Lab 1

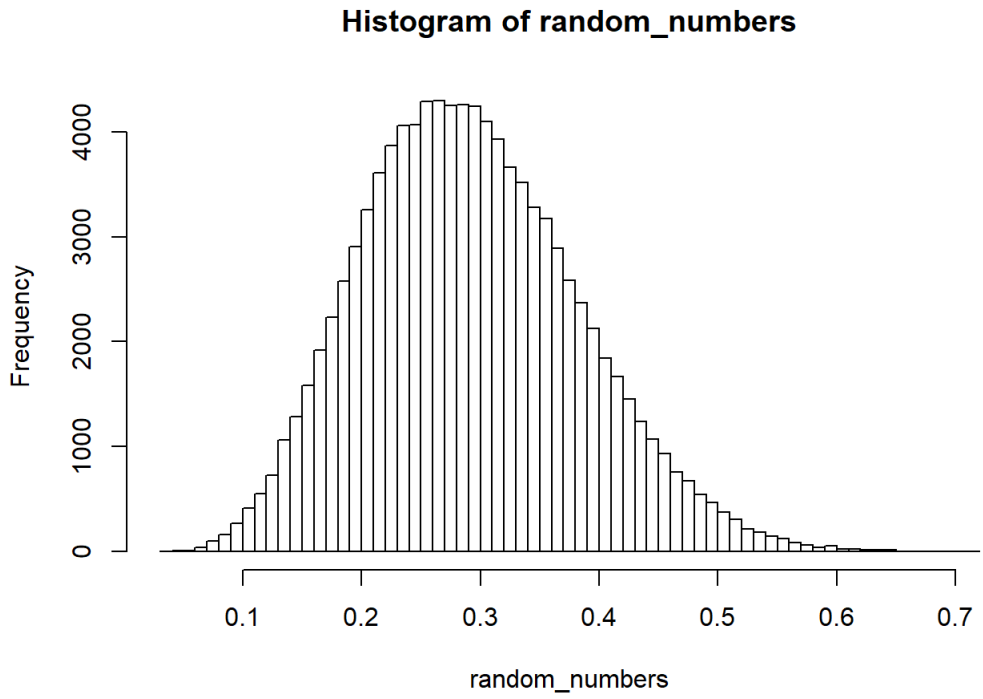
Part 1

Bernoulli ... again. Let $y_1, \ldots, y_n | \theta \sim \text{Bern}(\theta)$, and assume that you have obtained a sample with s = 5 successes in n = 20 trials. Assume a Beta(α_0, β_0) prior for θ and let $\beta_0 = \alpha_0 = 2$.

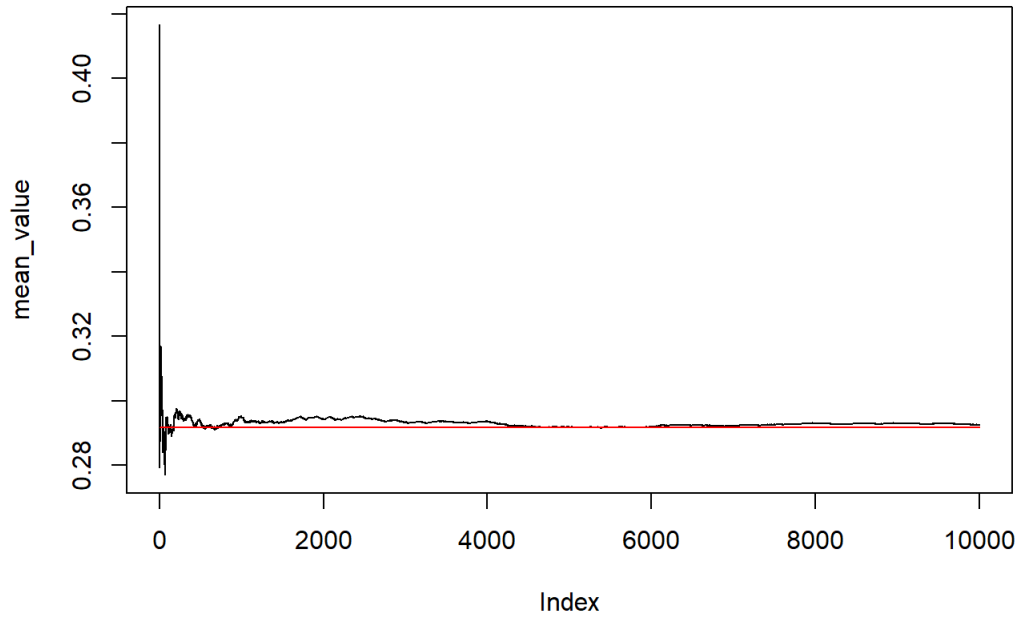
a)

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.

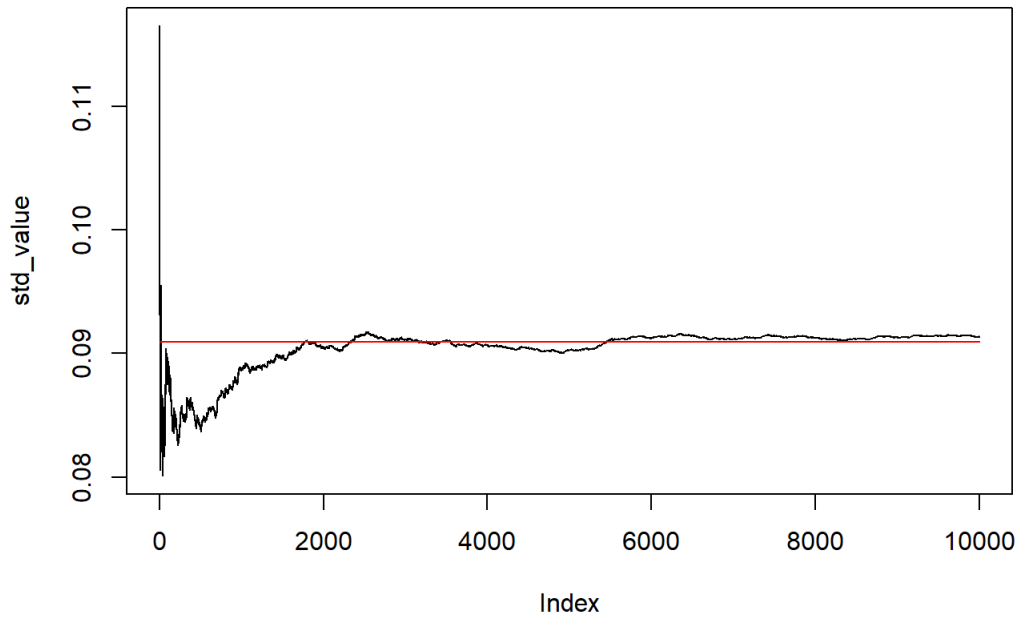
```
post_beta = 2;
post_alpha = 2;
s = 5
f= 20-s
alpha = post_alpha+s
beta = post_beta+f

beta_mean = (alpha)/(alpha+beta);
beta_std = sqrt((alpha*beta)/(((alpha+beta)^2)*(alpha+beta+1)))

n = 100000
random_numbers = rbeta(n, alpha, beta)
hist(random_numbers, breaks = 50)
```




```
mean_value = rep(0, 10000)
std_value = rep(0, 10000)

for(i in 1:10000){
  mean_value[i] = mean(random_numbers[1:i])
  std_value[i] =sqrt(var(random_numbers[1:i]))
}

plot(mean_value, type="l")
points(x=1:10000, y=rep(beta_mean,10000), type = "l", col="red")
```




```
plot(std_value, type="l")
points(x=1:10000, y=rep(beta_std, 10000), type="l", col="red")
```



b)

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

```
count = ifelse (random_numbers > 0.3, 1, 0)
count = sum(count);
#compare
posterior_prob = count/n
exact_value = 1-pbeta(0.3, alpha, beta)
posterior_prob
```

```
## [1] 0.4397
```

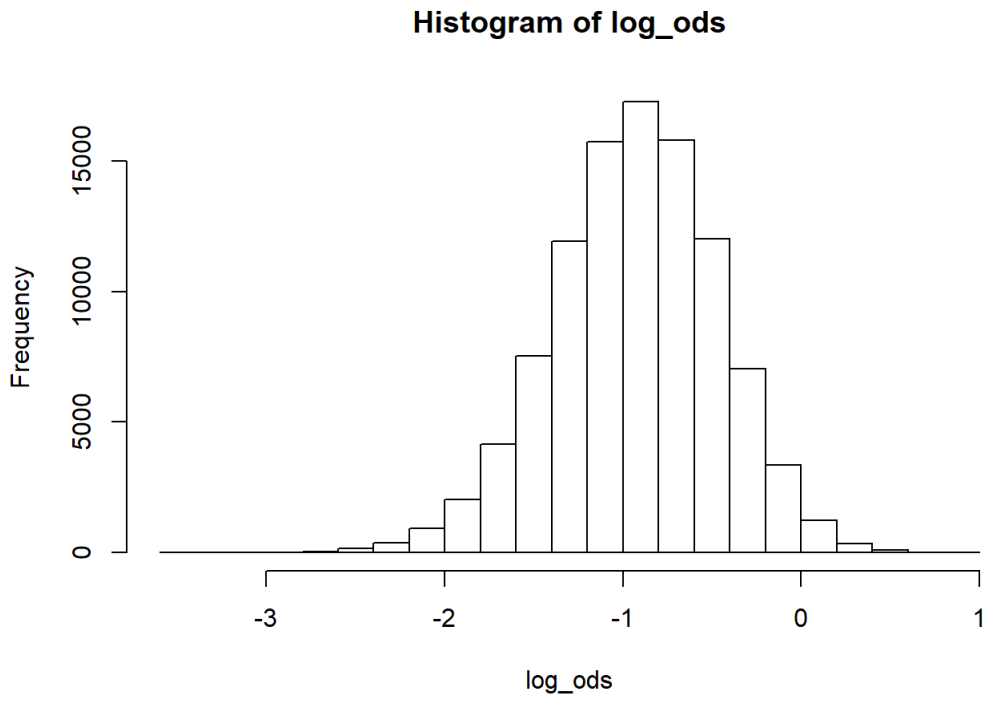
```
exact_value
```

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

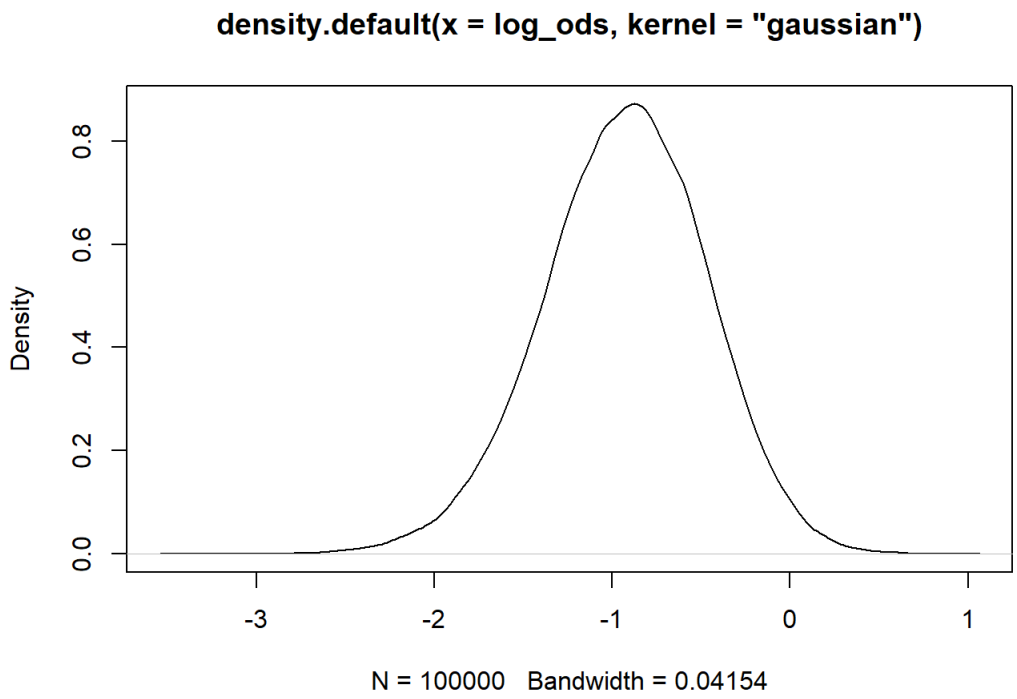
c)

Compute the posterior distribution of the log-odds ϕ = logθ by simulation 1 − θ (nDraws = 10000). [Hint: hist() and density() might come in handy]

```
log_ods = log(random_numbers/(1-random_numbers))
samp = mean(log_ods)
hist(log_ods)
```




```
fit = density(log_ods, kernel="gaussian")
plot(fit)
```




```
print(fit)
```

```
##
## Call:
##  density.default(x = log_ods, kernel = "gaussian")
##
## Data: log_ods (100000 obs.); Bandwidth 'bw' = 0.04154
##
##        x                  y
##  Min.   :-3.54276   Min.   :0.0000011
##  1st Qu.:-2.39103   1st Qu.:0.0010163
##  Median :-1.23931   Median :0.0396982
##  Mean   :-1.23931   Mean   :0.2168532
##  3rd Qu.:-0.08758   3rd Qu.:0.3870995
##  Max.   : 1.06414   Max.   :0.8716849
```

Part 2

Log-normal distribution and the Gini coefficient. Assume that you have asked 10 randomly selected persons about their monthly in- come (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

$p(y|\mu, \sigma) = 1y\sqrt{2\pi\sigma^2}\exp(-12\sigma^2(\log(y)-\mu)^2)$ for $y > 0$, $\mu > 0$ and $\sigma^2 > 0$. The log-normal distribution is related to the normal distribution as follows: if $y \sim \log N(\mu, \sigma^2)$. then $\log y \sim N(\mu, \sigma^2)$. Let $y_1, \ldots y_n|\mu, \sigma^2 \sim \log N(\mu, \sigma^2)$, where $\mu = 3.7$ is assumed to be known but σ is unknown with non-informative prior $p(\sigma) \propto 1/\sigma$. The posterior for σ is the Inv−χ^2(n,\tau^2) distribution, where

\tau^{2} = \frac{\sum_{i=1}^{n}\left(\log{y_{i}} - \bar{y}\right)^{2}}{n}

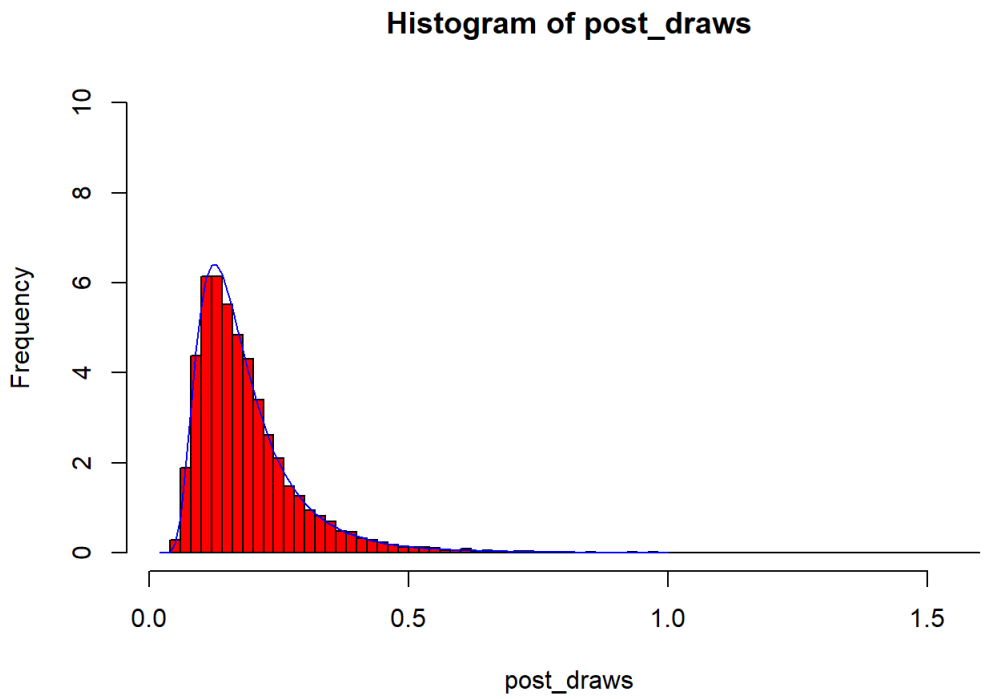
a)

Simulate 10, 000 draws from the posterior of \sigma^2 (assuming \mu = 3.7) and compare with the theoretical Inv−χ^2(n,\tau^2) posterior distribution.

```
y = c(44, 25, 45, 52, 30, 63, 19, 50, 34, 67)
my = 3.7
n=10
tau = sum((log(y)-my)^2)/length(y)
tau
```

```
## [1] 0.1499412
```

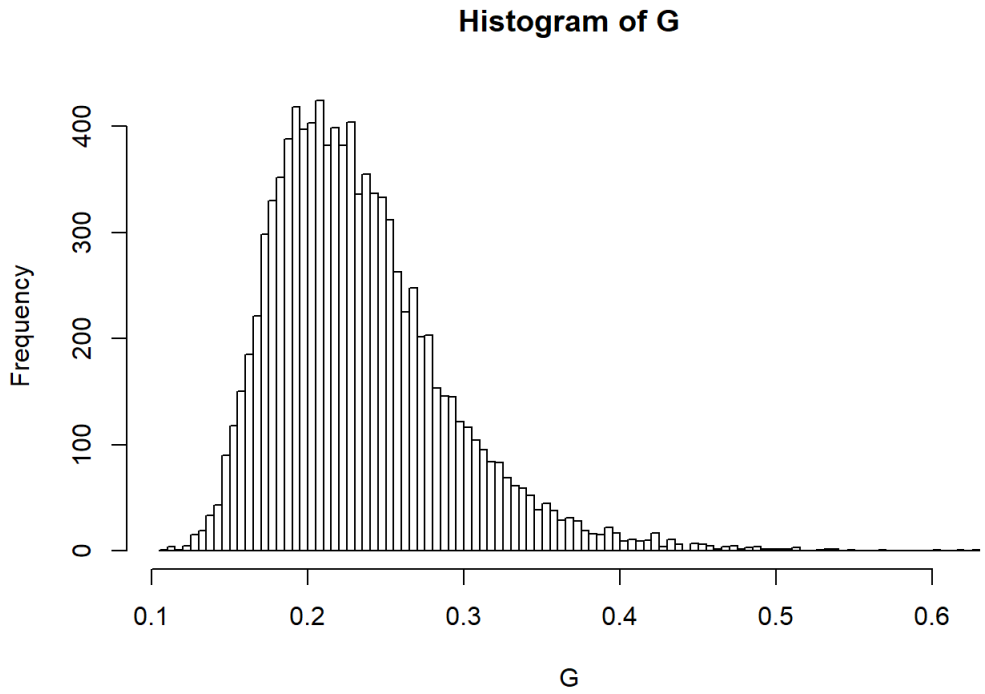
```
x_draw = rchisq(10000,n)
post_draws = ((n)*tau)/x_draw #scaled inv chi 2

h=hist(post_draws,breaks=100, plot=FALSE)
h$counts=h$counts*50/sum(h$counts)
plot(h, col="red",ylim=c(0,10))

library(LaplacesDemon)
scaled_inv_chi_2 = dinvchisq(seq(0.02, 1, by=0.01), n, tau)
points(seq(0.02, 1, by=0.01),scaled_inv_chi_2,col="blue",type="l")
```



b)

The most common measure of income inequality is the Gini coefficient, G, where 0 ≤ G ≤ 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(\sigma/\sqrt{2})-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.

```
G = 2*pnorm(sqrt(post_draws)/sqrt(2))-1
hist(G, breaks=100)
```



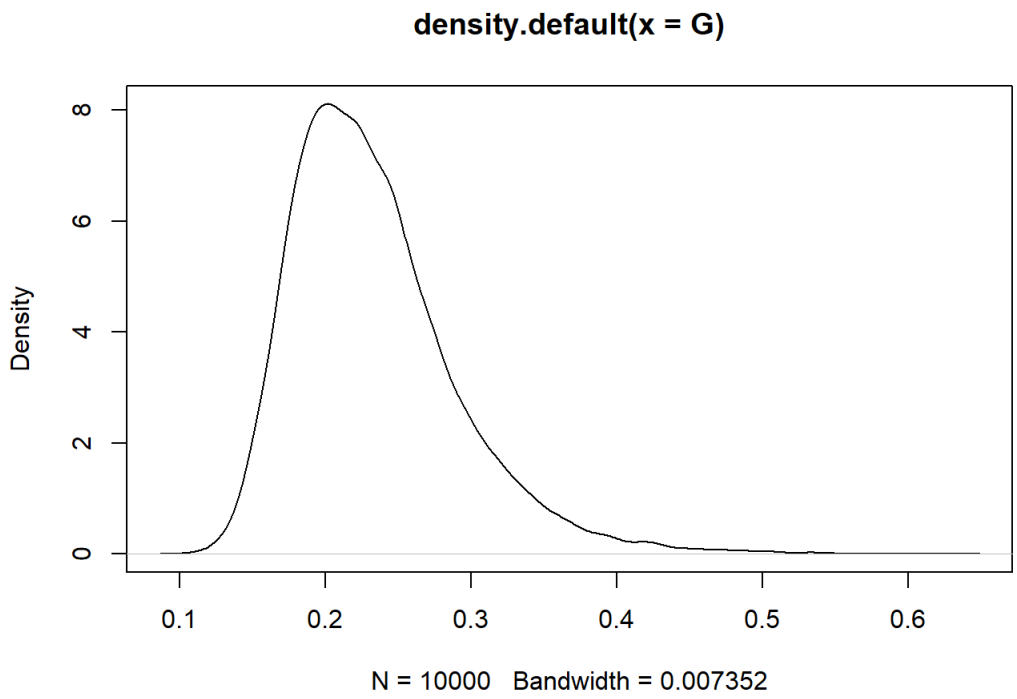
c)

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.

```
sort_g = sort(G)
low_tail = sort_g[length(G)*0.5]
high_tail = sort_g[length(G)*0.95]
c(low_tail, high_tail)
```

```
## [1] 0.2242372 0.3389215
```

```
fit_G = density(G)
plot(fit_G)
```




```
library(HDInterval)
hid_90 = hdi(fit_G, credMass=0.90)
hid_90
```

```
##     lower     upper
## 0.1479508 0.3175286
## attr(,"credMass")
## [1] 0.9
## attr(,"height")
## [1] 1.738229
```

Part 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 $-\pi < y \leq \pi$. 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

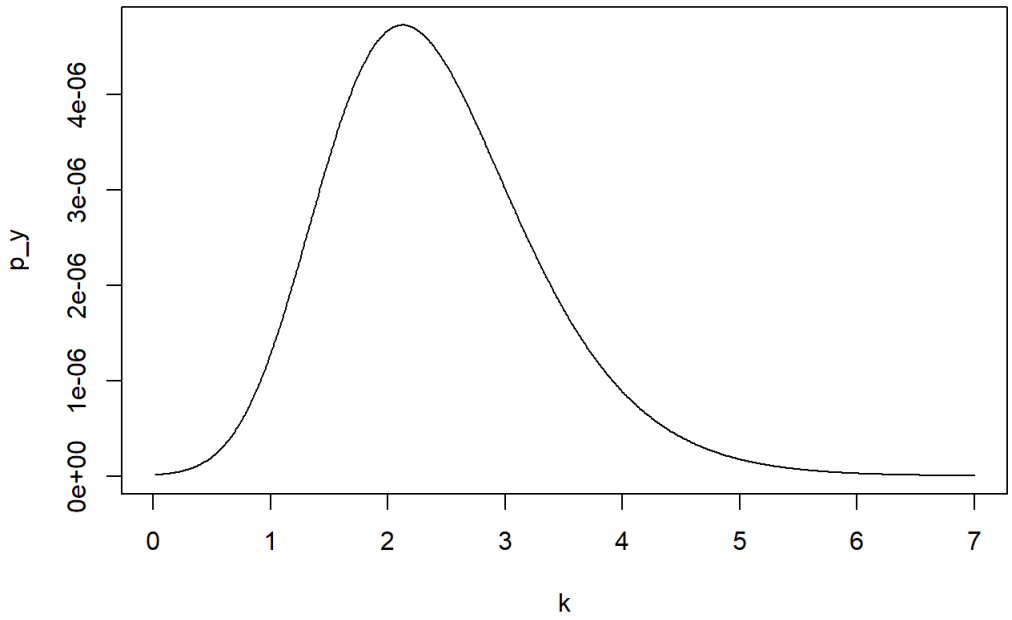
$$p(y|\mu, \kappa) = \frac{\exp [\kappa \cdot \cos(y - \mu)]}{2\pi I_0(\kappa)}, -\pi \leq y \leq \pi$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind of order zero [see ?bessell 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 $1/\lambda$).

a)

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

```
wind_d = c(40, 303, 326, 285, 296, 314, 20, 308, 299, 296)
wind_r = c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)

#?Bessel
k=seq(0.01, 7, by=0.01) #exp(lambda=1) as prior for k
my=2.39
#mises_distr =exp(k*cos(wind_r-my))/(2*pi*besselI(k,0))
p_y = rep(0, length(k))
for(i in 1:length(k)){
  p_y[i] = prod(exp(k[i]*cos(wind_r-my))/(2*pi*besselI(k[i],0)) )*dexp(k[i])
}
plot(k,p_y, type="l")
```



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

```
#find mode: moest frequent value of k.
k[which.max(p_y)]
```

```
## [1] 2.12
```