Beysian Lab11

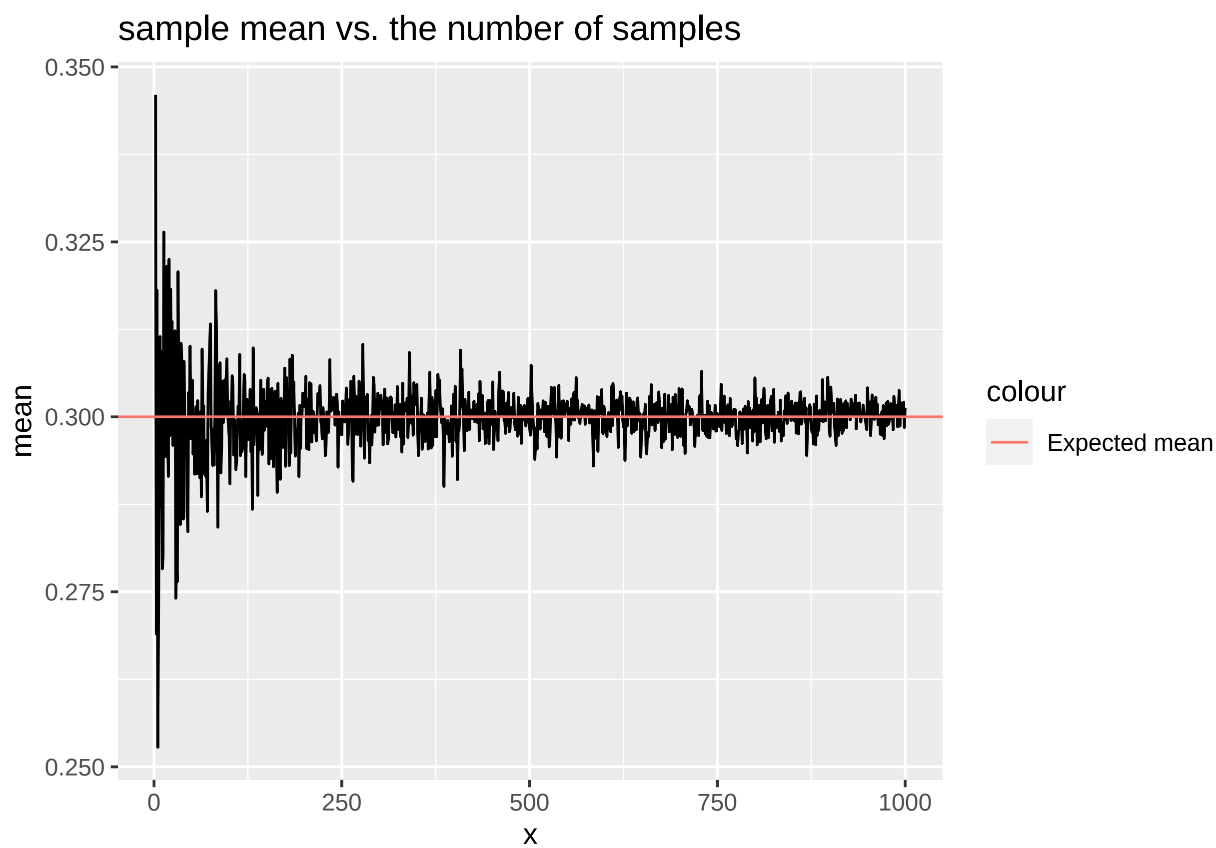
Wuhao Wang

4/7/2022

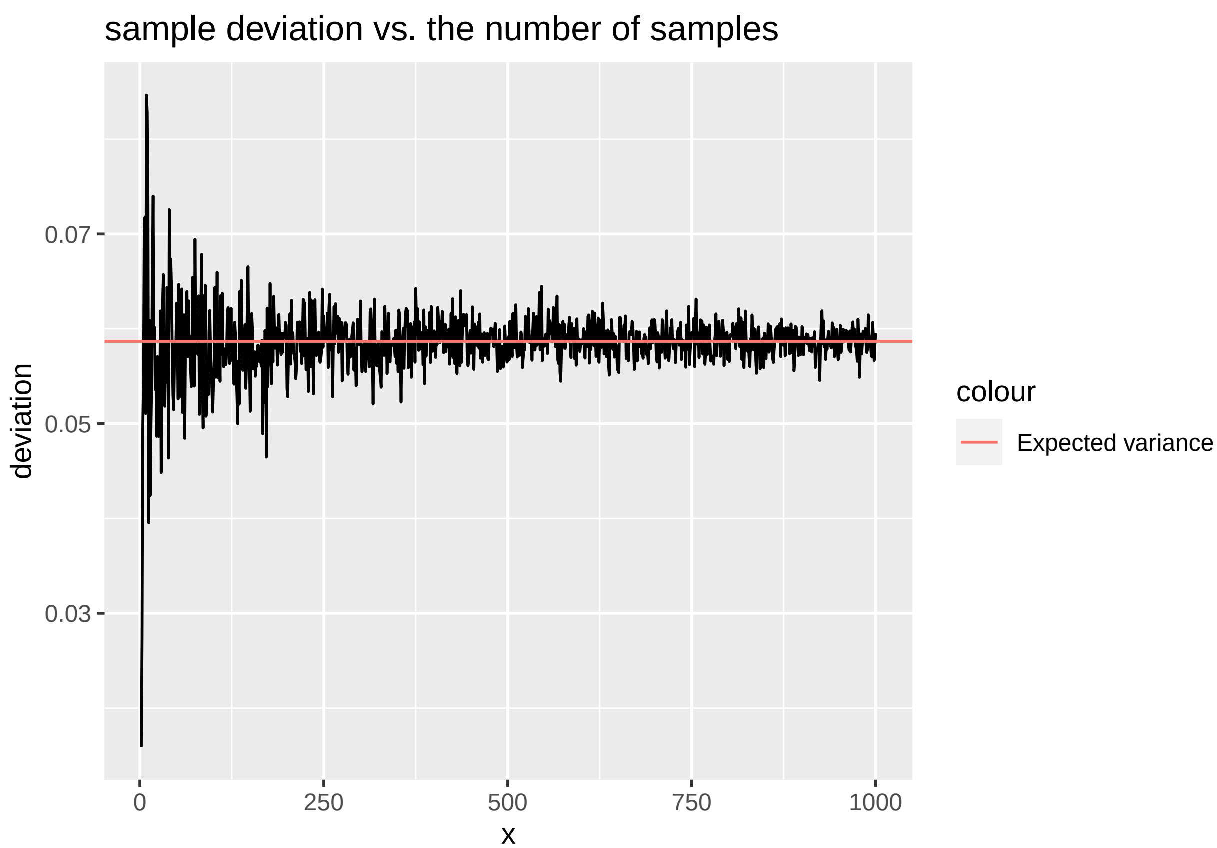
Q1 Daniel Bernoulli

question a)

```
set.seed(12345)
alpha = 5
beta = 5
n = 50
s = 13
f = n - s
alpha = alpha + s
beta = beta + f
exp_mean = alpha/(alpha+beta)
exp_var =  sqrt(alpha*beta/(alpha+beta)**2/(alpha+beta+1))
n_list = 2:1000
sample_mean = c()
sample_var = c()
for (i in n_list) {
  sample_mean = c(sample_mean,mean(rbeta(i,alpha,beta)))
  sample_var = c(sample_var,sqrt(var(rbeta(i,alpha,beta))))
}
df = data.frame(x=n_list,mean=sample_mean,deviation=sample_var)
p1 = ggplot(df)+
  geom_line(aes(x,mean))+
  ggtitle('sample mean vs. the number of samples')+
  geom_hline(aes(yintercept = exp_mean,color = 'Expected mean'))
print(p1)
```




```
p2 = ggplot(df)+
  geom_line(aes(x,deviation))+
  ggtitle('sample deviation vs. the number of samples')+
  geom_hline(aes(yintercept = exp_var,color = 'Expected variance'))
print(p2)
```



question b)

```
nDraws = rbeta(10000,alpha,beta)
expected_p = pbeta(0.3,alpha,beta)
sample_p = sum(nDraws<0.3)/length(nDraws)
print(paste0('the expected probility is : ',expected_p,'.'))
```

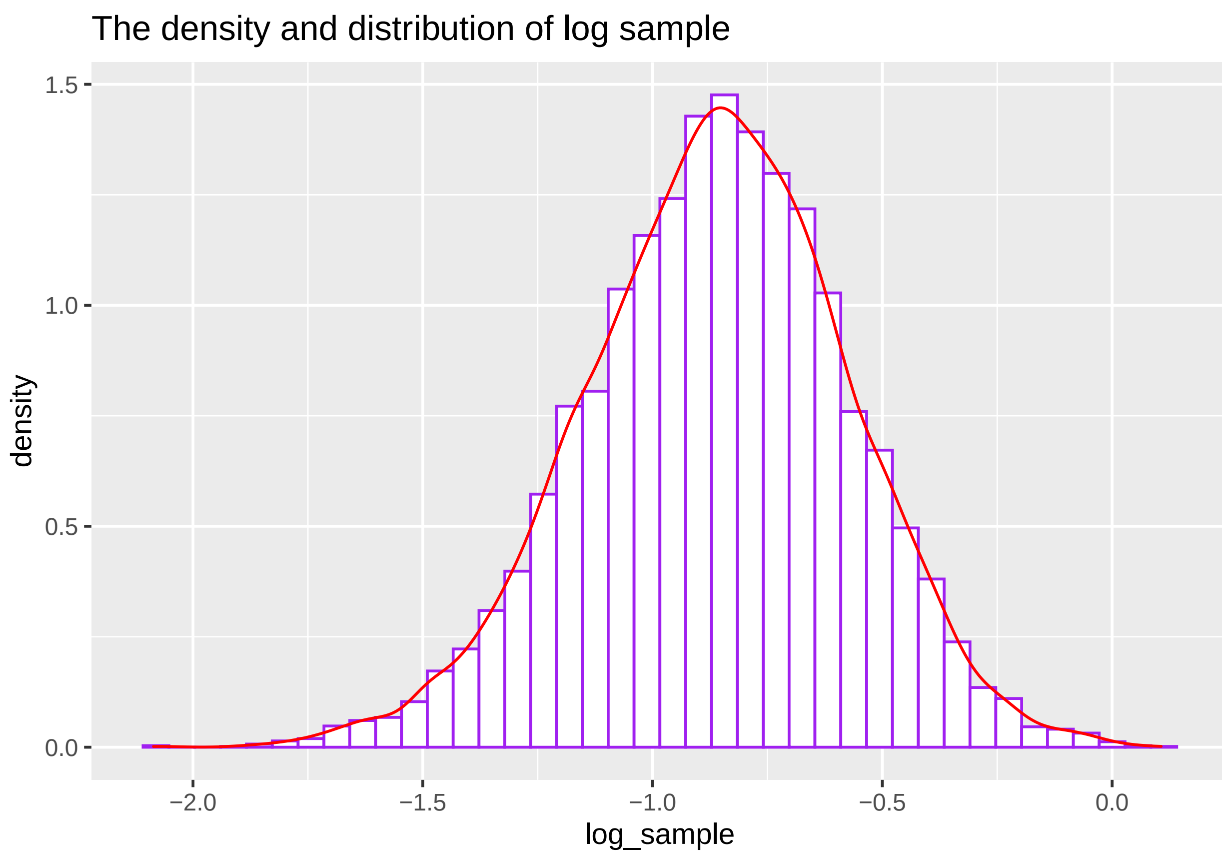
```
## [1] "the expected probility is : 0.515022634950062."
```

```
print(paste0('the calculated probility is : ',sample_p,'.'))
```

```
## [1] "the calculated probility is : 0.5121."
```

question c)

```
log_sample = log(nDraws/(1-nDraws))
df = data.frame(n=1:10000,theta=log_sample)
p3 = ggplot(df,aes(log_sample))+
  geom_histogram(aes(y=..density..),bins=40,color='purple',fill='white')+
  geom_density(alpha=0.3,color='red')+
  ggtitle('The density and distribution of log sample')
print(p3)
```

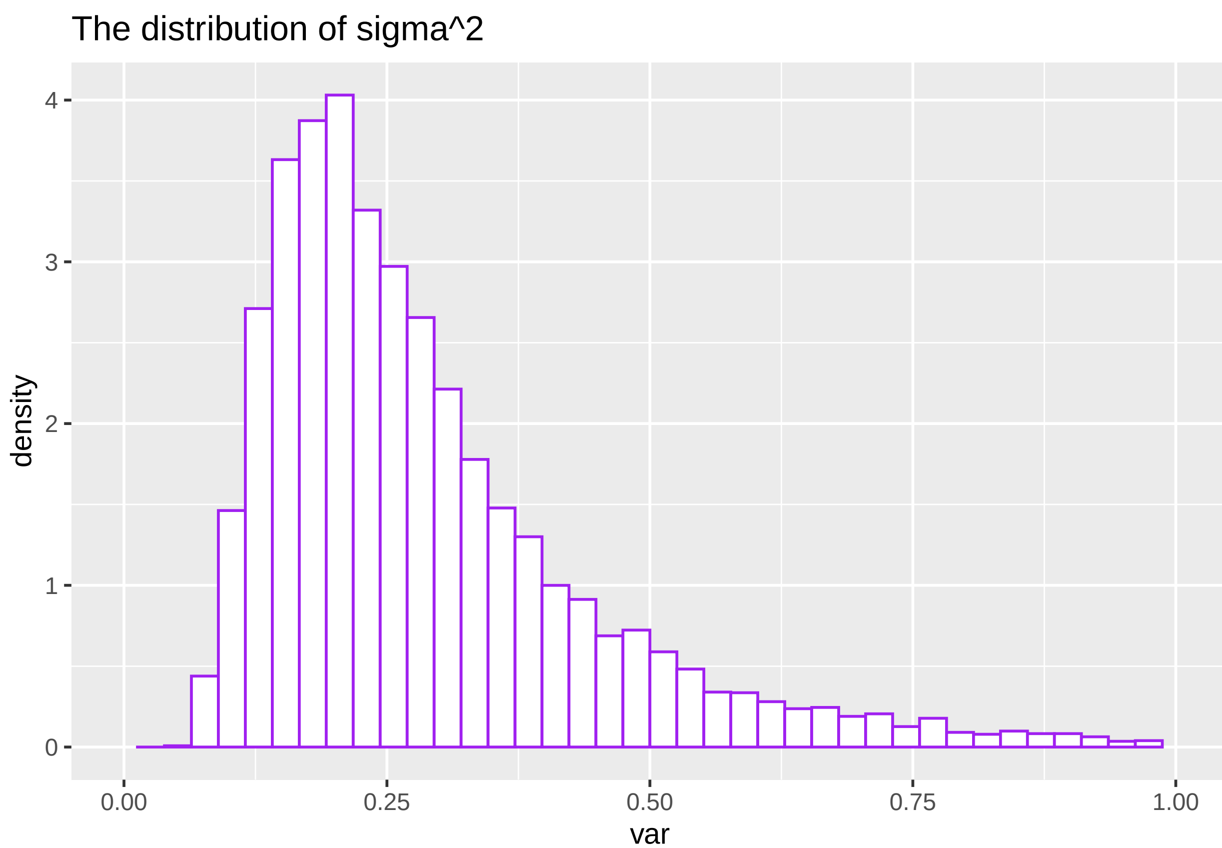


Q2 Log-normal distribution and the Gini coefficient

question a)

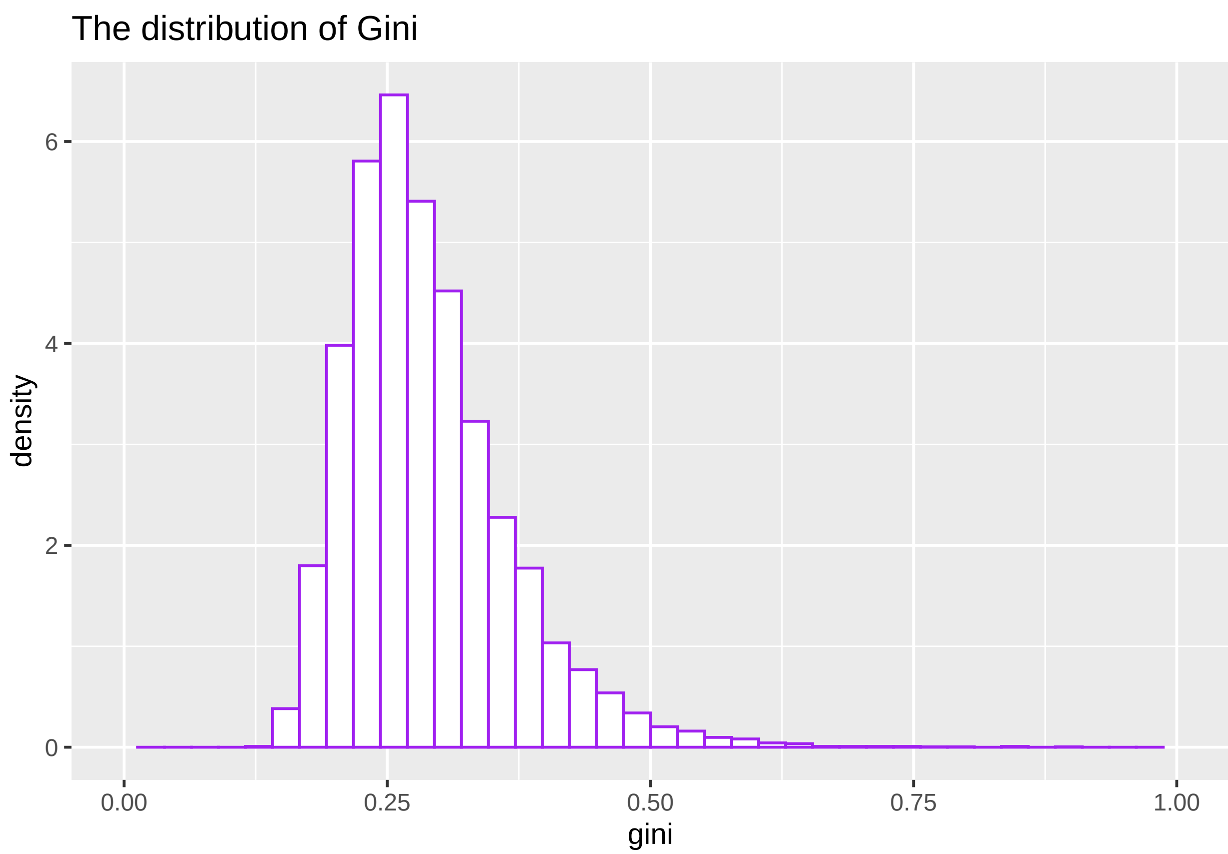
```
sample = c(33,24,48,32,55,74,23,76)
len = length(sample)
Myscale<-function(input,mean_value=3.5)
{
  return(sum(log(input)-mean_value)/length(input))
}
sample_scaled = Myscale(sample)
#print(paste0('The expected mean value is:', len*sample_scaled/(len-2)))
#print(paste0('The expected variance value is:', len*sample_scaled/(len-2)))
cal_var = c()
for (i in 1:10000)
{
  temp = rchisq(1,len)
  cal_var = c(cal_var,len*sample_scaled/temp)
}
df = data.frame(n=1:10000,var=cal_var)
p4 = ggplot(df,aes(var))+
  geom_histogram(aes(y=..density..),
                 color = 'purple',
                 fill = 'white',
```

```
                 bins = 40
                 )+
  ggtitle('The distribution of sigma^2')+
  scale_x_continuous(limits = c(0,1))
print(p4)
```



question b)

```
Gini_coe = 2*pnorm(sqrt(cal_var/2),mean = 0,sd=1)-1
df = data.frame(n=1:10000,gini=Gini_coe)
p5 = ggplot(df,aes(gini))+
  geom_histogram(aes(y=..density..),
                 color = 'purple',
                 fill = 'white',
                 bins = 40
                 )+
  ggtitle('The distribution of Gini')+
  scale_x_continuous(limits = c(0,1))
print(p5)
```



question c)

We have 10000 data in total. If we sort all the Gini-coefficients, we can find the number located in 250 and 9750. The interval between these two number is the request confidence interval.

```
sort_Gini = sort(Gini_coe)
print(paste0('The 95% confidence interval is: (',round(sort_Gini[250],3),',',round(sort_Gini[9750],3),'
```

```
## [1] "The 95% confidence interval is: (0.178,0.477)"
```

question d)

We use the function `density()` to find the estimated density value. The result contains 512 values. We first sort all the values and find the threshold which ranks 25 (512*0.05). Then according to the definition, all the density value within the interval will be greater then this threshold.

```
d=density(Gini_coe)
estimated = d$y
sort_estimated = sort(estimated)
threshold = sort_estimated[floor(length(sort_estimated)*0.05)]
interval=c()
for (i in 1:(length(d$x)-1))
{
  if((d$y[i]-threshold)*(d$y[i+1]-threshold)<=0)
```

```
    interval=c(interval,d$x[i])
}
print(interval[1:2])
```

```
## [1] 0.1097022 0.8611178
```

From the result above, we can see that the interval from HPDI is wider than previous one.

Q3 Bayesian inference

question a)

The posterior is proportional to prior multiplying likelihood.

According to material, T

he prior is exponentional distribution.

$$\pi(x) = \lambda exp(-\lambda x)$$

Here we have $\lambda = 1$,so the prior is

$$\pi(k) = exp(-k)$$

The likelihood is coming from Mises distribution

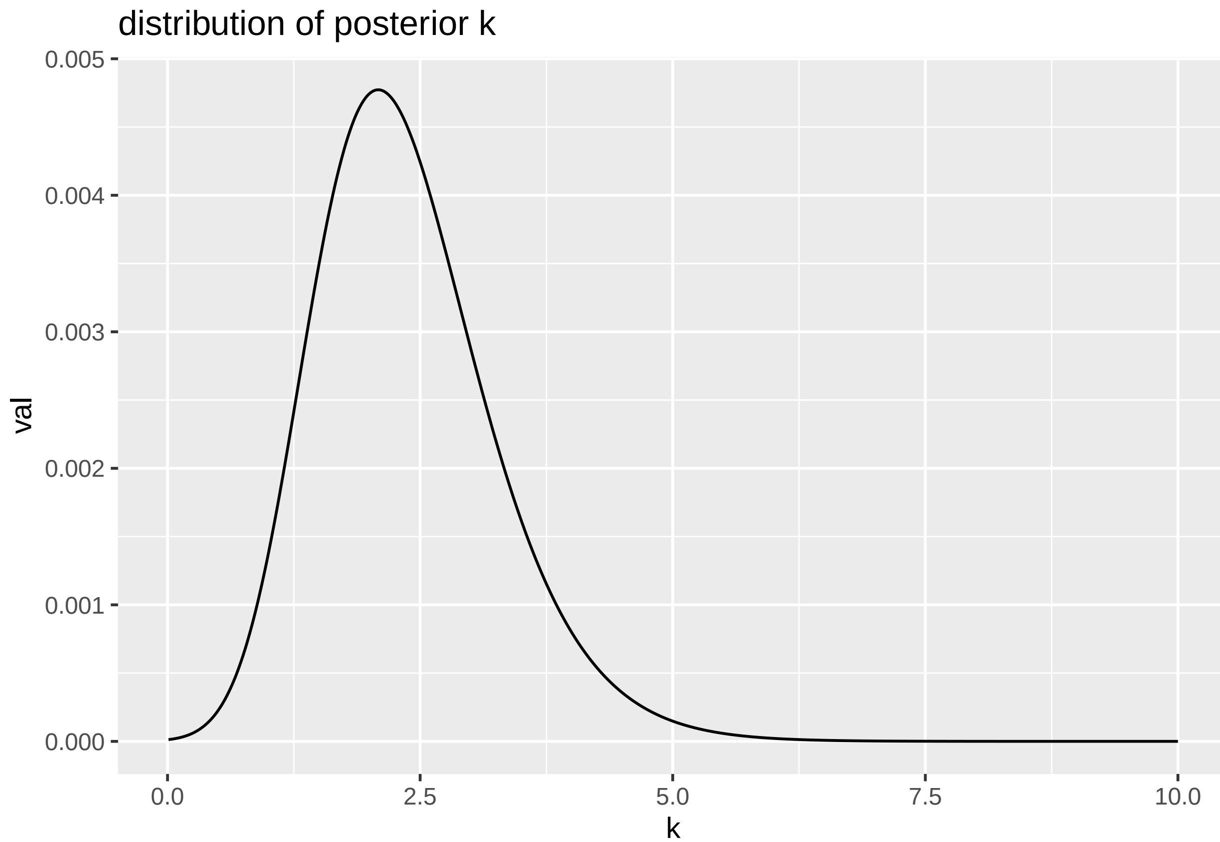
$$\begin{aligned}
L(y|k,\mu) &= \prod_{i=1}^{n} \frac{1}{2\pi} \frac{1}{I_0(k)} exp(k * cos(y_i - \mu)), -\pi \leq y \leq \pi \\
&= \frac{1}{(2\pi)^n} \frac{1}{I_0(k)^n} exp(k \sum_{i=1}^{n} cos(y_i - \mu))
\end{aligned}$$

So, the posterior would be:

$$\begin{aligned}
p(k|y,\mu) &\propto L(y|k,\mu) * \pi(k) \\
&\propto \frac{1}{I_0(k)^n} * exp(k \sum_{i=1}^{n} cos(y_i - \mu)) * exp(-k) \\
&= \frac{1}{I_0(k)^n} * exp((k \sum_{i=1}^{n} cos(y_i - \mu)) - k)
\end{aligned}$$

Now, we can plot the distribution of the posterior of k.

```
d = c(1.83,2.02,2.33,-2.79,2.07,2.02,-2.44,2.14,2.54,2.23)
k = (1:1000)/100
mu = 2.51
res = exp(k*sum(cos(d-mu))-k)/besselI(x = k, nu=0)^length(d)
res = res/sum(res)
df = data.frame(k=k,val=res)
p6 = ggplot(df, aes(k,val))+
  geom_line()+
  ggtitle('distribution of posterior k')
print(p6)
```



question b)

The posterior mode is the point with maximum posterior probability density value .

So the expected k is from

$$\hat{k}_m = argmax_k p(k|y,\mu)$$

```
print(k[which.max(res)])
```

```
## [1] 2.09
```