, col= "red", lwd=2)
legend("topright", inset=.02, legend=c("95% Equal Tail Cred. Int."), col=c("red"), lty=1)
```



The 95% equal tail credible interval is [0.14038, 0.4180197], marked by the red vertical lines on the plot above.

2.d

Use the posterior draws from b) to compute a 95% Highest Posterior Density Interval (HPDI) for G. Compare the two intervals in (c) and (d). [Hint: 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 the HPDI. Note that you need to order/sort the estimated density values to obtain the HPDI.].

```
kernel_estimate = density(Gs)
x = kernel_estimate$x
y = kernel_estimate$y

y_sorted = sort(y)

#Estimating the HDPI using the Reimann sum method for approximating the area under the curve:
y_val = 7.0 #starting value
#by default, density() returns 512 points, thus Reimann sum with 512 boxes.
#box width is range(x)/ # of boxes
box_width = (max(kernel_estimate$x)-min(kernel_estimate$x))/512
```

```
delta = 0.001 #step size for y_val

while(sum(kernel_estimate$y[kernel_estimate$y >= y_val]*box_width) < .95){
  #cat("y=", y_val, "\r")
  y_val = y_val - delta
}

print(paste("y_val=",y_val))
```

```
## [1] "y_val= 0.673999999999329"
```

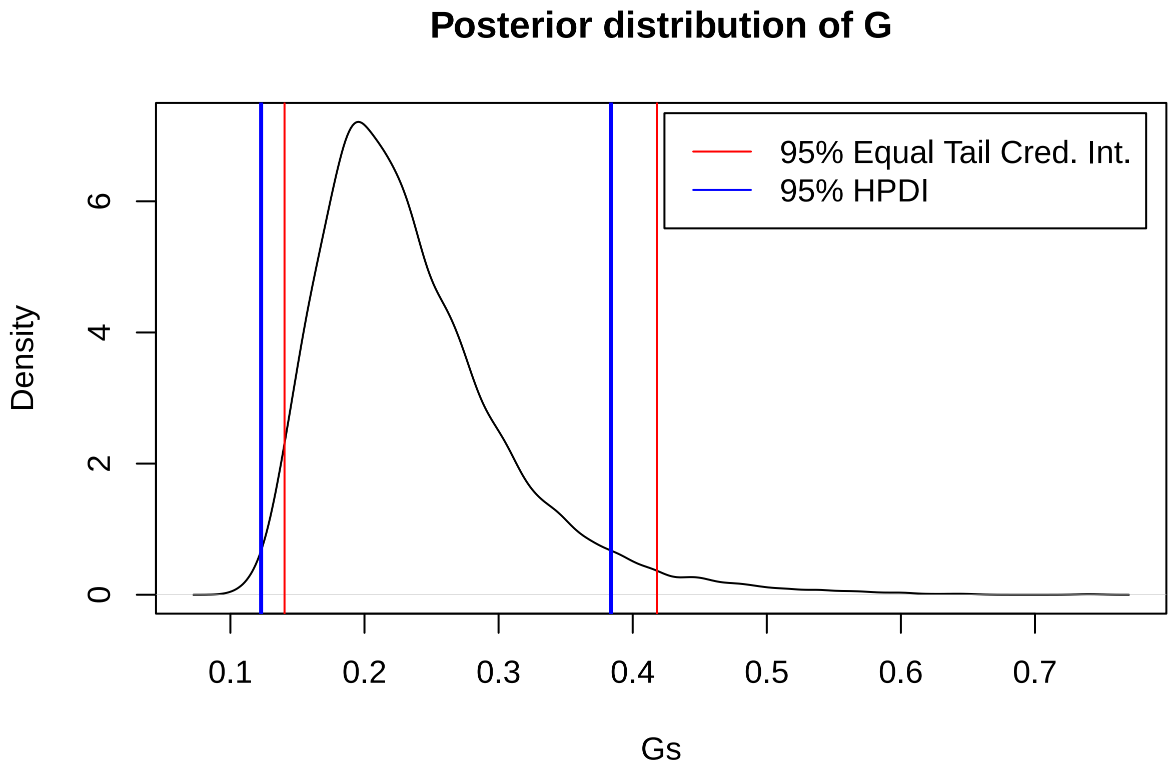
```
x_left_bound = min(kernel_estimate$x[kernel_estimate$y >= y_val])
print(paste("x_left_bound",x_left_bound))
```

```
## [1] "x_left_bound 0.122963640155895"
```

```
x_right_bound = max(kernel_estimate$x[kernel_estimate$y >= y_val])
print(paste("x_right_bount",x_right_bound))
```

```
## [1] "x_right_bount 0.383732050143403"
```

```
plot(kernel_estimate, main = "Posterior distribution of G", xlab = "Gs")
abline(v=x_left_bound, col = "blue", lwd=2)
abline(v=x_right_bound, col = "blue", lwd=2)
abline(v = G250, col = "red", lwd=1)
abline(v = G9750, col= "red", lwd=1)
legend("topright", inset=.02, legend=c("95% Equal Tail Cred. Int.", "95% HPDI"),
       col=c("red", "blue"), lty=c(1,1))
```



The 95% Highest Posterior Density Interval is [0.1229636, 0.3837321], marked by the blue vertical lines on the plot above. It is shifted to the left compared to the 95% equal tail credible interval in red, which makes sense since the posterior distribution is skewed right.

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:

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

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

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

. Assume that these data points conditional on (μ, κ) are independent observations from the following von Mises distribution:

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

where $I_0(\kappa)$ is the modified Bessel function of the first kind of order zero [see ?besselI in R]. The parameter $\mu(-\pi \leq \mu \leq \pi)$ 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 2.4. Let $\kappa \sim Exponential(\lambda = 0.5)$ a priori, where λ is the rate parameter of the exponential distribution (so that the mean is $1/\lambda$).

3.a

Derive the expression for what the posterior $p(\kappa|y,\mu)$ is proportional to. Hence, derive the function $f(\kappa)$ such that $p(\kappa|y,\mu) \propto f(\kappa)$. Then, plot the posterior distribution of κ for the wind direction data over a fine grid of κ values. [Hint: you need to normalize the posterior distribution of κ so that it integrates to one.]

To find the posterior distribution, we first need to find the Likelihood:

$$L = \prod_{i=1}^{n} \frac{\exp(\kappa * \cos(y_i - \mu))}{2\pi I_0(\kappa)} = \frac{\exp(\sum_{i=1}^{n} \kappa * \cos(y_i - \mu))}{(2\pi I_0(\kappa))^n}$$

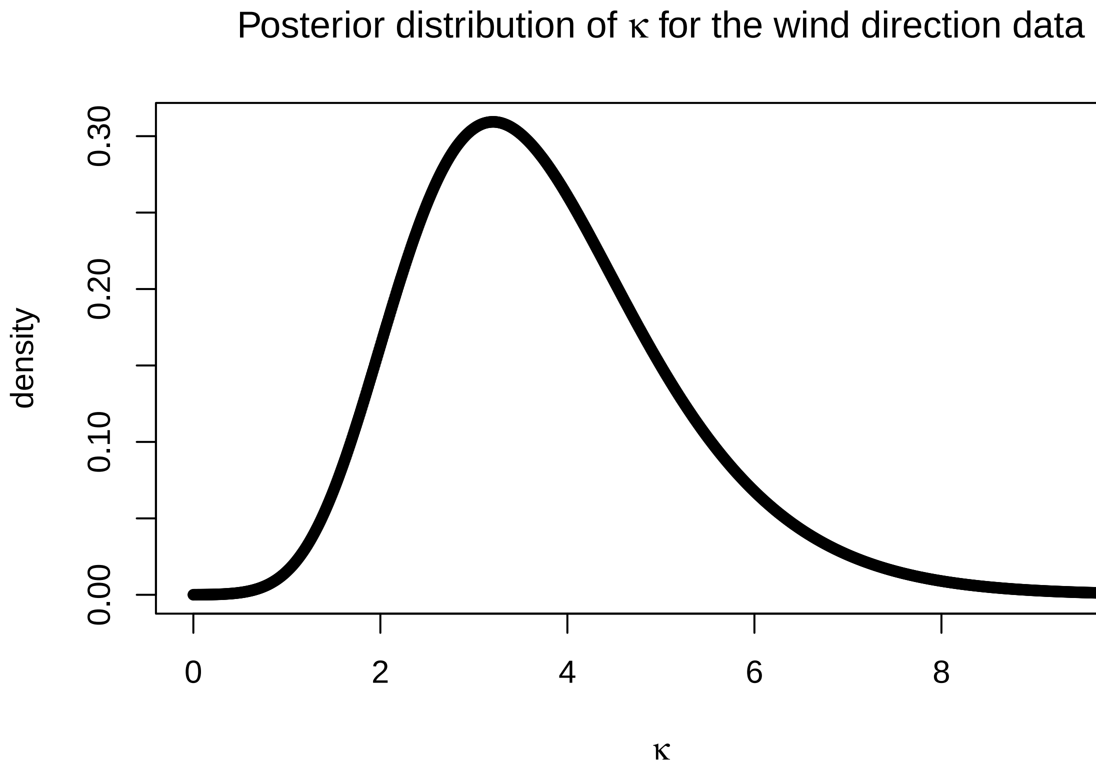
hence:

$$p(\kappa|y,\mu) \propto L * p(\kappa) = \frac{\exp(\sum_{i=1}^{n} \kappa * \cos(y_i - \mu))}{(2\pi I_0(\kappa))^n} * \frac{1}{2}\exp(-\kappa/2)$$

which can be simplified to :

$$p(\kappa|y,\mu) \propto f(\kappa) = \frac{\exp(\kappa * \sum_{i=1}^{n} \cos(y_i - \mu) - \kappa/2)}{(I_0(\kappa))^n}$$

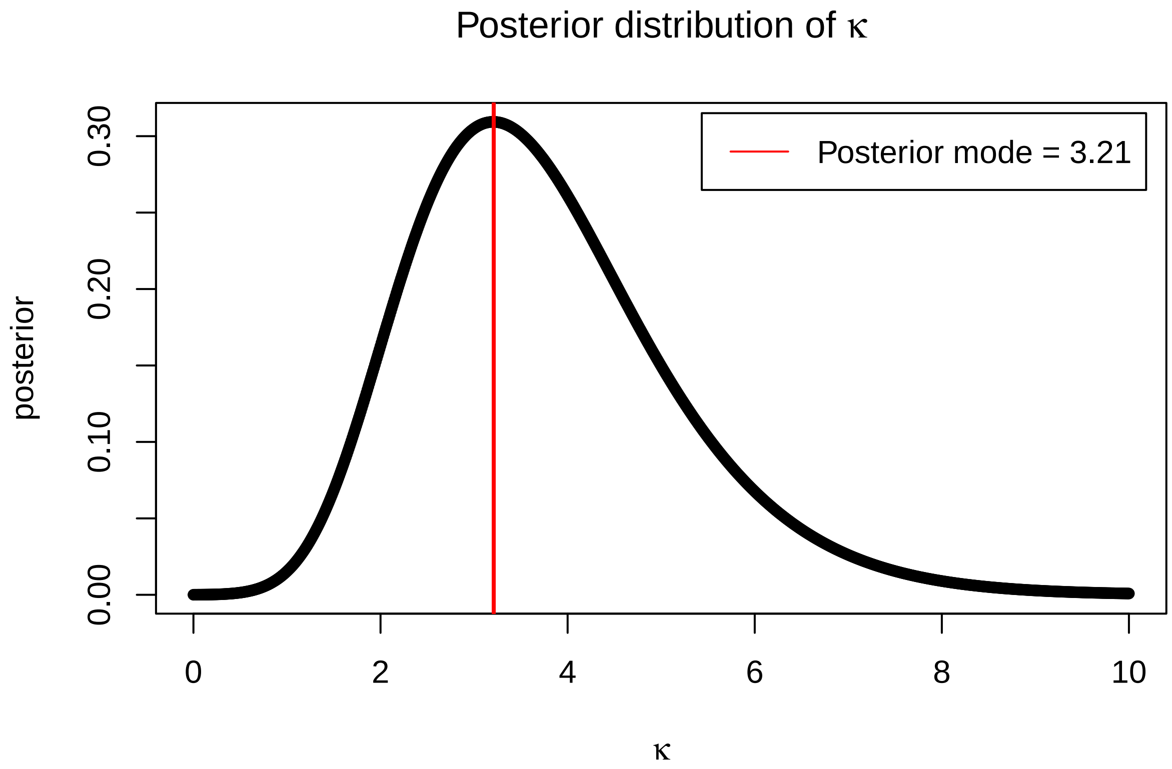
```r
obs = c(-2.79, 2.33, 1.83, -2.44, 2.23, 2.33, 2.07, 2.02, 2.14, 2.54)
n = length(obs)
mu = 2.4

minimum = 0
maximum = 10
delta = 0.01
kappaGrid <- seq(minimum, maximum, by = delta)

posterior_func <- function(kappa, mu, y){
  n = length(y)
  (exp(kappa * sum(cos(y - mu))) - kappa/2)/(besselI(kappa, nu = 0))^n
}
posterior = posterior_func(kappaGrid, mu, obs)
#normalizing:
posterior = posterior/(sum(posterior)*delta)

plot(posterior ~ kappaGrid, pch = 20,
     main = expression(paste("Posterior distribution of ", kappa, " for the wind direction data")),
     xlab=expression(kappa), ylab = "density")
```



3.b

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

```
idx_mode = which.max(posterior)
post_mode = kappaGrid[idx_mode]

plot(posterior ~ kappaGrid, pch = 20,
     xlab = expression(kappa), main = expression(paste("Posterior distribution of ", kappa)))
abline(v = post_mode, col = "red", lwd=2)
legend("topright", inset=.02, legend=paste0("Posterior mode = ", post_mode), col=c("red"), lty=1)
```



The posterior mode can be approximated by finding the point kappa with the highest density. The posterior mode is approximated to be 3.21, visualized by the red vertical line on the plot.