732A91 - Lab 1

Joris van Doorn || Weng Hang Wong

09 April 2020

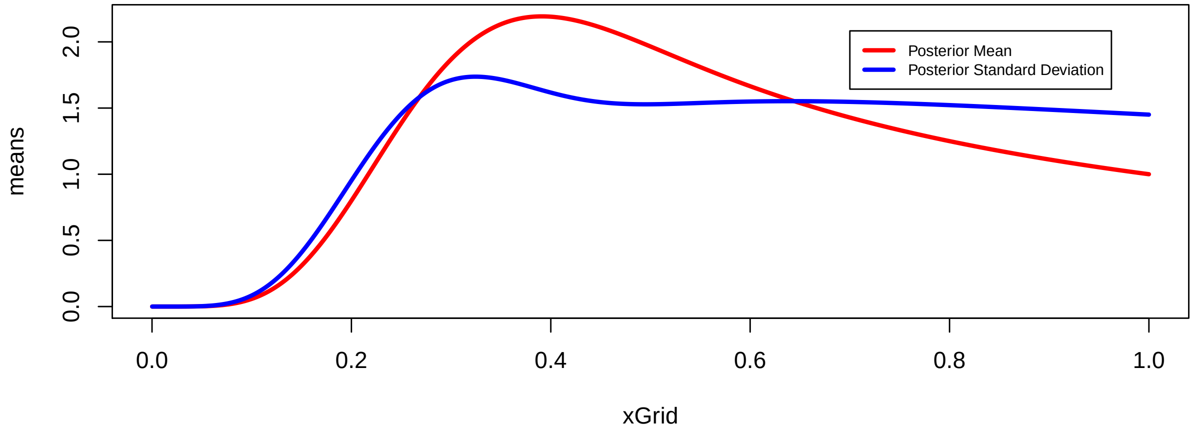
1. Bernoulli . . . again.

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

```
## Number of Draws:  10000
## Posterior Mean:  1
## Posterior Standard Deviation:  1.450244
```



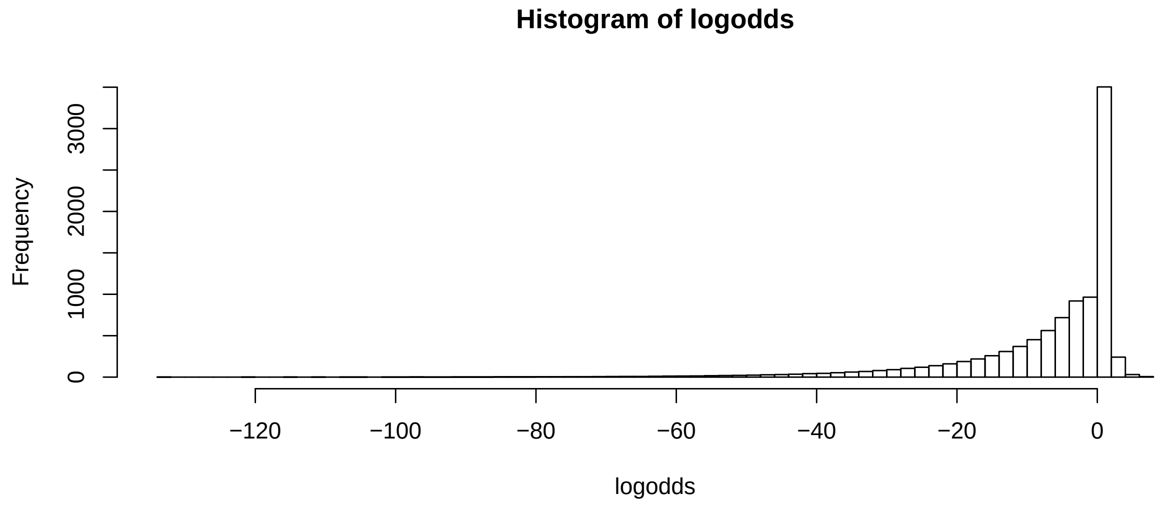
b.

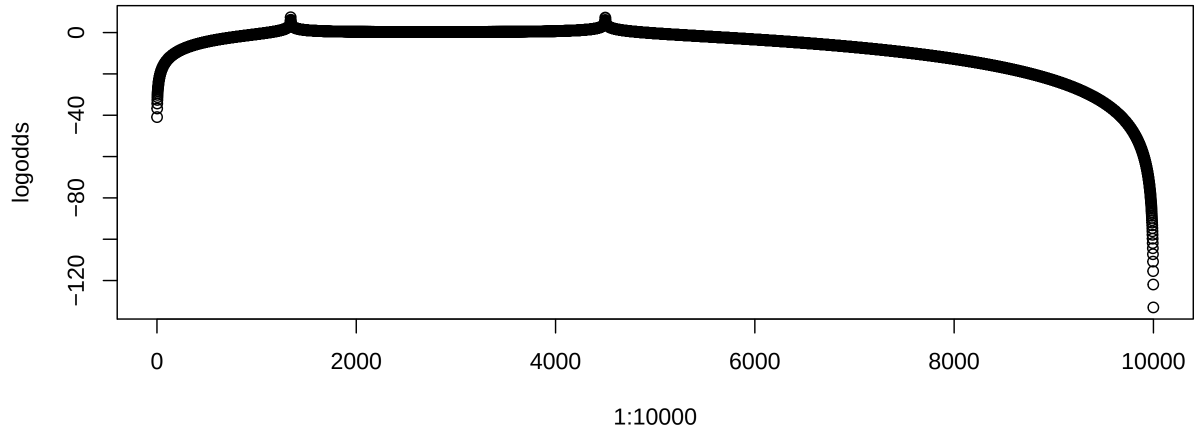
Use simulation (nDraws = 10000) to compute the posterior probability $Pr(\theta > 0.3|y)$ and compare with the exact value

```
## The probability of theta being larger than 0.3 is:  0.7613
```

c.

Compute the posterior distribution of the log-odds ϕ





```
##
## Call:
##  density.default(x = logodds)
##
## Data: logodds (10000 obs.);  Bandwidth 'bw' = 1.154
##
##        x                  y
##  Min.   :-136.47   Min.   :0.000e+00
##  1st Qu.: -99.63   1st Qu.:4.515e-05
##  Median : -62.80   Median :3.945e-04
##  Mean   : -62.80   Mean   :6.783e-03
##  3rd Qu.: -25.96   3rd Qu.:3.620e-03
##  Max.   :  10.88   Max.   :1.309e-01
```

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:*

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

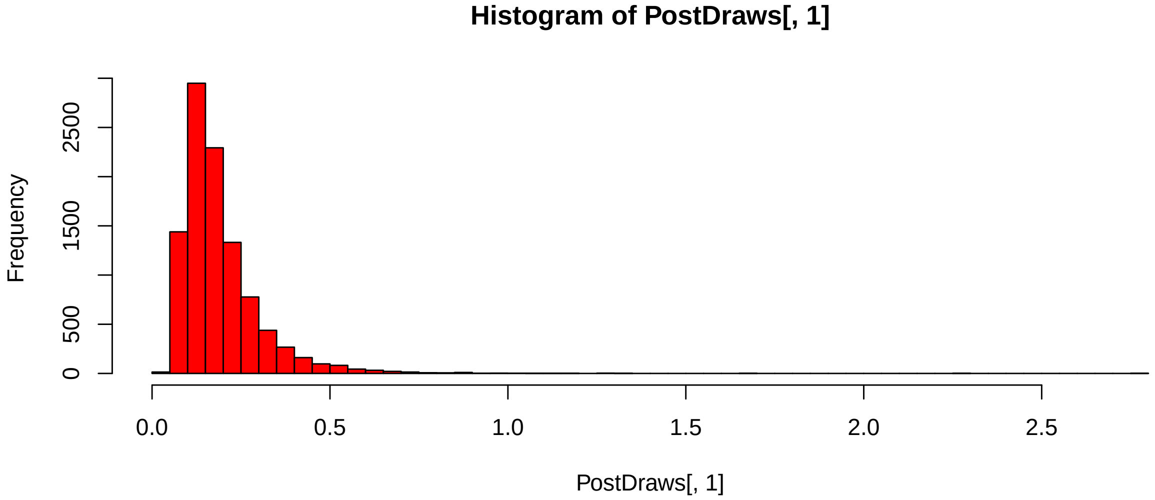
For y>0, $\mu > 0 and \sigma^2 > 0$. *The log-normal distribution is related to the normal distribution as follows: if y* $logN(\mu, \sigma^2)$ *then* $log(y)$ $N(\mu, \sigma^2)$. *Let* $y_1, ..., y_n|\mu, \sigma^2 \ {}^{iid} log(N(\mu, \sigma^2))$, *where* $\mu = 3.7$ *is assumed to be known but* σ^2 *is unknown with non-informative prior* $p(\sigma^2) \propto \frac{1}{\sigma^2}$. *The posterior for the* σ^2 *is the* $Inv - \chi^2(n, \tau^2)$ *distribution, where*

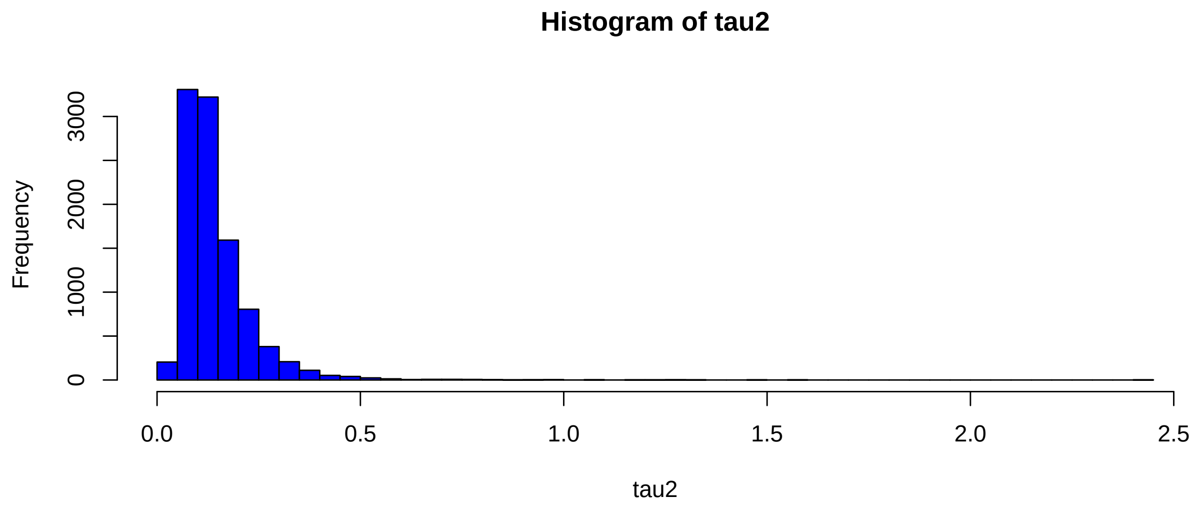
$$\tau^2 = \frac{\sum_{i=1}^{n}(log(y_i) - \mu)^2}{n}$$

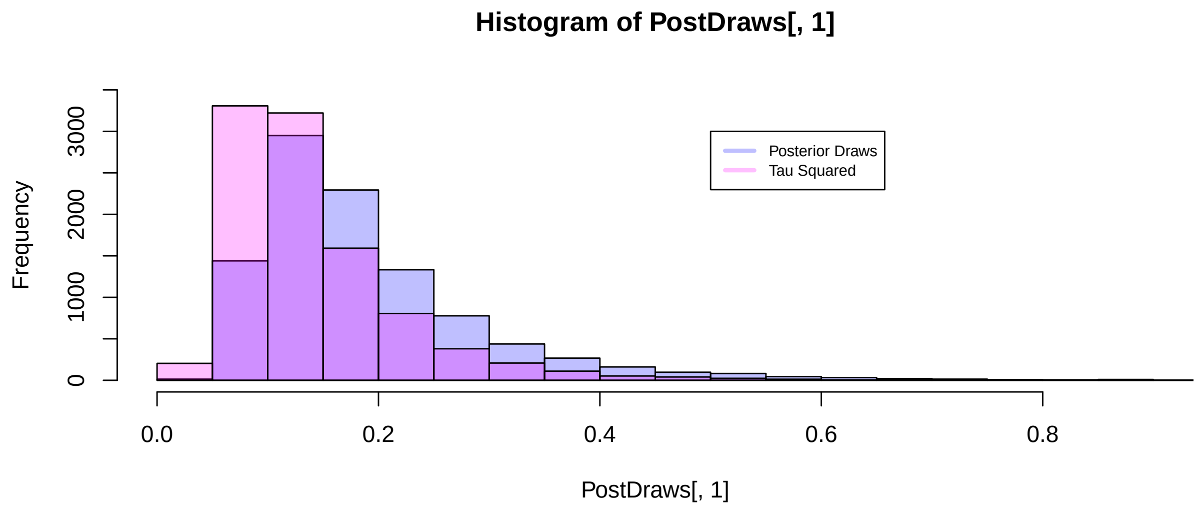
a.

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.*

We worked with the log of the data, assuming that is approximately normally distributed with $N(\mu, \sigma^2)$







b.

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. See Wikipedia for more information. It can be shown that* $G = 2\phi(\frac{\sigma}{\sqrt{2}} - 1)$ *when incomes follow a* $logN(\mu, \sigma^2)$ *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.*

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width=9, fig.height = 4.1)
library(tidyverse)
library(dplyr)
library(knitr)
library(LaplacesDemon)
RNGversion("3.6.2")
set.seed(12345)
# --------------------------------------------------------
# 1a

a = b = 2
n = 20
s = 5
nDraws = 10000
#xGrid <- seq(0.001, 0.999, by=0.001)
#posterior = dbeta(xGrid, a+s, b+(n-s))

means <- c()
sds <- c()

set.seed(12345)

for(i in 1:nDraws){
  xGrid <- seq(1/nDraws, i/nDraws, by=1/nDraws)
  posterior = dbeta(xGrid, a+s, b+(n-s))
  means[i] <- mean(posterior)
  sds[i] <- sd(posterior)

  #at("\nNumber of Draws: ", i , "\nMean: ", mean(posterior), "\nStandard Deviation: ", sd(posterior))
}

cat("Number of Draws: ", nDraws , "\nPosterior Mean: ", means[nDraws], "\nPosterior Standard Deviation:

plot(xGrid, means, type = 'l', lwd = 3, col = "red")
lines(xGrid, sds, lwd = 3, col = "blue")
legend(x = max(xGrid)*0.70, y = 0.95*max(means), legend = c("Posterior Mean", "Posterior Standard Devia
# --------------------------------------------------------
# 1b

xGrid <- seq(1/nDraws, nDraws/nDraws, by=1/nDraws)
posterior = pbeta(xGrid, a+s, b+(n-s)) # Ask for the difference between pbeta & dbeta

prob_03 <- posterior[posterior > 0.3]
prob <- length(prob_03)/nDraws

cat("The probability of theta being larger than 0.3 is: ", prob)
# --------------------------------------------------------
# 1c

xGrid <- seq(1/nDraws, nDraws/nDraws, by=1/nDraws)
```

```
posterior = dbeta(xGrid, a+s, b+(n-s))

logodds <- c()

for(i in 1:length(posterior)){
  logodds[i] <- log(abs(posterior[i]/(1-posterior[i])))
}

hist(logodds, breaks = 100)
plot(1:10000,logodds)
density(logodds)
# ---------------------------------------------------------
# 2a
y <- c(44,25,45,52,30,63,19,50,34,67)
n <- length(y)
logy <- log(y)
nDraws <- 10000
mu <- 3.7

# Code based on Mattias Villani's NormalNonInfoPrior.R, found on https://github.com/STIMALiU/BayesLearn
LogNormalNonInfoPrior <- function(nDraws, data, mu){
  datamean <- mean(data)
  n <- length(data)
  tau2 <- sum((data-mu)^2) / n
  PostDraws <- matrix(0,nDraws,2)
  PostDraws[,1] <- ((n-1)*tau2)/rchisq(nDraws,n-1)
  PostDraws[,2] <- datamean+rnorm(nDraws,0,1)*sqrt(PostDraws[,1]/n)
  return(PostDraws)
}

#tau2 <- sum((logy-mu)^2) / n
tau2<-rinvchisq(nDraws, n-1, scale=1/(n-1))

PostDraws<-LogNormalNonInfoPrior(10000,logy, mu)

p1<-hist(PostDraws[,1], breaks = 50, col = "red")            # Plotting the histogram of mu-draws
p2<-hist(tau2, breaks = 50, col = "blue")
plot(p1, col=rgb(0,0,1,1/4), ylim = c(0,3500), xlim = c(0,0.90))
plot(p2, col=rgb(1,0,1,1/4), add=T)
legend(x = 0.5, y = 3000, legend = c("Posterior Draws", "Tau Squared"), col = c(rgb(0,0,1,1/4), col=rgb
# ---------------------------------------------------------
# 2b
```