732A91: Lab 1
Bayesian Learning

Sarah Alsaadi, Carles Sans Fuentes

April 13, 2017

Question 1

Let $y_1, \ldots, y_n | \theta \sim \text{Bern}(\theta)$, and assume that you have obtained a sample with $s = 14$ successes in $n = 20$ trials. Assume a $\text{Beta}(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 2$.

- Draw random numbers from posterior $\theta | y_1, \ldots, y_n \sim \text{Bern}(\alpha_0 + s, \beta_0 + f)$ and verify graphically that the posterior mean and standard deviation converges to the true mean $E[\theta] = \frac{\alpha_0+s}{\alpha_0+s+\beta_0+f} \approx 0.66$ and true standard deviation $Var(\theta) = \frac{(\alpha_0+s)*(\beta_0+f)}{(\alpha_0+s+\beta_0+f)^2*(\alpha_0+s+\beta_0+f+1)} \approx 0.09$. Figure 1 and 2 shows how the mean and standard deviation of samples randomly drawn converges to the true mean and true standard deviation with larger samples.

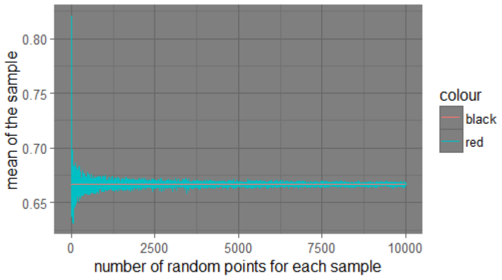

Figure 1: A plot representing the mean of 10000 samples where the sample size goes from 1 to 10000. The theoretical mean is plotted as a line.

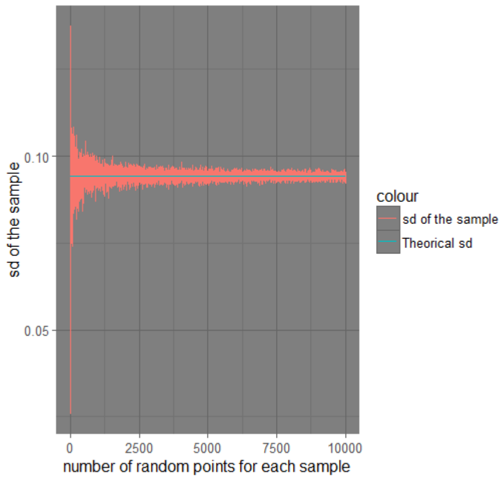

Figure 2: A plot representing the standard deviation of 10000 where the sample size goes from 1 to 10000. The theoretical standard deviation (sd) is plotted as a line.

- Use simulation (nDraws= 10000) to compute the posterior probability $Pr(\theta < 0.4|y)$ and compare with the exact value $Pr(\theta < 0.4|y) = 0.00397$. The simulated value is given below, we get 0.0036 which is close enough to the exact value.

```
pbeta(0.4,16,8)
[1] 0.003972563
simulated pbeta
[1] 0.0036
```

- Compute the posterior distribution of the log-odds $\phi = log(\frac{\theta}{1-\theta})$ by simulation with nDraws= 10000. Figure 3 below shows the histogram together with the kernel density of the data simulated from the posterior distribution of the log-odds $\phi = log(\frac{\theta}{1-\theta})$ with nDraws= 10000.

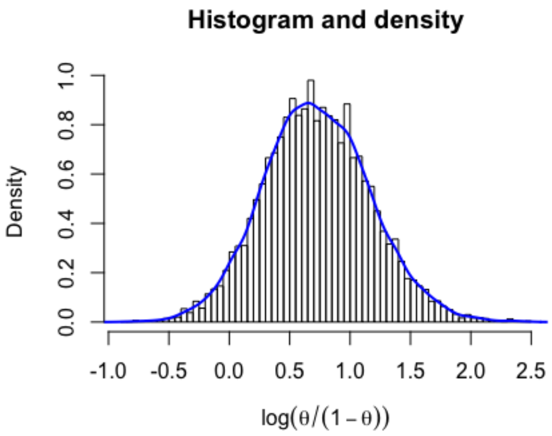

Figure 3: Histogram and kernel density of of the data simulated from the posterior distribution of the log-odds $\phi = log(\frac{\theta}{1-\theta})$ with nDraws= 10000.

Question 2

Assume that you have asked 10 randomly selected persons about their monthly income (in thousands Swedish Krona) and obtained the following 10 observations $(y_1, \ldots, y_{10}) = (14, 25, 45, 25, 30, 33, 19, 50, 34, 67)$. Assume $y_1, \ldots, y_n|\mu, \sigma^2 \sim logNormal(\mu, \sigma^2)$,

$\mu = 3.5$ and a non-informative prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$. It can be shown that the posterior for σ^2 is $Inv - \chi^2(n, \tau^2)$ (scaled), where

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

.

- Simulate 10000 draws from the posterior of σ^2 and compare with the theoretical $Inv - \chi^2(n, \tau^2) = Inv - \chi^2(10, 0.198)$. Figure 4 shows the histogram of the data simulated from $Inv-\chi^2(10, 0.198)$ and the density of the theoretical posterior distribution $Inv-\chi^2(10, 0.198)$. Since the simulated sample is so big, the histogram and the density is quite alike.

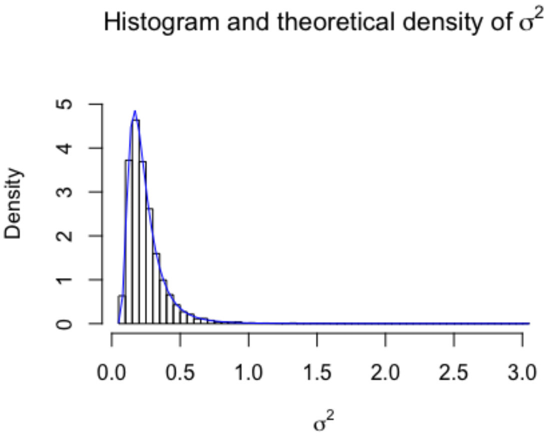

Figure 4: Histogram of simulated data and theoretical density of the σ^2 with draws= 10000.

- The most common measure of income inequality is the Gini coefficient, G, where $0 \leq G \leq 1$. G= 0 means a completely equal income distribution, whereas G= 1 means complete income inequality. It can be shown that $G = 2\Phi(\frac{\sigma}{\sqrt{2}}) - 1$ when incomes follow a $logNormal(\mu, \sigma^2)$ distribution. Use the draws in a) to compute the posterior of the Gini coefficient for the current data set. Figure 5 shows the histogram of G and the kernel density. It shows that G is far from 1 so the income distribution seems to be closer to equal than inequal.

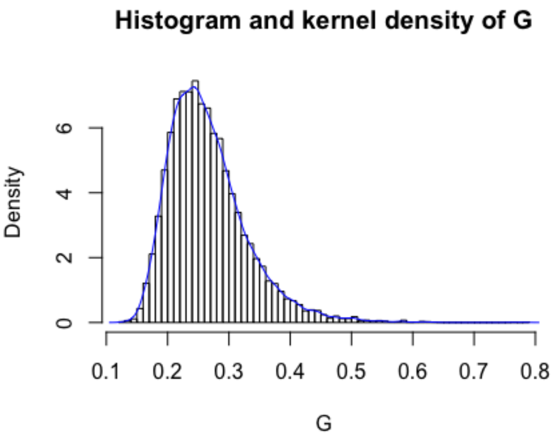

Figure 5: Histogram of simulated data and kernel density of G with draws= 10000.

- Use the posterior draws from b) to compute a 95% equal tail credible set for G. Also, do a kernel density estimate of the posterior of G and use it to compute a 95% Highest Posterior Density set for G. The Highest Posterior Density set (HPD) is a credible set that consists of

θ-values having the Highest Posterior Density. Below we have a 95% equal tail credible set for G and and HPD. For a symmetric distribution, the intervals would be the same but since the distribution of G is slightly skewd, the intervals differ.

```
quantile(G,  probs = c(0.025, 0.975))
#     2.5%      97.5%
# 0.1739323 0.4152031
hdi(density(G),credMass=0.95)
#lower      upper
#0.1601422 0.3918670
```

Question 3

The following data is the observed wind directions at a given location on 10 different days. The data recorded in radians:

$$(y_1, \ldots, y_{10}) = (-2.44, 2.14, 2.54, 1.83, 2.02, 2.33, -2.79, 2.23, 2.07, 2.02)$$

where $-\pi \leq y \leq \pi$.
Assume that the observations are independent and follow the von Mises distribution

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

$-\pi \leq y \leq \pi$.
Furthermore, assume that $\mu = 2.39$ and let $\kappa \sim \text{Exp}(\lambda = 1)$ apriori.

- Plot the posterior distribution of κ for the data given above. The posterior is given by the following:

$$p(\kappa|\mu, y) = exp(-\kappa) \prod_{i=1}^{10} \frac{exp(\kappa * cos(y_i - 2.39))}{2\pi I_0(\kappa)}$$

  Figure 6 shows the posterior of κ.

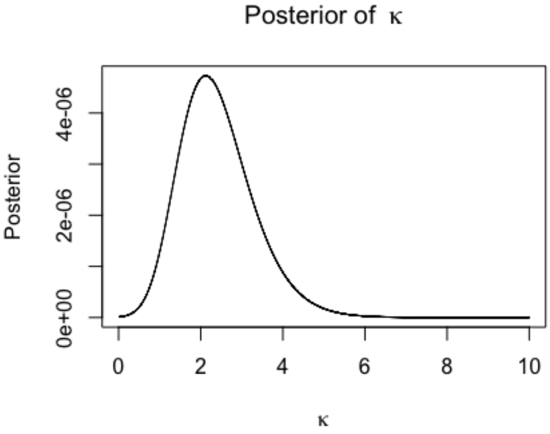

Figure 6: Posterior of κ .

- The approximate point where the posterior has its mode is $(2.125, 4.7 * 10^{-6})$.

Contributions

All results and comments presented have been developed and discussed together by the members of the group.

Appendix

Question 1

```

install.packages("geoR")
install.packages("HDInterval")
library("geoR")
library("HDInterval")
#lab 1 Bayesian learning

#1a
x<-rbeta(10000,2+14,2+6)
h<-hist(x,breaks = seq(0,1,0.02))
xfit<-seq(min(x),max(x),length=100)
yfit<-dbeta(xfit,16,8)
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)

#1b
pbeta(0.4,16,8)
y<-x[x<0.4]
prob<-length(y)/length(x)


#1c
xodds<-log(x/(1-x))
hist(xodds,breaks = 50,prob=TRUE, main = ’Histogram and density’, xlab = expression(paste(log(theta/(1-theta)))))
lines(density(xodds), col="blue", lwd=2)

#2a

data<-c(14,25,45,25,30,33,19,50,34,67)

tao2<-function(data)
{
  sum((log(data)-3.5)^2)/length(data)

}

tao<-tao2(data=data)

sigma2<-rinvchisq(10000,df=10,scale=tao)
hist(sigma2,breaks = 100, prob=TRUE)
x<-seq(from=0, to=10000, by=0.001)
hx<-dinvchisq(x,df=10,scale=tao)

sigma2.histogram = hist(sigma2, breaks = 100, freq = F)
sigma2.ylim.normal = range(0, sigma2.histogram$density, dinvchisq(sigma2,df=10,scale=tao), na.rm = T)
hist(sigma2, breaks = 100, freq = F, ylim = c(0, 5.5), main=expression("Histogram and theoretical density of" ~ sigma^2), xlab = expression(paste(sigma^2)), ylab = ’Density’)
curve(dinvchisq(x,df=10,scale=tao), add = T,col="blue")

#2.b
sigma<-sqrt(sigma2)/sqrt(2)
G<-2*pnorm(sigma,0,1)-1
hist(G,freq=F,breaks=50, main=’Histogram and kernel density of G’)
lines(density(G),col="blue")

#2.c
quantile(G,  probs = c(0.025, 0.975))
#    2.5%      97.5%
# 0.1739323 0.4152031

hdi(density(G),credMass=0.95)

#lower     upper
#0.1601422 0.3918670

#3.a

windradians<-c(-2.44,2.14,2.54,1.83,2.02,2.33,-2.79,2.23,2.07,2.02)
posterior<-function(k)
{
  return(exp(-k)*prod(exp(k*cos(windradians-2.39))/(2*pi*besselI(x=k,nu=0))))

}


values<-sapply(k,posterior)
plot(k,values,type="l",main="Posterior of " ~ kappa, xlab=expression(paste(kappa)), ylab="Posterior")

```

```
#3.b
max(values)
#4.727694e-06
k[which.max(value)]
#2.125
```