Computer Lab 1

Dhyey Patel, Erik Anders

4/18/2020

##1. Bernoulli … again. Let y 1 , …, y n |θ ∼ Bern(θ), 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 α 0 = β 0 = 2.

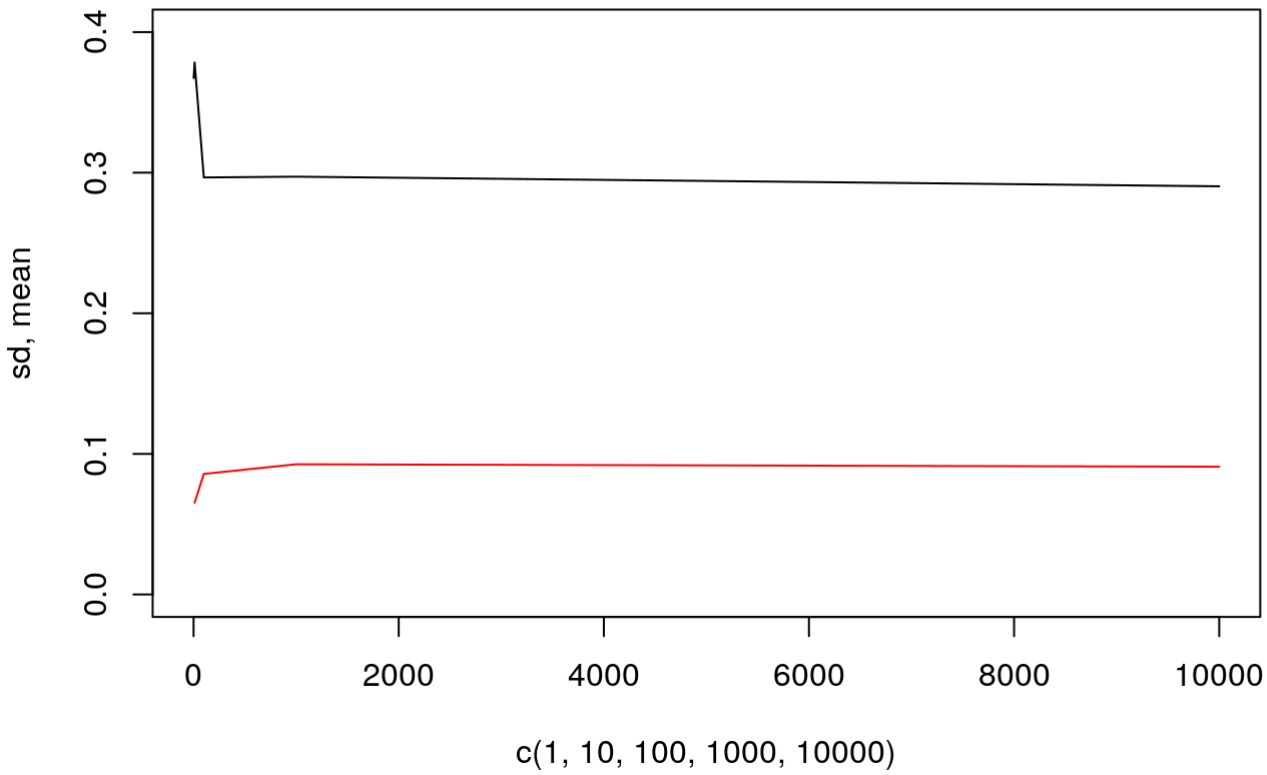
a. Draw random numbers from the posterior θ|y ∼ Beta(α 0 + s, β 0 + f), y = (y 1 , . . . , y n), and verify graphically that the posterior mean and standard de- viation converges to the true values as the number of random draws grows large.

```
## [1] "True mean"
```

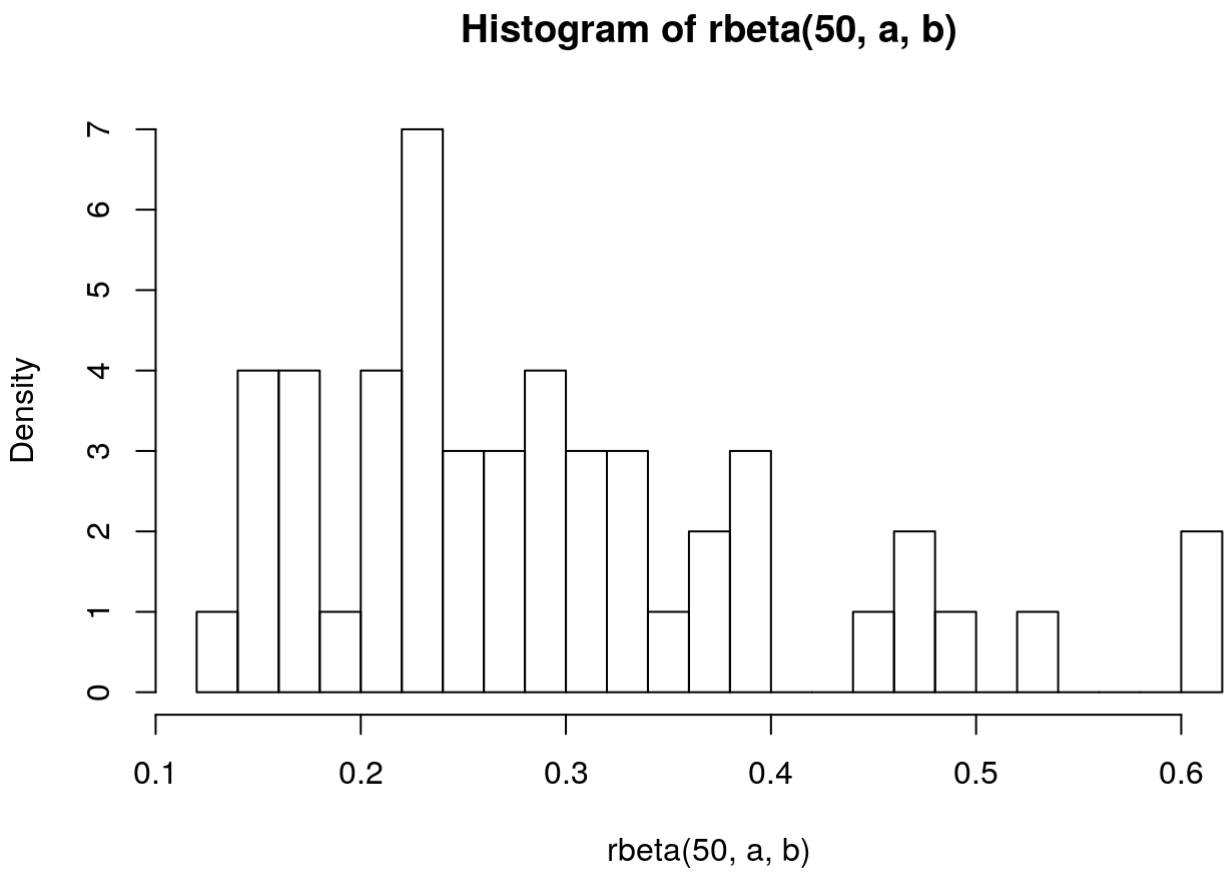
```
## [1] 0.2916667
```

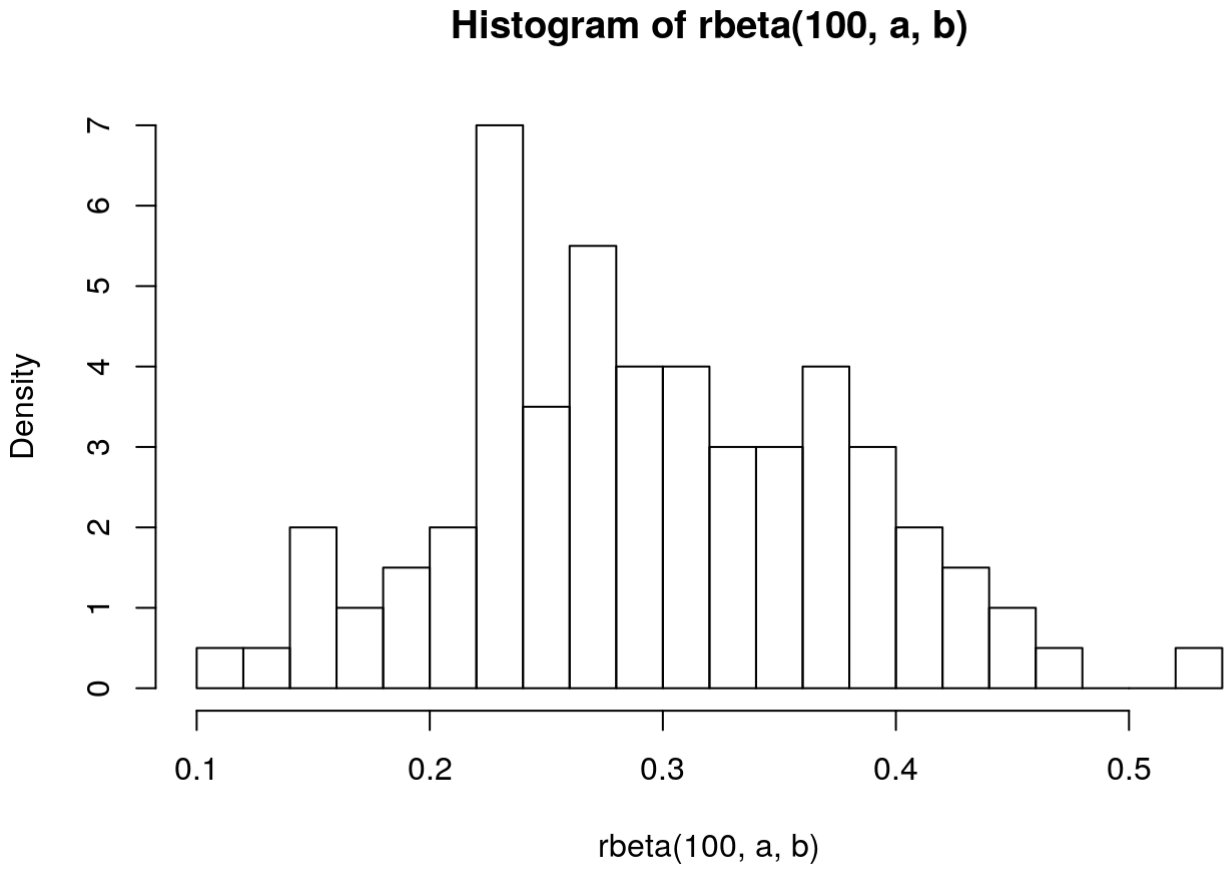
```
## [1] "True sd"
```

```
## [1] 0.008263889
```

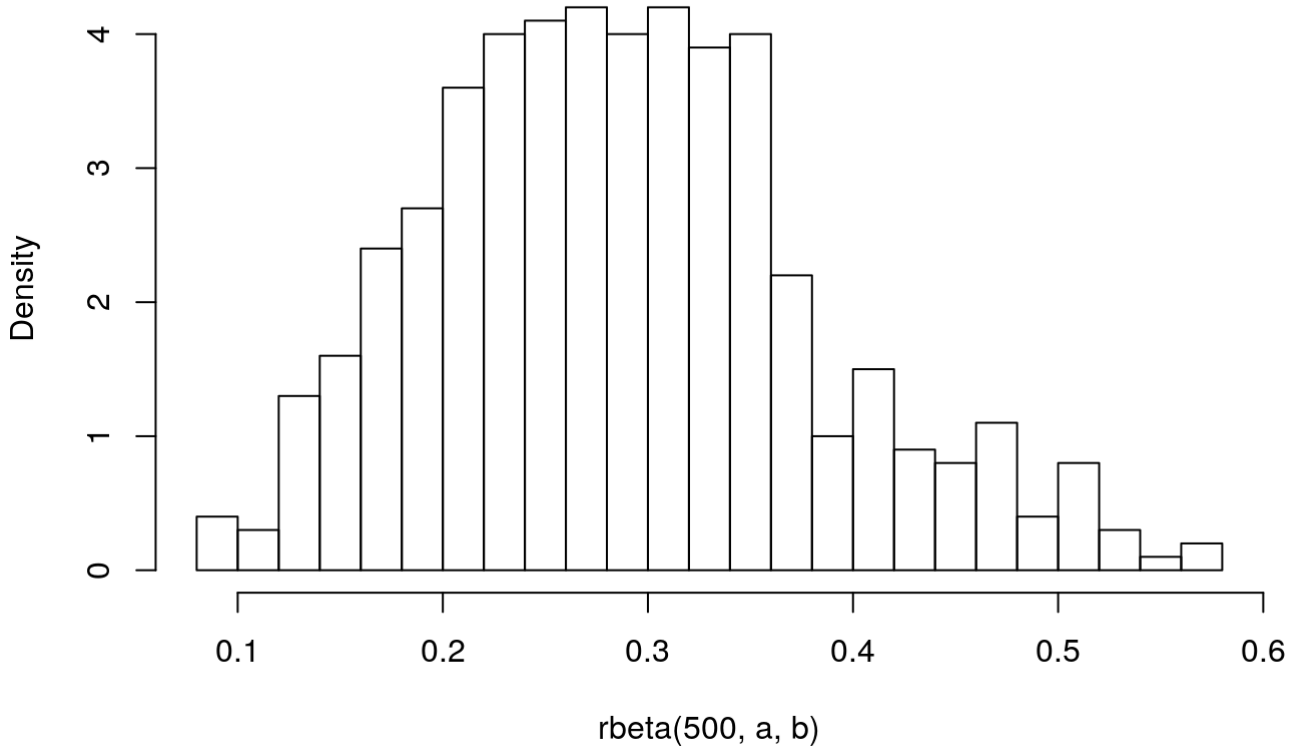


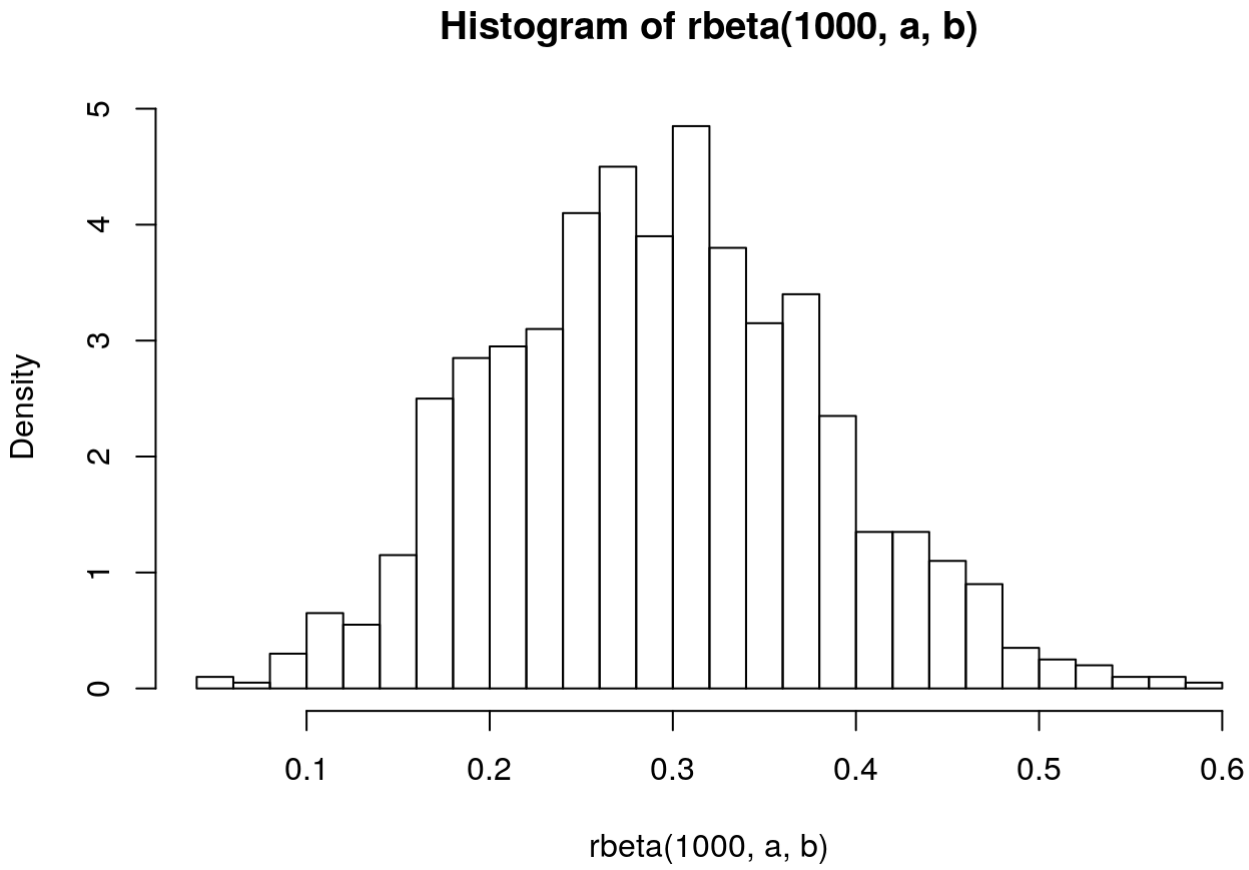
As shown above, the mean (black) and the standard deviation (red) converge very quickly to their true values.





Histogram of rbeta(500, a, b)





As we can see from the plots, as n increases from 50 to 1000, the mean visually converges to the true mean 0.29.

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

```
## [1] "simulated probability"
```

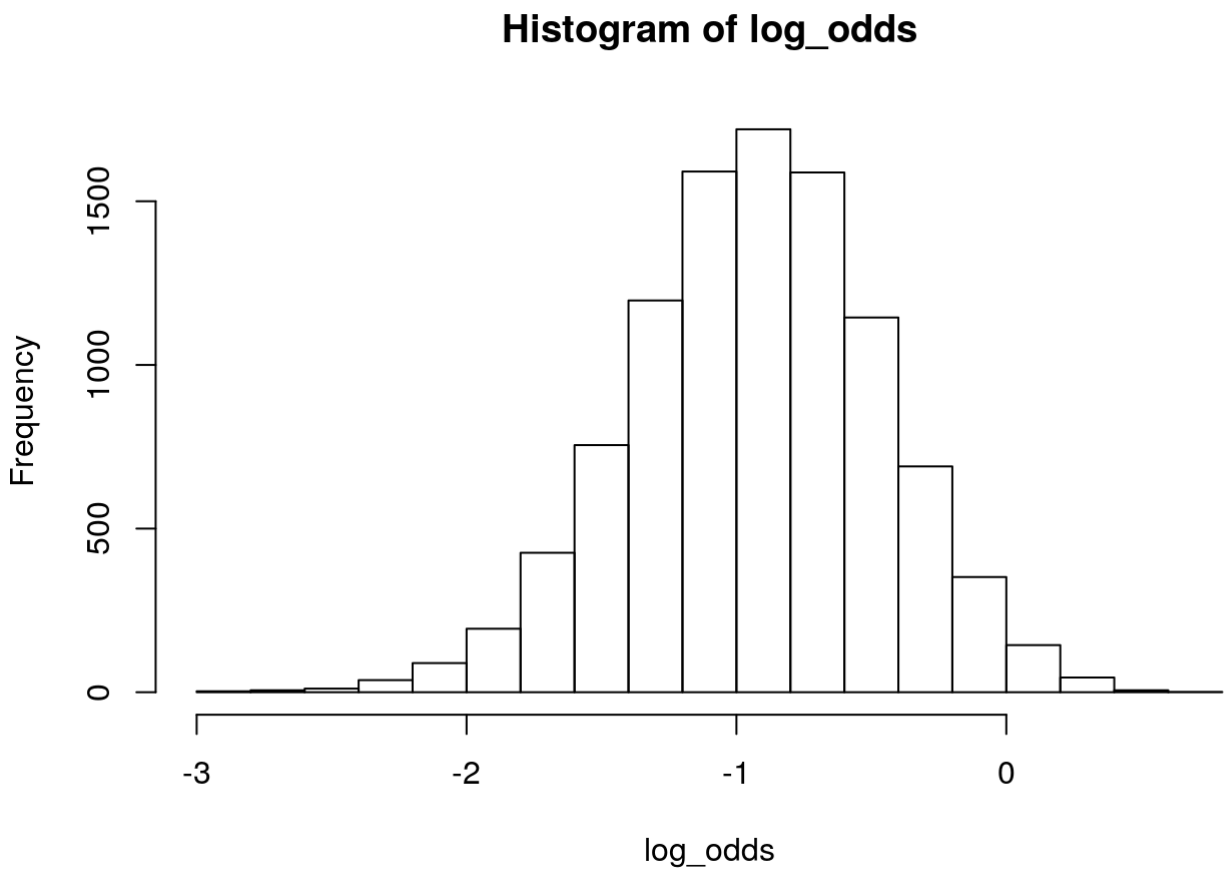
```
## [1] 0.4399
```

```
## [1] "exact probability"
```

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

Posterior probability by simulation comes up to be approximately 0.439. Posterior probability theoretically is 0.4399.

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




```
##
## Call:
##  density.default(x = log_odds)
##
## Data: log_odds (10000 obs.); Bandwidth 'bw' = 0.06606
##
##        x                 y
##  Min.   :-3.1946   Min.   :0.0000071
##  1st Qu.:-2.1832   1st Qu.:0.0045802
##  Median :-1.1718   Median :0.0806820
##  Mean   :-1.1718   Mean   :0.2469383
##  3rd Qu.:-0.1603   3rd Qu.:0.4631446
##  Max.   : 0.8511   Max.   :0.8595850
```

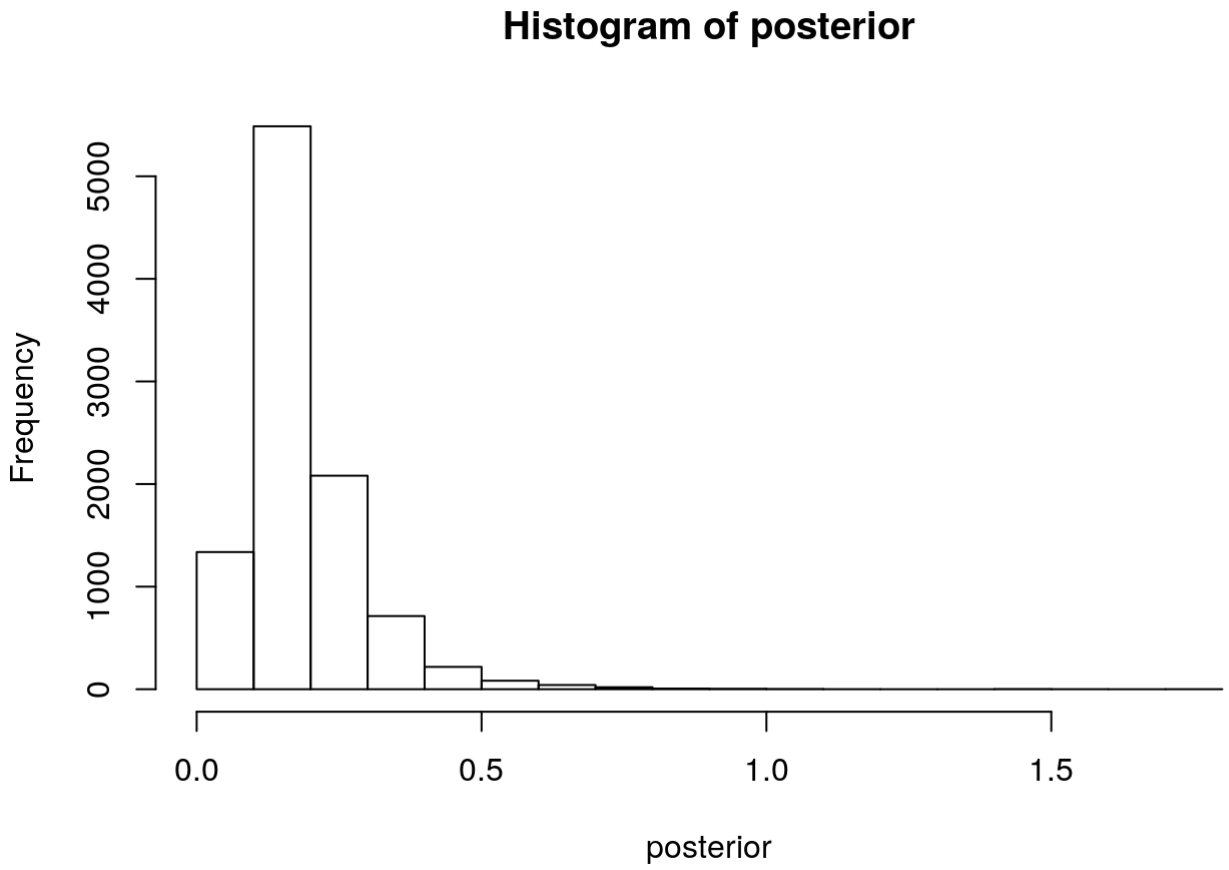
##2. Log-normal distribution and the Gini coefficient. Assume that you have asked 10 randomly selected persons about their monthly in- come (in thousands Swedish Krona) and obtained the following ten observations: 44, 25, 45, 52, 30, 63, 19, 50, 34 and 67. A common model for non-negative continuous variables is the log-normal distribution. The log-normal distribution log N (μ, σ 2) has density function

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

for y > 0, μ > 0 and σ 2 > 0. The log-normal distribution is related to the normal distribution as follows: if y ∼ log N (μ, σ 2) then log y ∼ N (μ, σ 2). Let iid y 1 , …, y n |μ, σ 2 ∼ log N (μ, σ 2), where μ = 3.7 is assumed to be known but σ 2 is unknown with non-informative prior p(σ 2) ∝ 1/σ 2 . The posterior for σ 2 is the Inv − χ 2 (n, τ 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 μ = 3.7) and com- pare with the theoretical Inv − χ 2 (n, τ 2) posterior distribution.




```
## [1] "Simulated mean"
```

```
## [1] 0.1862659
```

```
## [1] "Simulated variance"
```

```
## [1] 0.01083515
```

```
## [1] "Theoretical mean"
```

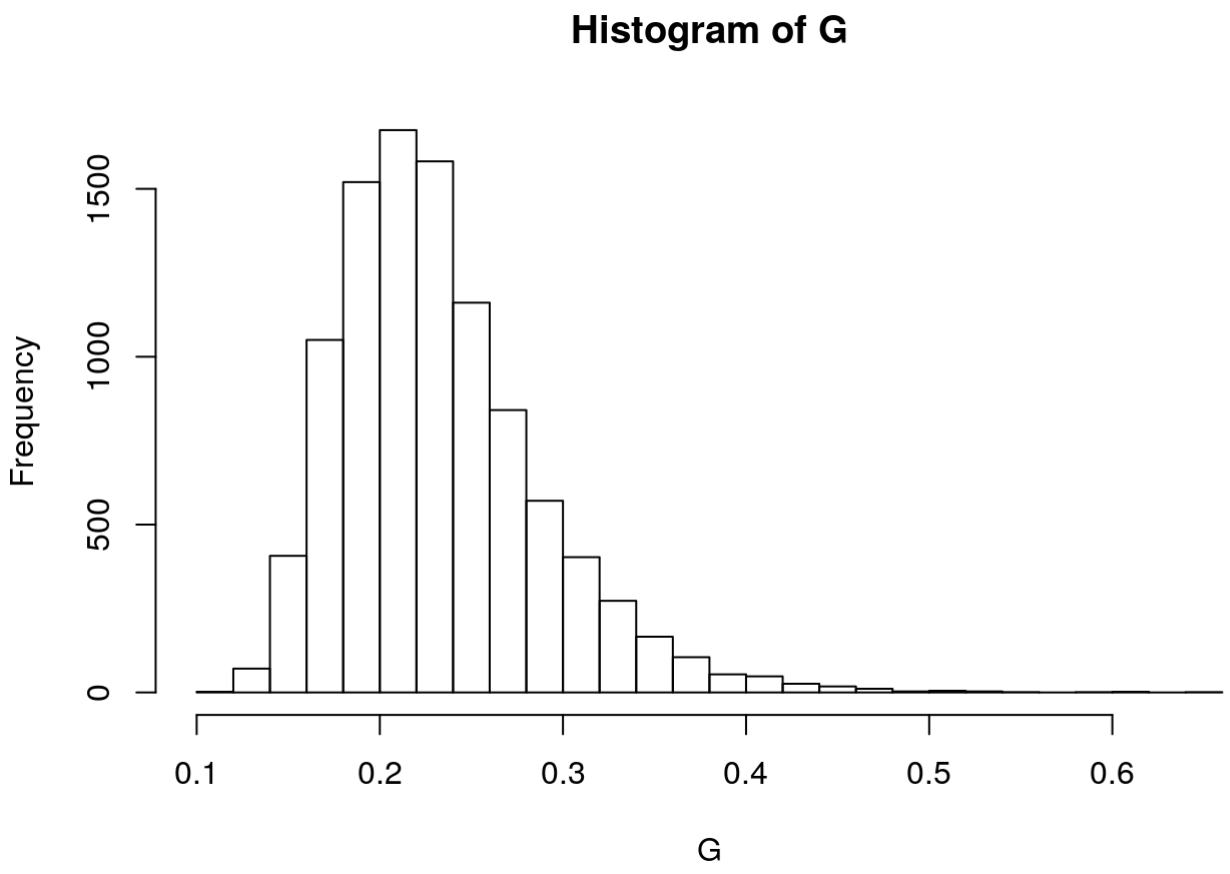
```
## [1] 0.1874264
```

```
## [1] "Theoretical variance"
```

```
## [1] 0.01170956
```

To compare the simulated and theoretical distribution we chose the mean and standard deviation. As shown in the output above, they are both very close to each other.

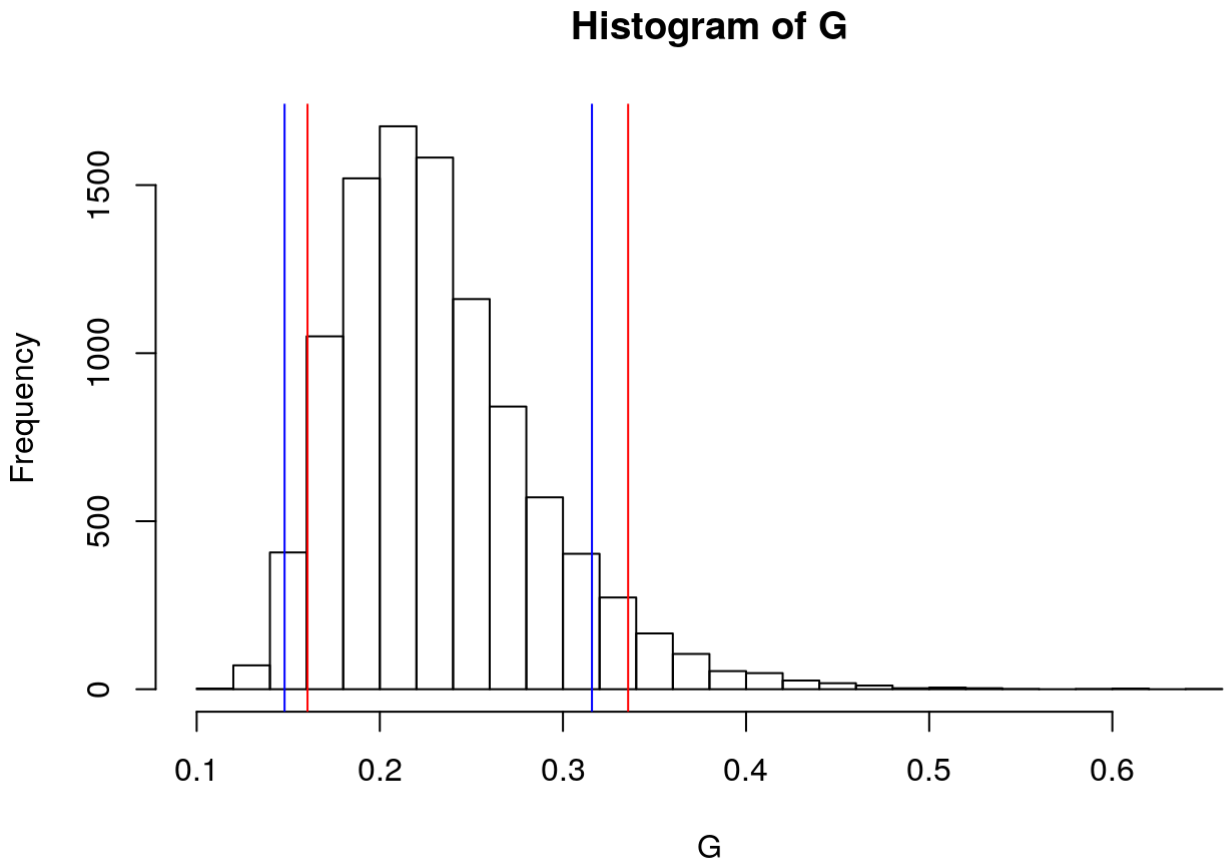
b. The most common measure of income inequality is the Gini coefficient, G, where 0 ≤ G ≤ 1. G = 0 means a completely equal income distribution, whereas G = 1 means complete income inequality. √ ◁ See Wikipedia for more information. It can be shown that G = 2Φ σ/ 2 − 1 when incomes follow a log N (μ, σ 2) distribution. Φ(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.

```
## [1] "Highest Posterior Density interval is :"
```

```
## [1] 0.1479902 0.3158322
```

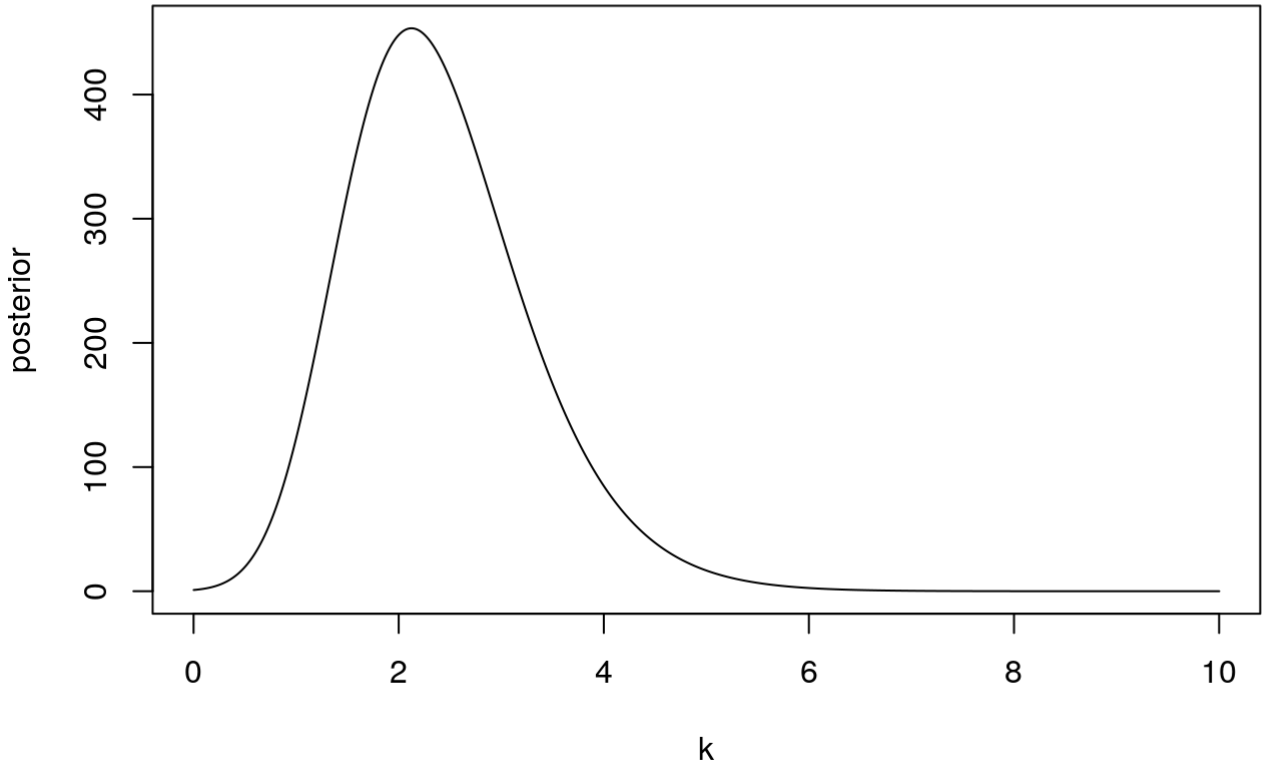


The red lines indicate the 90 % equal tail credible interval while the blue lines indicate the highest posterior density interval. We can see that the highest posterior density interval is smaller than the 90% equal tail credible interval.

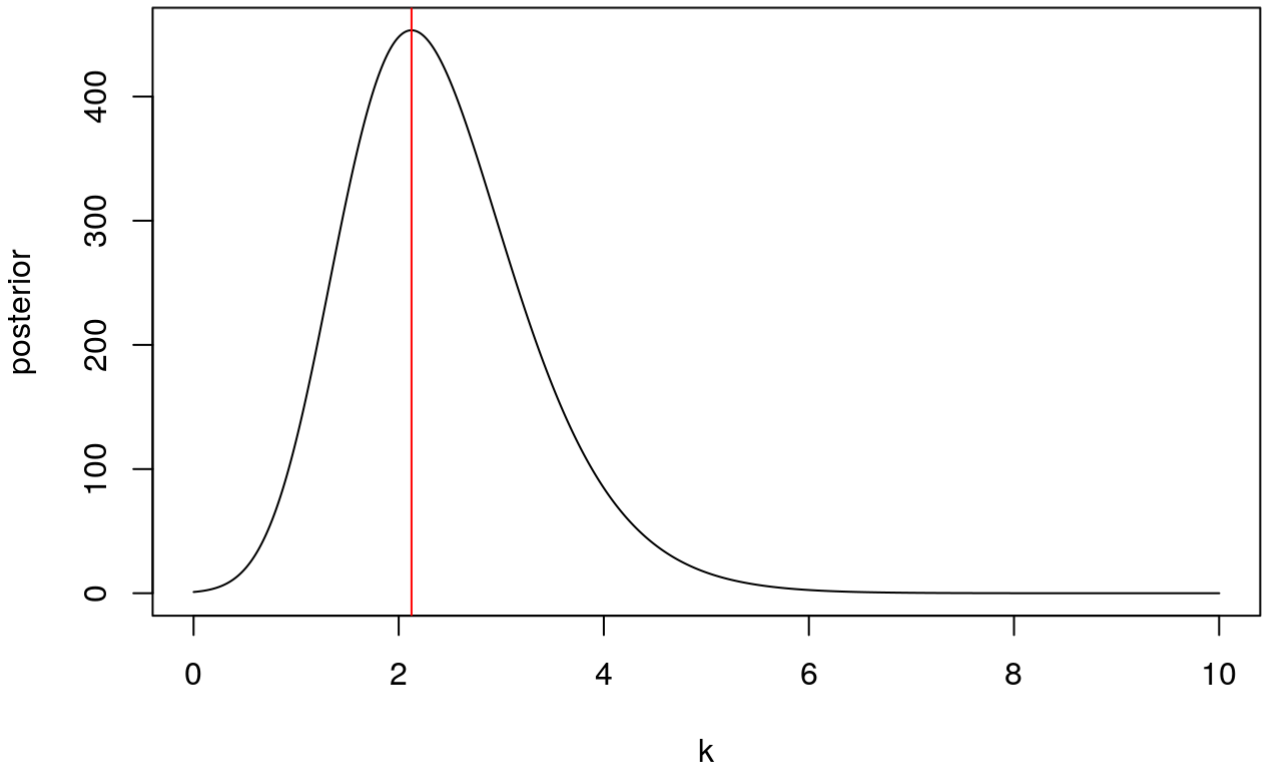
##3. Bayesian inference for the concentration parameter in the von Mises distribution. This exercise is concerned with directional data. The point is to show you that the posterior distribution for somewhat weird models can be obtained by plotting it over a grid of values. The data points are observed 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 −π ≤ y ≤ π . 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

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

where I 0 (κ) is the modified Bessel function of the first kind of order zero [see ?bessell in R]. The parameter μ (−π ≤ μ ≤ π) is the mean direction and κ > 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 κ ∼ Exponential(λ = 1) a priori, where λ is the rate parameter of the exponential distribution (so that the mean is 1/λ). (a) Plot the posterior distribution of κ for the wind direction data over a fine grid of κ values.



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



##Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#1.
#a).

avg <- c()
sdv <- c()
for (n in c(1,10,100,1000,10000)) {
  dist <- rbeta(n = n,shape1 = 2+5,shape2 = 2+15)
  #plot(dist)
  avg <- c(avg,mean(dist))
  sdv <- c(sdv, sd(dist))
}

a = 7
b = 17
mean_th = a/(a+b)
sd_th = a*b/(((a+b)^2) * (a+b+1))

print("True mean")
print(mean_th)
print("True sd")
print(sd_th)

#E = (alpha + s)/(alpha + s +beta+ f)
# = (alpha + 5)/(alpha + 5+beta + 15)
# = (alpha + 5)/(alpha + beta + 20)
# = 7/24 = 0.2916667

plot(c(1,10,100,1000,10000),avg, type = "l", ylim = c(0,0.4), ylab = "sd, mean")
lines(c(1,10,100,1000,10000),sdv, col="red")

hist(rbeta(50,a,b),freq = FALSE, breaks = 20)

hist(rbeta(100,a,b),freq = FALSE, breaks = 20)

hist(rbeta(500,a,b),freq = FALSE, breaks = 20)

hist(rbeta(1000,a,b),freq = FALSE, breaks = 20)

#b).
dist <-rbeta(n = 10000,shape1 = 2+5,shape2 = 2+15)
print("simulated probability")
sum(dist>0.3)/length(dist)
print("exact probability")
1-pbeta(0.3,shape1 = 2+5,shape2 = 2+15)

#c).
log_odds<-log(dist/(1-dist))
hist(log_odds)
density(log_odds)

#2.
#a).
obs <- c(44,25,45,52,30,63,19,50,34,67)

n <- 10000
tau2 <- sum((log(obs)-3.7)^2)/(length(obs))
```

```
posterior <- ((length(obs))*tau2)/rchisq(n = n, df = length(obs))
hist(posterior)

#We compare mean of simulated posterior and theoretical value
#simulated:
print("Simulated mean")
mean(posterior)
print("Simulated variance")
var(posterior)

#theoretical:
print("Theoretical mean")
(length(obs)*tau2)/(length(obs)-2)
print("Theoretical variance")
(2*length(obs)^2*tau2^2)/(((length(obs)-2)^2) * (length(obs)-4))

#b).
G <- 2* pnorm(q = sqrt(posterior)/sqrt(2), mean = 0, sd = 1)-1
hist(G, breaks = 30)

#c).
#90% equal tail credible interval
G_sort <- sort(G)
G_cut <- G_sort[(length(G_sort)*0.05):(length(G_sort)*0.95-1)]
hist(G, breaks = 30)
abline(v=min(G_cut), col = "red")
abline(v=max(G_cut), col = "red")

abc <- density(G)
cdf <- data.frame(abc$x, abc$y)
cdf$abc.y <- cdf$abc.y / sum(cdf$abc.y)
cdf <- cdf[order(cdf$abc.y),]
cdf$y <- cumsum(cdf$abc.y)
cdf_90 <- cdf[cdf$y > 0.1,]
print('Highest Posterior Density interval is :')

print(range(cdf_90$abc.x))
abline(v=min(cdf_90$abc.x), col = "blue")
abline(v=max(cdf_90$abc.x), col = "blue")
#3.
#a).
wind_dir <- c(40,303,326,285,296,314,20,308,299,296)
radians <- c(-2.44,2.14,2.54,1.83,2.02,2.33,-2.79,2.23,2.07,2.02)

mu <- 2.39
k <- seq(0.001, 10.001, by = 0.001)
n <- length(radians)
ml <- exp(k*sum(cos(radians-mu)))/(besselI(k, 0))^n
prior <- exp(-k)
#posterior <- exp(k*sum(cos(radians-mu))-k)/(besselI(k, 0))^n
posterior <- ml * prior #distribution the same but values very small
plot(k,posterior, type = "l")

#b).
plot(k,posterior, type = "l")
abline(v=k[which(posterior==max(posterior))], col = "red")
```