Bayesian Learning: Computer Lab 1

Mowniesh Asokan(mowas455), Shahrukh Iqbal(shaiq681)

April 14, 2021

1 Daniel Bernoulli

1.1 (a)

Given, the posterior is $\theta|y \sim \text{Beta}(\alpha_0 + s, \beta_0 + f)$ where, $y = (y_1, ..., y_n)$,$\alpha_0 = \beta_0 = 3$ and $s = 8, f = 16$.

Also, the theoretical value of Posterior Mean($\text{E}(\theta|y)$) and standard deviation($\sqrt{\text{var}(\theta|y)}$) is given as,

$$\text{E}(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n} \tag{1}$$

$$\sqrt{\text{var}(\theta|y)} = \sqrt{\frac{\text{E}(\theta|y)[1 - \text{E}(\theta|y)]}{\alpha + \beta + n + 1}} \tag{2}$$

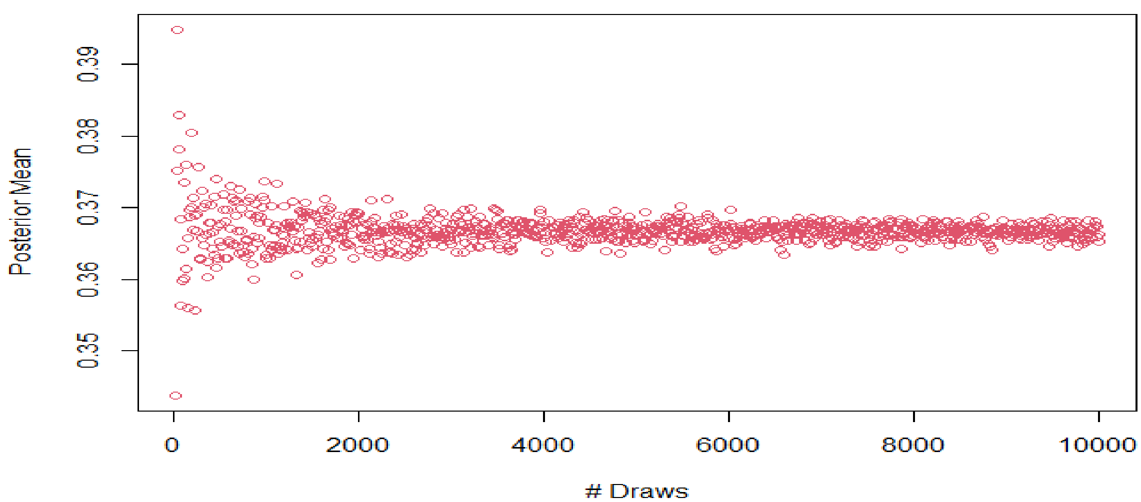

Figure 1: Posterior Mean

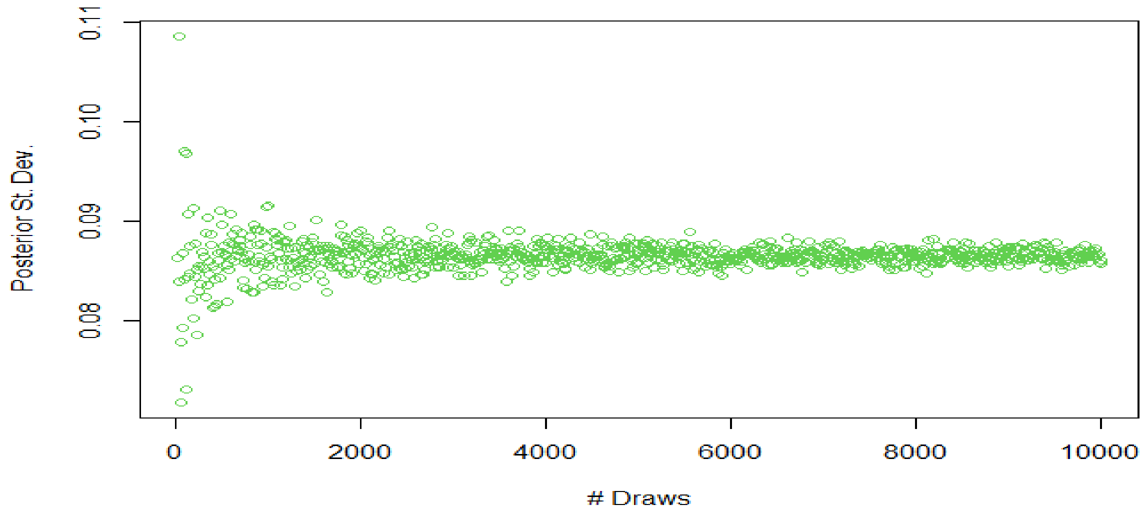

Figure 2: Posterior Standard Deviation

Using results from (1) and (2), we compute the statistical summaries of the posterior distribution i.e., $\mathrm{E}(\theta|y) = 0.366$ and $\sqrt{\mathrm{var}(\theta|y)} = 0.0865$. Next, we analytically verify the results.

In the experiment, we perform a computer function call to calculate posterior mean and standard deviation of an array of one-thousand random draws ranging from 10 to 10000 samples per draw. We present the findings in **Figure 1** and **Figure 2**. We observe that on increasing the sample-size the statistical summaries for the parameter θ converge to the theoretical summaries.

1.2 (b)

We implement an `R` function called as `simulation` to compute the posterior probability of the event $\Pr(\theta > 0.4|\mathrm{y})$ or equivalently $[1 - \Pr(\theta \leq 0.4|\mathrm{y})]$.
We analytically obtain the value of $\Pr(\theta > 0.4|\mathrm{y}) = 0.3406$ and the theoretical value of $\Pr(\theta > 0.4|\mathrm{y}) = 0.3426$. Thus, on comparing both the values we can state that the simulated draws for `nDraws=10000` is nearly equivalent to the expected theoretical value.

1.3 (c)

We compute the posterior distribution of log-odds ϕ where ϕ represents the logistic transformation of θ from $(0, 1)$ to $(-\infty, \infty)$ in the parameter space. The computation is performed through `R` function called as `log.odds`. We present the results in **Figure 3**. On visual inspection, we observe that majority of the distribution for ϕ lies between the support *-1.5* and *0.5*

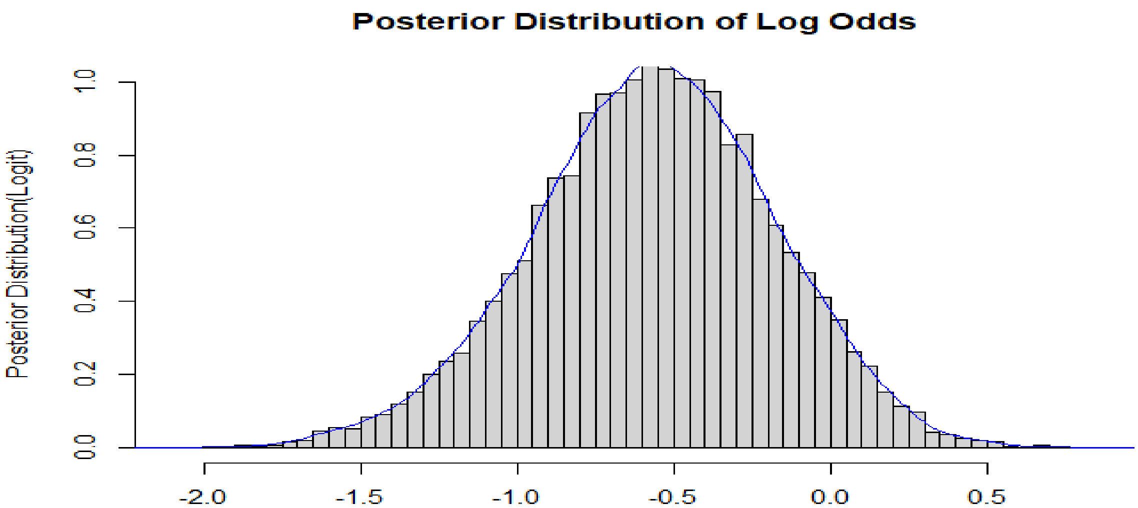

Figure 3: Distribution of ϕ

2 Log-normal distribution and the Gini coefficient

2.1 (a)

In the current problem, firstly we simulate the posterior of σ^2 with $\mu = 3.8$ and compare it's results summary of position and variation with theoretical Inv - $\chi^2(n, \tau^2)$.
Given, $p(\theta) \sim Inv$ - $\chi^2(n, \tau^2)$,

$$\mathrm{E}(\theta) = \frac{n}{n-2}\tau^2 \tag{3}$$

$$\mathrm{var}(\theta) = \frac{2n^2}{(n-2)^2(n-4)}\tau^4 \tag{4}$$

Using results (3) and (4) from [1, p.577], we calculate the theoretical mean and variance for Inv - $\chi^2(n, \tau^2)$. The results are as shown in **Table 1** below. Based on the observations in **Table 1**,it seems that the simulated values are able to generalize the posterior distribution well.

	Simulated Statistic	Theoretical Statistic
$E(\sigma^2)$	0.3273	0.3263
$var(\sigma^2)$	0.03596	0.03549

Table 1: Comparison of Simulated and Theoretical summaries of *Inv* - $\chi^2(n, \tau^2)$

2.2 (b)

In the current problem Gini coefficient, G is calculated using equation (5) for the sampled values of posterior σ.

$$G = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1 \tag{5}$$

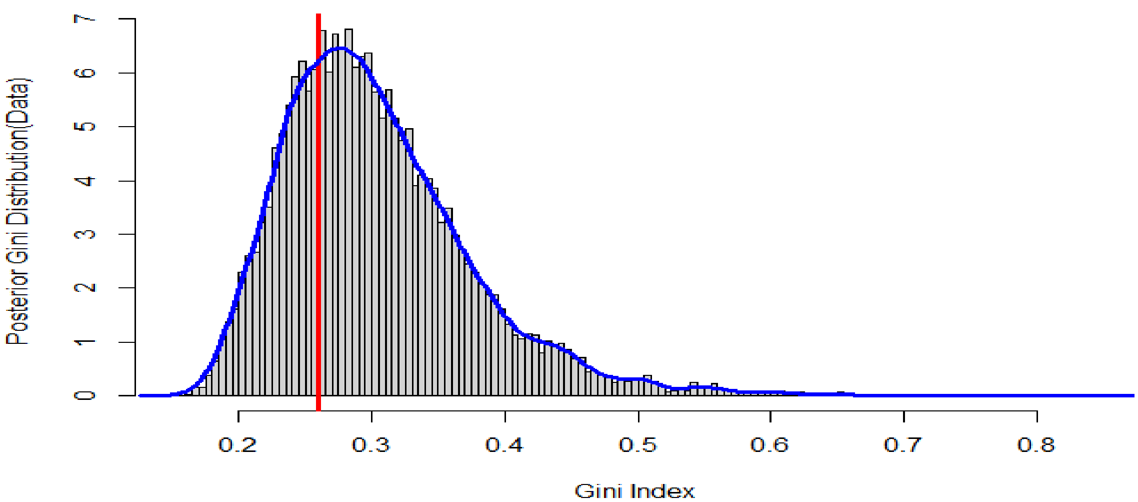

Figure 4: Posterior Distribution of Gini-Index G using current data

From **Figure 4**, we can infer that the distribution is skewed towards 0 i.e., G's distribution suggests that the evidence for income inequality is very low.

2.3 (c)

The Posterior draws from **(b)** is used to generate a 90% equal tail credible interval for G. The summary of posterior uncertainty using the central 90% interval is presented in **Figure 5**. Meanwhile, **Figure 6** shows the posterior uncertainty using the highest posterior density region. The *highest posterior density region* is calculated using the `R` function `hdi` [2]. On comparing the results from both the figures, we observe that the regions do not differ substantially. This is expected as the posterior distribution is unimodal and not highly skewed.

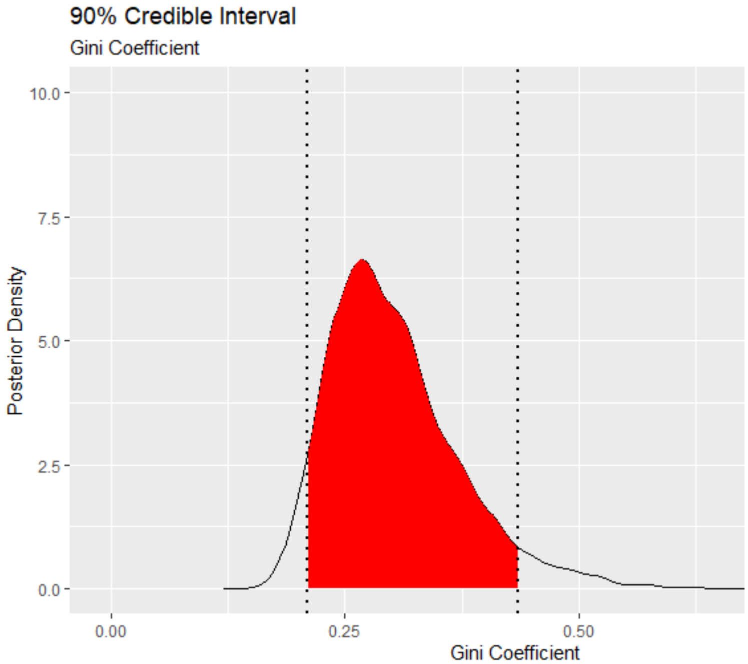

Figure 5: Central Posterior Interval

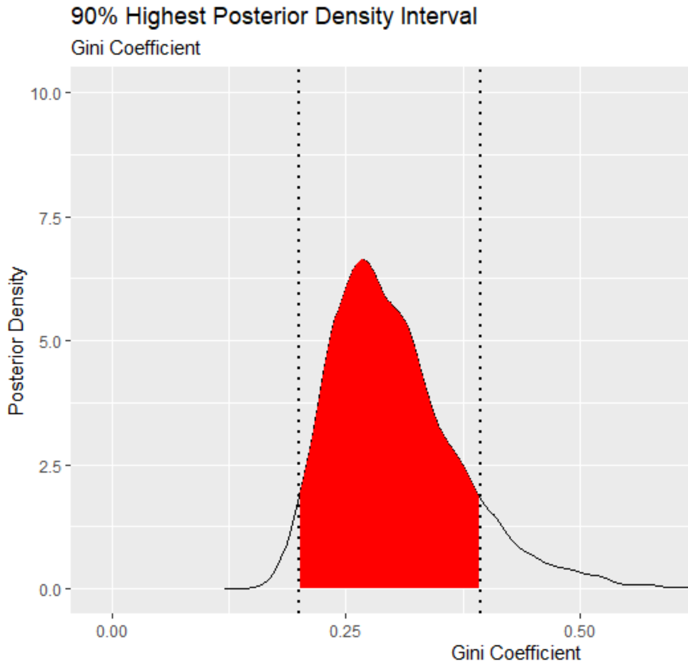

Figure 6: Highest Posterior Density Region

3 Bayesian inference for the concentration parameter in the von Mises distribution

3.1 (a)

In the current problem, we obtain the posterior distribution of κ over a grid of values that are generated from its prior distribution. We are provided with the likelihood of a single data-point as,

$$\mathrm{p}(\mathrm{y}|\theta) = \frac{exp\left[\kappa.cos(y-\mu)\right]}{2\pi I_{\circ}(\kappa)} \tag{6}$$

where $I_0(\kappa)$ is the modified Bessel function of order zero. Using (6) and assuming all data-points to be independent observations, we obtain the complete data likelihood as,

$$\mathrm{p}(\mathrm{y}_1..., \mathrm{y}_{10}|\mu, \theta) \overset{\text{i.i.d}}{=} \prod_{i=1}^{i=10} \frac{exp\left[\kappa.cos(y_i-\mu)\right]}{2\pi I_{\circ}(\kappa)} \qquad \{\text{Given: } \mu = 2.39\} \tag{7}$$

Also, it is given that κ has an Exponential($\lambda = 1$) prior distribution. Therefore, we can parametrize the family of κ as,

$$\begin{aligned} &\kappa \sim \text{Exponential}(\lambda = 1) \\ &\mathrm{p}(\kappa) = \lambda \exp(-\lambda\kappa) \qquad \{\text{Given: } \lambda = 1\} \end{aligned} \tag{8}$$

From Bayes Theorem, we can obtain the *unnormalized posterior density* of κ as,

$$\begin{aligned} \mathrm{p}(\kappa|\mathrm{y}_1..., \mathrm{y}_{10}) &\propto \mathrm{p}(\kappa).\mathrm{p}(\mathrm{y}_1...\mathrm{y}_{10}|\mu, \kappa) \\ &= exp(-\lambda\kappa).\prod_{i=1}^{i=10} \frac{exp\left[\kappa.cos(y_i - \mu)\right]}{2\pi I_\circ(\kappa)} \\ &= exp(-\lambda\kappa).\frac{exp\left[\kappa.\sum_{i=1}^{i=10} cos(y_i - \mu)\right]}{\left(2\pi I_\circ(\kappa)\right)^{10}} \\ &= \frac{exp\left[\kappa\left(-1 + \sum_{i=1}^{i=10} cos(y_i - \mu)\right)\right]}{\left(2\pi I_\circ(\kappa)\right)^{10}} \end{aligned} \tag{9}$$

We simulate 10000 values from prior of κ using the results from (8) and use the same as a grid of values over the support of $\kappa > 0$. The generated values are stored in `prior_kappa`. Next, we calculate the unnormalized posterior density using the results from (9) in `posterior`. In **Figure 7** we show the posterior distribution of κ for the wind direction data over a fine grid of κ values.

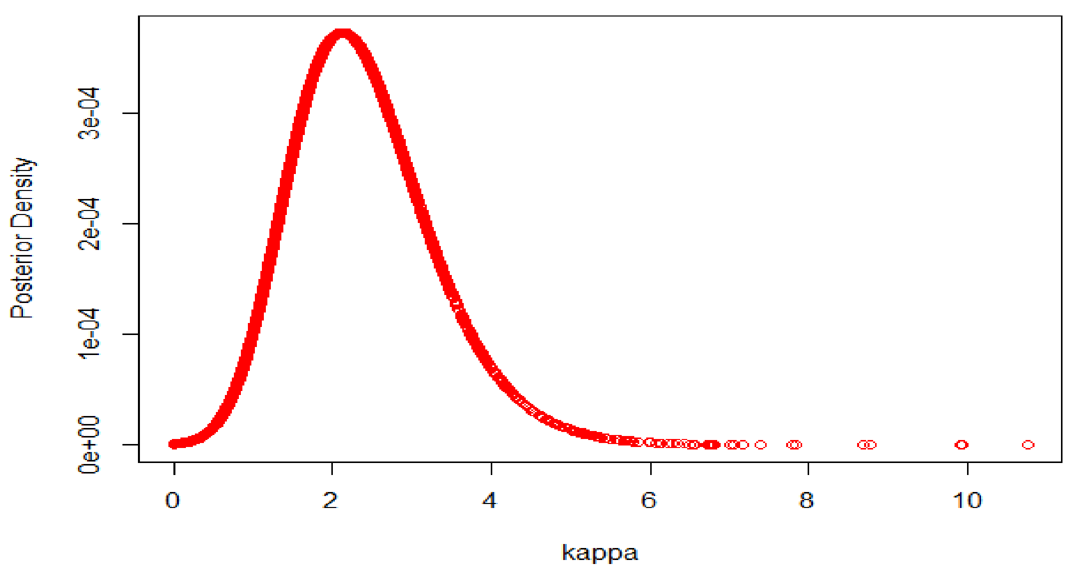

Figure 7: Posterior Distribution of κ over a fine grid of κ values

3.2 (b)

Given the data, the mode is calculated as the single 'most likely' value of the posterior density. It is found to be approximately `2.1246`

4 R code

```
library(ggplot2)

##### 1
## (a)
set.seed(2020-04-10)
s = 8; n=24;f=n-s
alpha = 3; beta=3
### Theoretical mu
e_theta = (alpha+s)/(alpha+beta+n)
### theoretical standard deviation
sqrt((e_theta*(1- e_theta))/(alpha+beta+n+1))
# Posterior Draws
bernoulli.posterior<- function(nDraws, s, f, a, b){
  shape1 = a+s
  shape2 = b+f

  dist = rbeta(n = nDraws, shape1 = shape1,shape2 = shape2)
  posterior_mean = mean(dist)
  posterior_variance = var(dist)
  return(c(posterior_mean,posterior_variance))
}

draws = seq(10,10000, by = 10) # Sequence of total draws

result_matrix = matrix(0,2,length(draws))
# Matrix to store Mean, Variance for draws
result_matrix = sapply(draws, FUN = bernoulli.posterior,
                       s=s,f=f,a=alpha,b=beta)

# Mean Plot
plot(result_matrix[1,], col=2, x = draws,
     xlab="# Draws", ylab = "Posterior Mean")
# Variance Plot
plot(sqrt(result_matrix[2,]), col=3, x = draws,
     xlab="# Draws", ylab = "Posterior St. Dev.")

## (b)
```

```
set.seed(2020-04-10)
simulation<- function(nDraws, s, f, a, b){
  shape1 = a+s
  shape2 = b+f

  dist = rbeta(n = nDraws, shape1 = shape1,
               shape2 = shape2)
  return(dist)
}
nDraws = 10000
result_sim = simulation(nDraws = nDraws,
                        s=s, f=f, a=alpha, b=beta)

length(result_sim[result_sim>0.4])/nDraws # Analytical calculation
# Theoretical Probability
1 - pbeta(0.4, shape1 = alpha+s,
          shape2 = beta+f)

## (c)
set.seed(2020-04-10)
log.odds<-function(nDraws, s, f, a, b){
  shape1 = a+s
  shape2 = b+f

  dist = rbeta(n = nDraws, shape1 = shape1,
               shape2 = shape2)
  odds = dist/(1-dist)
  log_odds = sapply(odds, FUN = log)
  return(log_odds)
}

res_log_odds = log.odds(nDraws = 10000,
                        s = s, f = f, a=alpha, b= beta)

# plot
hist(res_log_odds, breaks=100, ylim=c(0,1),
     freq=F, main="Posterior Distribution of Log Odds",
     xlab="", ylab="Posterior Distribution(Logit)")
lines(density(res_log_odds), col = "blue")

##### 2 ####
data = c(38,20,49,58,31,70,18,56,25,78)
## (a)
log.norn<-function(y,mu,sd){
  v = sd**2
```

```
  return((1/(y*sqrt(2*pi*v)))*exp(-0.5/v*((log(y) - mu)**2)))
}

log_data = log(data)
n = length(data)
mu = 3.8
tao_sq = sum((log_data - mu)**2)/n
# Simulating from Posterior
# Lecture-3, pg. 5
set.seed(2020-04-10)
sigma_sq = (n*tao_sq)/rchisq(10000,n)
hist(sigma_sq, breaks = 1000, xlim = c(0,2))


mean(sigma_sq) # Simulated Mean
(10*tao_sq)/8 ## Theoretical mu (BDA pg. 577)
var(sigma_sq) # Simulated Variance
(2*(10*tao_sq*10*tao_sq))/(64*6) # Theoretical Variance (BDA pg. 577)

## (b)
#sqrt(2)
gini_dist = (pnorm(sqrt(sigma_sq/2),mean = 0, sd = 1)*2)-1
hist(gini_dist, freq = F,breaks=100, xlab="Gini Index",
     ylab= "Posterior Gini Distribution(Data)",
     main = "")
abline(v=0.26,col="red",lwd=3)
lines(density(gini_dist), col="blue", lwd=3)

## (c)
# 90% Credible Interval
ci_90 = quantile(gini_dist, probs = c(0.05,0.95))

densities = density(gini_dist, kernel = "gaussian")

dat <- with(densities, data.frame(x, y))
# Plot
ggplot(data = dat, mapping = aes(x = x, y = y)) +
  geom_line()+
  geom_area(mapping = aes(x = ifelse(x>ci_90['5%'] &
                            x< ci_90['95%']  , x, 0)), fill = "red")+
  ylim(c(0,10)) +
  labs(x = "Gini Coefficient", y = "Posterior Density",
       title = "90% Credible Interval",
       subtitle = "Gini Coefficient")+
  geom_vline(xintercept = ci_90['5%'],
```

```
             color = "black", size=1, linetype="dotted")+
  geom_vline(xintercept = ci_90['95%'],
             color = "black", size=1,linetype="dotted")

## Highest Posterior Density Interval
# Reference
#https://stats.stackexchange.com/questions/381520/
#how-can-i-estimate-the-highest-posterior-density-interval-from-a-set-of-x-y-

hdi = function(x,y,coverage){
  l_x = length(x)
  best = 0
  for(ai in 1:(l_x-1))
  {
    for(bi in (ai +1): l_x){
      mass = sum( diff(x[ai:bi]) * y[(ai+1):bi])
      if (mass >= coverage && mass/(x[bi] - x[ai]) > best)
      {
        best = mass / (x[bi] - x[ai])
        ai.best = ai
        bi.best = bi
      }
    }
  }
  c(x[ai.best], x[bi.best])
}

hdci_90 = hdi(x = densities$x, y = densities$y, coverage = 0.9)
# Plot HPDI
ggplot(data = dat, mapping = aes(x = x, y = y)) +
  geom_line()+
  geom_area(mapping = aes(x = ifelse(x>hdci_90[1]
                          & x< hdci_90[2] , x, 0)), fill = "red")+
  ylim(c(0,10)) +
  labs(x = "Gini Coefficient", y = "Posterior Density",
       title = "90% Highest Posterior Density Interval",
       subtitle = "Gini Coefficient") +
  geom_vline(xintercept = hdci_90[1],
             color = "black", size=1, linetype="dotted") +
  geom_vline(xintercept = hdci_90[2],
             color = "black", size=1, linetype="dotted")

##### 3

## (a)
```

```r
set.seed(2020-04-10)
data_radian = c(-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)

mu = 2.39 # Given
n = length(data_radian)
nDraws = 10000
# Draw Prior Kappa values
prior_kappa = rexp(nDraws, rate = 1)
# y_i - mu
const2 = sum(cos(data_radian-mu))

# Calculate Posterior
posterior = exp(prior_kappa* (const2 - 1))/ (2*pi*besselI(x =prior_kappa, nu
post = posterior/sum(posterior)# Normalize

plot(x = prior_kappa, y = post,
     col="red", xlab = "kappa",
     ylab = "Posterior Density")

## (b)
prior_kappa[which.max(post)]
```

References

[1] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[2] Kodiologist. *Estimating HPDI in R*. https://stats.stackexchange.com/questions/381520/how-can-i-estimate-the-highest-posterior-density-interval-from-a-set-of-x-y-valu. Accessed: 2021-04-10. 20.