Computer Lab 1
732A54 - Bayesian Learning

Jooyoung Lee - joole336
Zuxiang Li - zuxli371

April 28, 2020

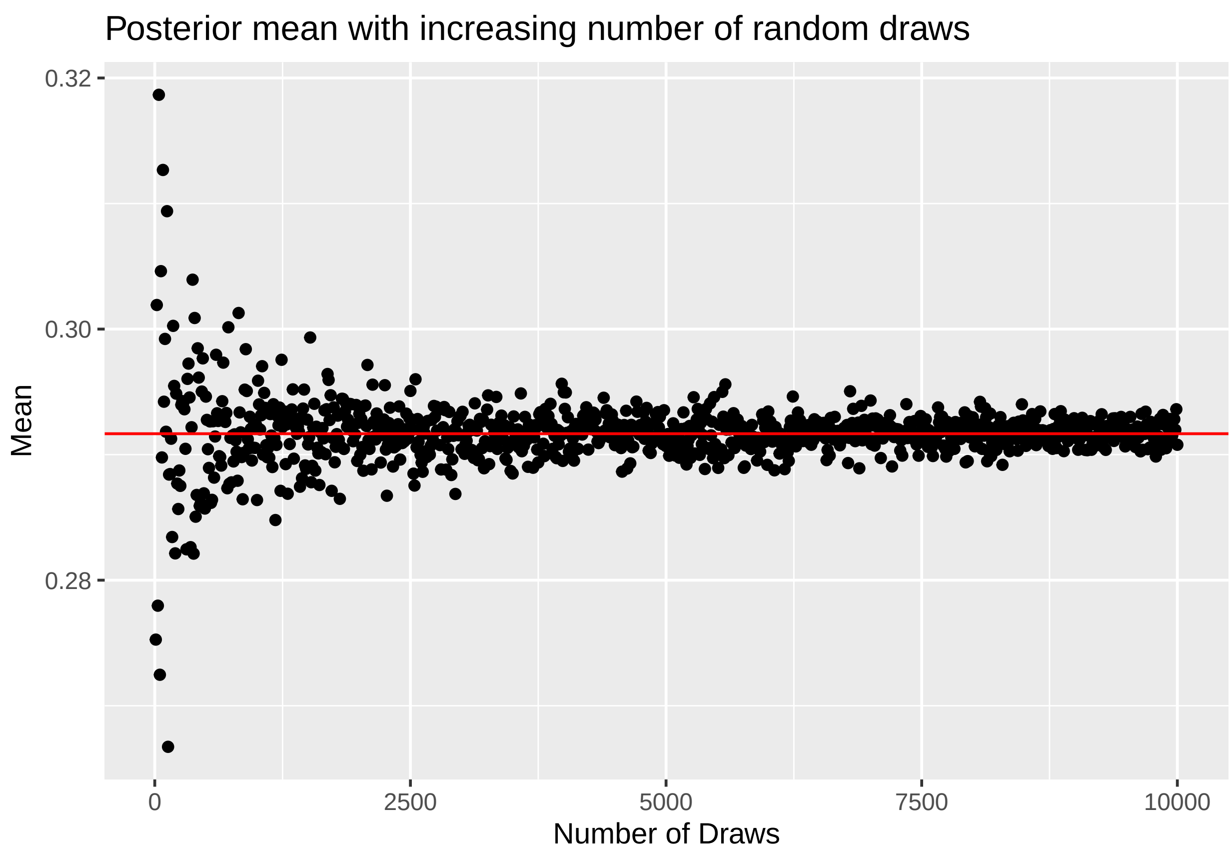
1. Bernoulli ... again.

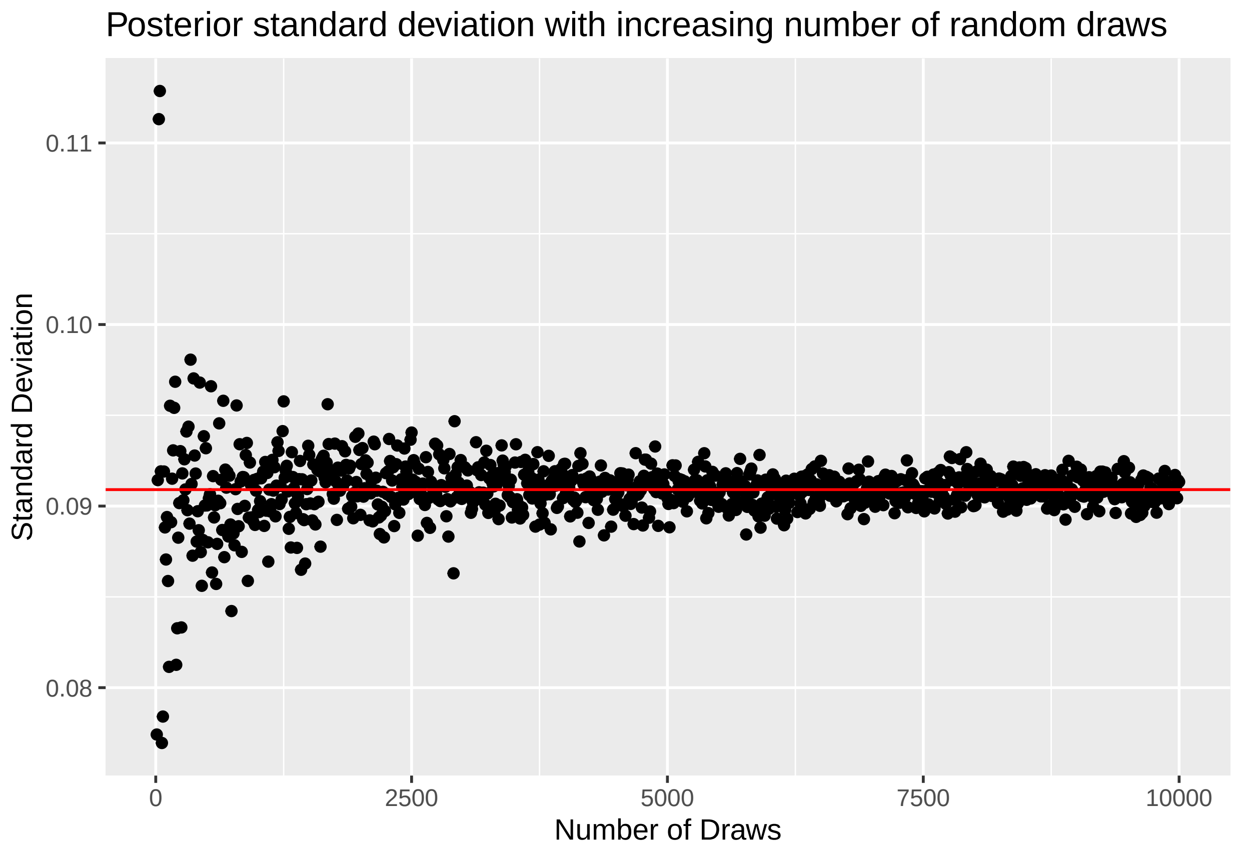
Let $y_1, \cdots, y_n|\theta \; Bern(\theta)$, and assume that you have obtained a sample with $s = 5$ successes in $n = 20$ trials. Assume a $Beta(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 2$.

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

```
## true mean= 0.2916667

##
## true standard deviation= 0.09090593
```





As we can see from the plots above, for both mean and stanard deviation, at first the values are distributed around two side of the red line. But as the number of random draws grows, the posterior mean and standard deviation converge to the red line, which is the true mean and true stanard deviation based on numerical calculation using the posterior distribution.

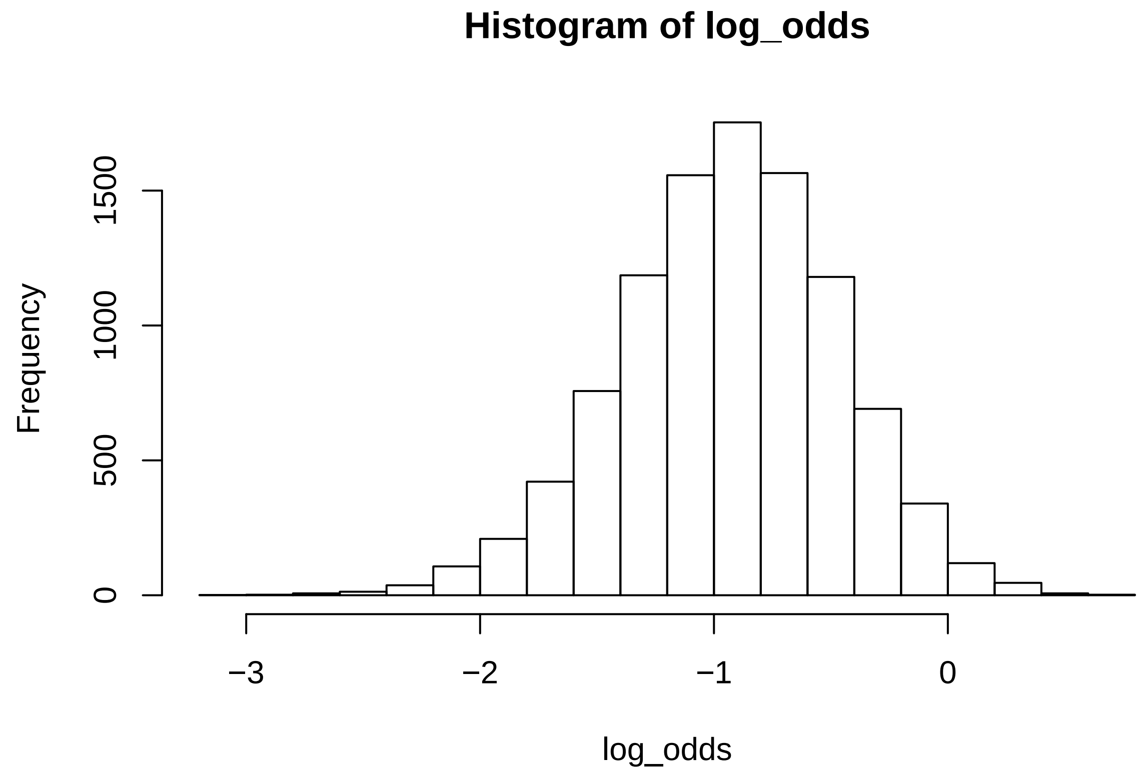
(b) Use simulation (nDraws = 10000) to compute the posterior probability $Pr(\theta > 0.3|y)$ and compare with the exact value [Hint: pbeta()].

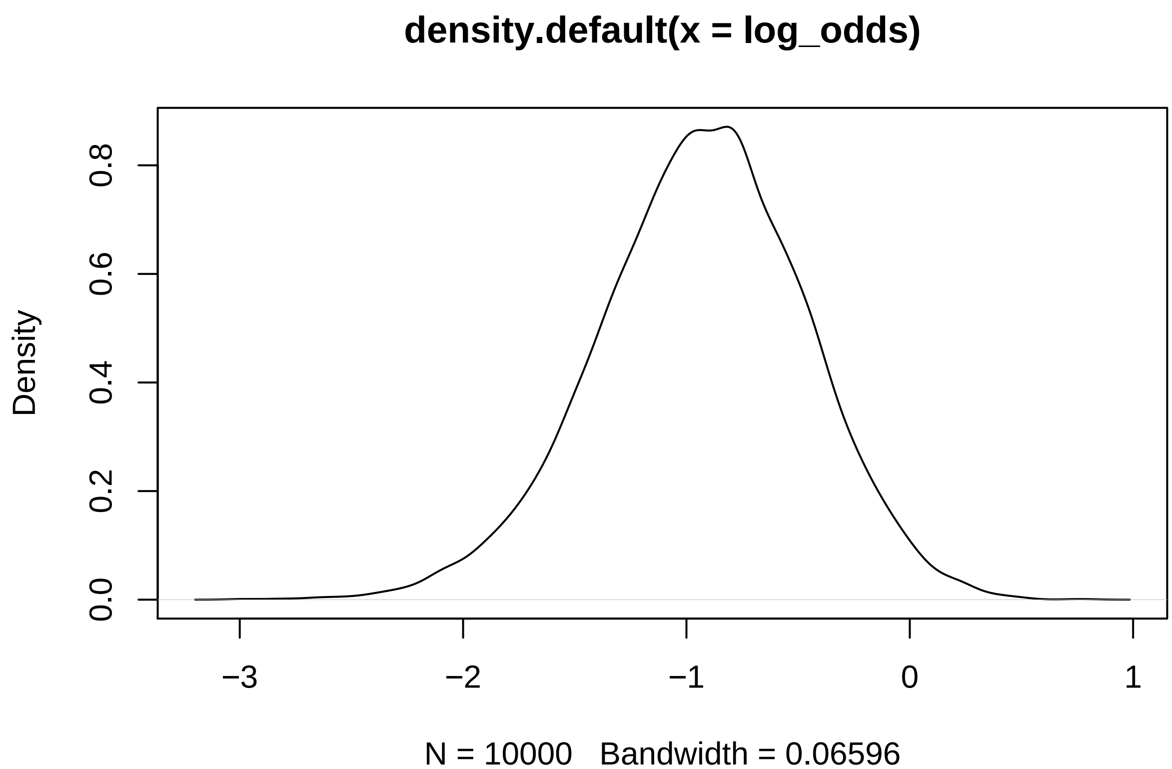
```
## sample prob= 0.4402
```

```
##
## true prob= 0.4399472
```

We use the α and β from (a) with function rbeta(), generated 10000 draws. Then, count the number of thetas which are greater than 0.3 and divide it by the total number of simulated thetas. Function pbeta() is utilized to get the true probability.

(c) Compute the posterior distribution of the log-odds $\phi = log\frac{\theta}{1-\theta}$ by simulation (nDraws = 10000). [Hint: hist() and density() might come in handy].





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

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

for $y > 0, \mu > 0, \sigma^2 > 0$ The log-normal distribution is related to the normal distribution as follows: if $y \ \log N(\mu, \sigma^2)$ then log $N(\mu, \sigma^2)$. Let $y_1, \cdots, y_n|\mu, \sigma^2 \ \log(\mu, \sigma^2)$, 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

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

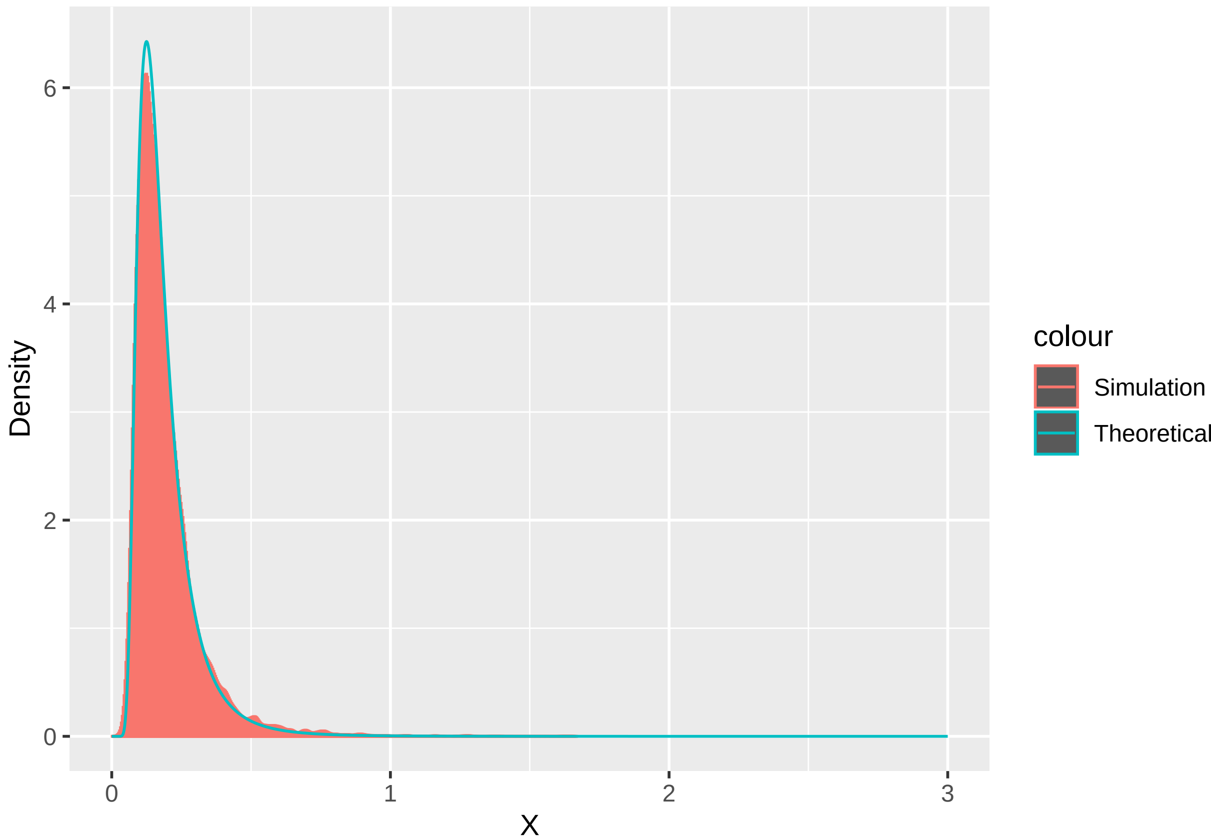
(a) Simulate 10,000 draws from the posterior of σ^2 (assuming $\mu = 3.7$) and compare with the theoretical $Inv - \chi^2(n, \tau^2)$ posterior distribution.

By checking the distribution table, we can find the theoretical mean and standard deviation for Inv-chisq distribution

$$E(\theta) = \frac{1}{v-2}, v > 2$$

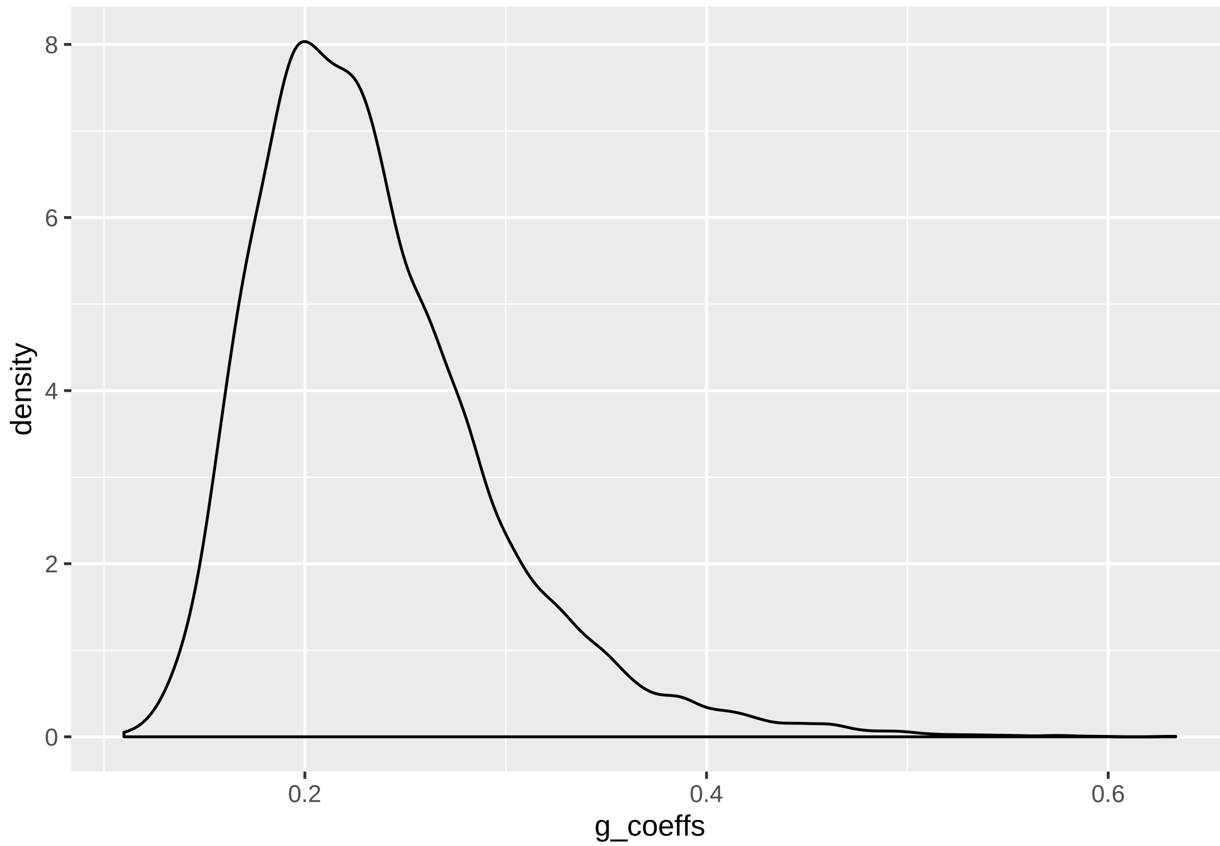
$$var(\theta) = \frac{2}{(v-2)^2(v-4)}, v > 4$$

```
##
## THEORETICAL
## Mean 0.125
##
## Standard deviation 0.005208333
##
## SIMULATION
## Mean: 0.1909554
##
## Standard deviation 0.1164052
```



After plotting the histogram of simulated σ^2 and theoretical probability density of the posterior distribution of σ^2, it is possible to spot that these two are very similar to each other.

(b) The most common measure of income inequality is the Gini coefficient, G, where $0 \leqslant G \leqslant 1$. $G = 0$ 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 $\log N(\mu, \sigma^2)$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.



By generating 10000 draws and plotting the result, we can observe that the simulated Gini-coefficients are right-skewed, which indicates this sample represents income inequality.

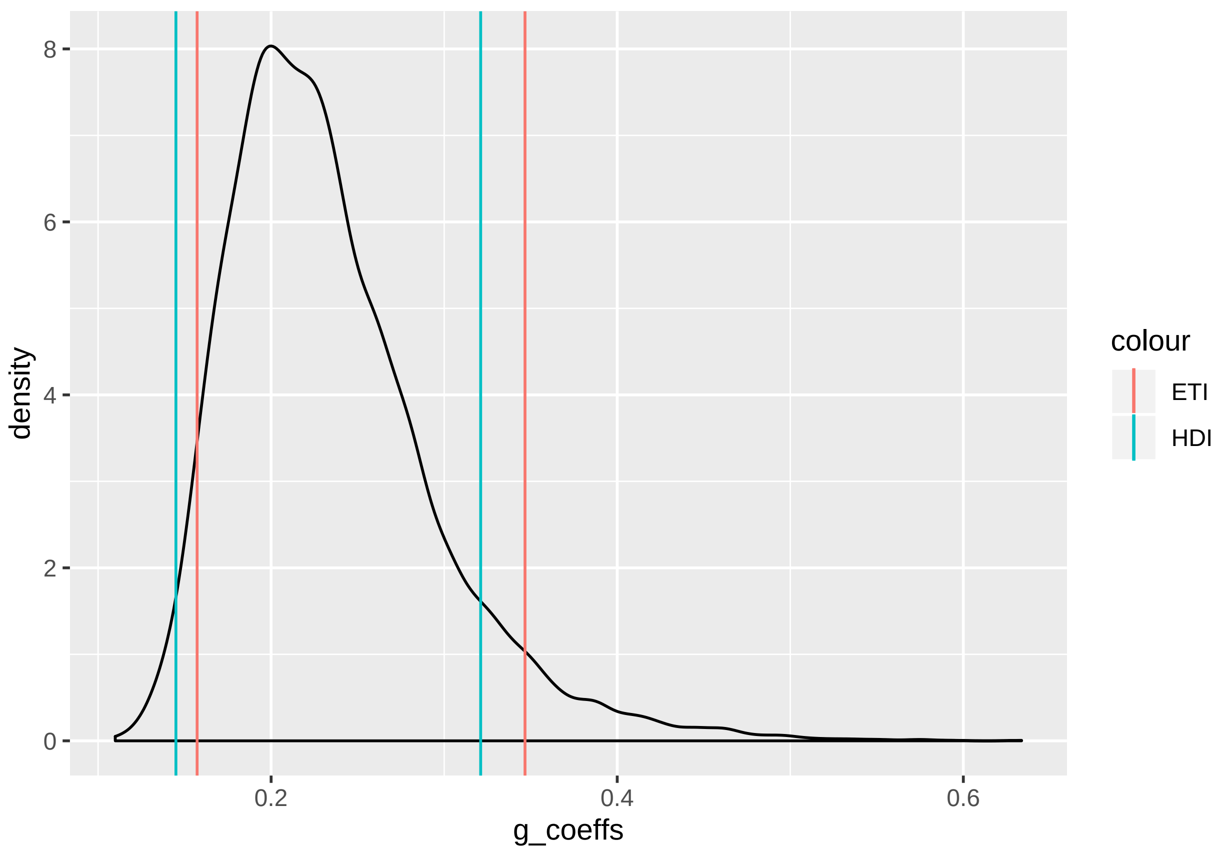
(c) Use the posterior draws from b) to compute a 90% equal tail credible interval for G. A 90% equal tail interval (a; b) cuts off 5% percent of the posterior probability mass to the left of a, and 5% to the right of b. Also, do a kernel density estimate of the posterior of G using the density function 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.

```
## 90% equal tail credible interval is:

## 0.1572372 0.3467172

## 90% Highest Posterior Density interval is:

## 0.1449768 0.3210816
```



By definition, HPD interval includes all values of x "for which the density is at least as big as some value W".(Kruschke, J. K. (2015). Doing Bayesian data analysis: a tutorial with R, Jags, and Stan. Amsterdam: Elsevier/AP.) Therefore it is expected that the end points of the interval to have the same density if the distribution is continuous - and on the graph above, it is possible to see that the end points of HPD interval have the same density.

We can find that in this case, HPD has a narrower interval compare to ETI. This is because the distribution is right-skewed. Right-skewed distribution indicates that more data points are placed on the right side of the peak of the distribution. Since ETI considers amount of data, while HPD interval considers density, it does make sense that ETI has a bigger value for the upper end point of the interval in this right-skewed distribution.

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 $-\pi \leqslant y \leqslant \pi$ 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

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

where $I_0(k)$ is the modified Bessel function of the first kind of order zero [see ?besselI in R]. The parameter $\mu(-\pi \leqslant y \leqslant \pi)$ is the mean direction and k > 0 iscalled the concentration parameter. Large k gives a small variance around μ, and vice versa. Assume that μ is known to be 2:39. Let $k\ Exp(\lambda = 1)$ 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 k for the wind direction data over a fine grid of k values.

Likelihood for von Mises distribution

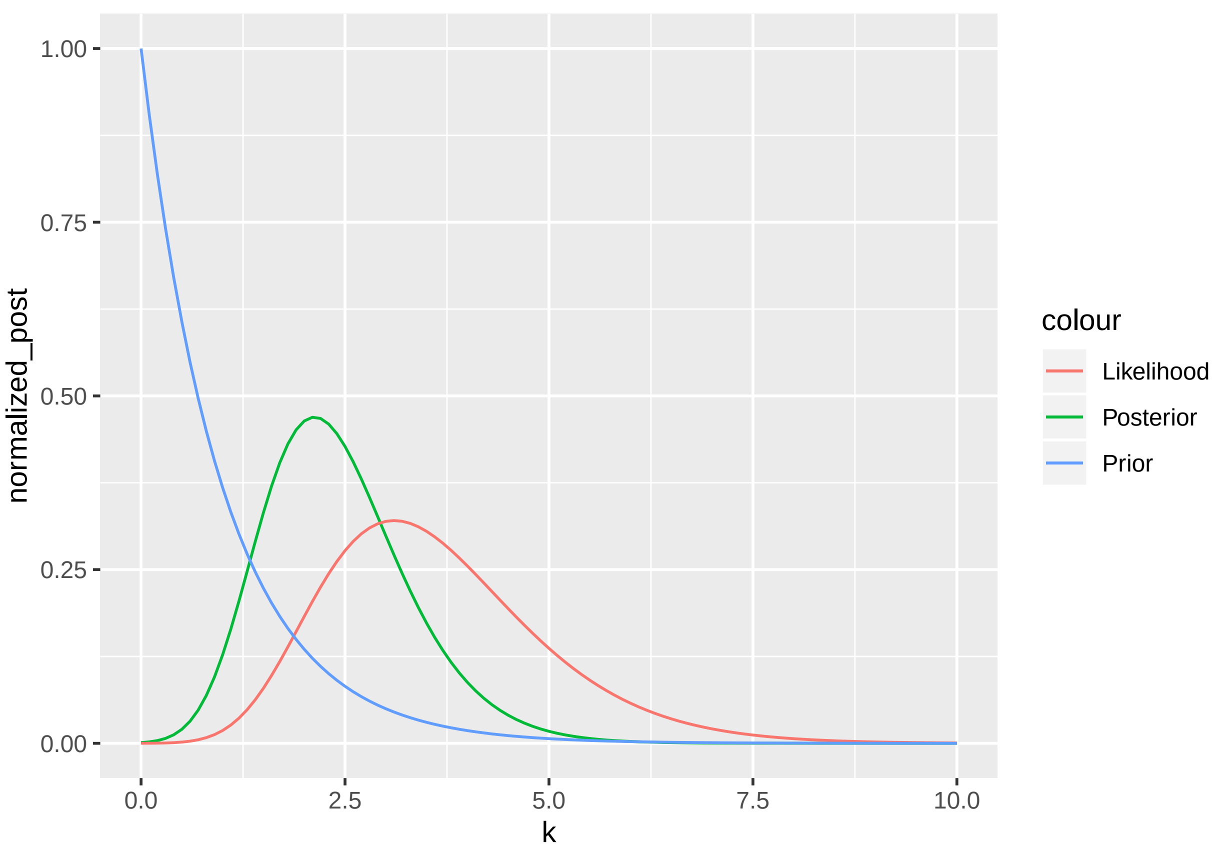
$$p(y|\mu, k) = \prod_{i=1}^{n} p(y|\mu, k)$$
$$= \prod_{i=1}^{n} \frac{exp[k \cdot \cos(y_i - \mu)]}{2\pi I_0(k)}$$
$$= (\frac{1}{2\pi I_0(k)})^n \cdot exp\{\sum_{i=1}^{n} k \cdot \cos(y_i - \mu)\}$$

Prior

$$p(k) \sim Exp[1]$$
$$p(k) = e^{-k}$$

Posterior = Likelihood x Prior

$$p(k|y, u) = p(y|\mu, k) \cdot p(k)$$
$$\propto \frac{1}{I_0(k)}^n \cdot exp(k\sum_{i=1}^{n} \cos(y_i - \mu)) \cdot exp(-k)$$
$$\propto \frac{1}{I_0(k)}^n \cdot exp(k\sum_{i=1}^{n} \cos(y_i - \mu) - k)$$



To show our results simultaneously on one graph, we need to normalize our likelihood and posterior. In order to do this, first, we integrate our posterior and likelihood to obtain the area under the curve. Then we use our likelihood and posterior divided by the value we got to normalize it. After normalizing our likelihood and posterior, we got one graph containing all three densities, which are integrated to 1.

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

```
## Approximate mode of K is:
```

```
## [1] 2.1
```

Mode is, by definition, the value that appears the most. (https://en.wikipedia.org/wiki/Mode_(statistics)) Therefore, the value of k with the highest posterior density will be equivelant to mode of the approximate posterior k.

Appendix A : Code Question 1

```
s=5
f=15
n=20
a0=2
beta0=2
true_a=a0+s
true_beta=beta0+f

means=c()
stds=c()

for(i in seq(10,10000,10)){
  pos=rbeta(i,true_a,true_beta)
  means=c(means,mean(pos))
  stds=c(stds,sd(pos))
}
true_mean=true_a/(true_a+true_beta)
true_var=(true_a*true_beta)/((true_a+true_beta+1)*(true_a+true_beta)^2)
cat("true mean=",true_mean)
cat("\ntrue standard deviation=",sqrt(true_var))
ggplot()+geom_point(aes(x=seq(10,10000,10),y=means))+geom_hline(yintercept=true_mean,col="red")+xlab("Nu
ggplot()+geom_point(aes(x=seq(10,10000,10),y=stds))+geom_hline(yintercept = sqrt(true_var),col="red")+xl
nDraws=10000
res=rbeta(nDraws,true_a,true_beta)
sample_prob=length(res[res>0.3])/nDraws
true_prob=1-pbeta(0.3,true_a,true_beta)
cat("sample prob=",sample_prob)
cat("\ntrue prob=",true_prob)
res=rbeta(nDraws,true_a,true_beta)
log_odds=log(pos/(1-pos))
hist(log_odds)
plot(density(log_odds))
```

Appendix B : Code Question 2

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

# simulating sigmasq from posterior
X=rchisq(10000,n-1)
simu_sigmas_sq=(n-1)*tausq/X
#hist(simu_sigmas,breaks=100,main = "Histogram of simulated sigma square from posterior")

cat("\nTHEORETICAL\n")
cat("Mean",1/(n-2))
cat("\nStandard deviation",2/((n-2)^2*(n-4)))

cat("\nSIMULATION\n")
cat("Mean:",mean(simu_sigmas_sq))
cat("\nStandard deviation",sd(simu_sigmas_sq))

# theoretical
sigmas_sq=dinvchisq(seq(0,3,0.001),nu=n,tau=tausq)

#plot(sigmas_sq,main = "Density of theoretical function")
d=density(simu_sigmas_sq)
ggplot()+geom_col(aes(x=d$x,y=d$y,colour="Simulation"))+geom_line(aes(x=seq(0,3,0.001),y=sigmas_sq,colo

# plot(x=1:10000, y=sigmas, main="Comparing Simulated Values and Real Value", xlab="Trial", ylab="Simul
# theoretical_val=n*tausq/(n-2)
# abline(h=theoretical_val, col="red", lwd = 2)
# legend("topright", legend = "Theoretical Sigma-Squared", col="red", lty=1, lwd=2)
# df=data.frame(Simulation_Mean=mean(sigmas), Theoretical_Sigmasq = theoretical_val, Difference_in_Perc
# print(df)
g_coeffs = 2*pnorm(sqrt(simu_sigmas_sq)/sqrt(2)) -1
#hist(g_coeffs, main="Histogram of Gini Coefficients from Simulation", xlab="Simulated Gini-coefficient
ggplot()+geom_density(aes(g_coeffs))
d=density(g_coeffs)

ETI_x=d$x
ETI_y=d$y
ETI_unsorted=cumsum(ETI_y)/sum(ETI_y)

lower <- which(ETI_unsorted>=0.05)[1]
upper <- which(ETI_unsorted>=0.95)[1]
cat("90% equal tail credible interval is:", "\n")
ETI=d$x[c(lower,upper)]
cat(ETI)

df=data.frame(x=d$x,y=d$y,idx=c(1:length(d$x)))
df=df[order(df$y,decreasing = TRUE),]
HPD_sorted=cumsum(df$y)/sum(df$y)
df$HPD_sorted=HPD_sorted
```

```
inx=df$idx[df$HPD_sorted<0.9]
lower=min(inx)
upper=max(inx)
cat("90% Highest Posterior Density interval is:", "\n")
HPD=c(d$x[lower],d$x[upper])
cat(HPD)

ggplot()+geom_density(aes(x=g_coeffs))+geom_vline(aes(xintercept =ETI[1],colour="ETI"),size=0.5)+geom_vl
```

Appendix C : Code Question 3

```
mu=2.39
data=c(-2.44,2.14,2.54,1.83,2.02,2.33,-2.79,2.23,2.07,2.02)

k=seq(0,10,0.1)

posterior=function(k){
  return(exp(k*sum(cos(data-mu))-k)*besselI(k,0)^(-length(data)))
}
llik_fn=function(k){
  return(exp(k*sum(cos(data-mu)))*besselI(k,0)^(-length(data)))
}
prior_fn=function(k){
  return(exp(-k))
}
inte=integrate(posterior,lower = 0,upper = 10)
prior=prior_fn(k)
llik=exp(k*sum(cos(data-mu)))*(besselI(k,0))^(-length(data))
normalized_llik=llik/integrate(llik_fn,0,10)$value
df=data.frame(k=k,normalized_post=(posterior(k)/inte$value),llik=normalized_llik,prior=prior)
ggplot(data=df)+geom_line(aes(x=k,y=normalized_post,colour="Posterior"))+geom_line(aes(x=k,y=llik,colou
#ggplot()+geom_line(aes(x=k,y=llik))
cat("Approximate mode of K is: ", "\n")
k[which.max(posterior(k))]
```