M200A Evolutionary Biology, ABC lab

Hayley Stansell

November 29, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

NUMBER OF #'S MATTER ###big #####small

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot. This also can keep data outputs from showing, or you can just comment it out

### Exercise 1

##### a) Assume that N can be any value b/w 100 and 100,000. Draw 1 million values from the prior dist. of N. Use a uniform distribution for N. Assume that N~U[100, 100000]

getwd()

## [1] "C:/Users/Hayley/eeb201hayleystansell"

setwd("C:\\Users\\Hayley\\eeb201hayleystansell")  
getwd()

## [1] "C:/Users/Hayley/eeb201hayleystansell"

#?runif  
#try simulating the coin flip thing from class  
#prior<- runif(1e6)  
#D\_sim<- rbinom(n=1e6, size=100, p=prior)  
#plot(D\_sim)  
#This produces a "black block", x-axis going from 0 to 1e6, y-axis going from 0 to 100....  
#?plot  
#how is a number the prior distribution????   
prior<- runif(n=1e6, min = 100, max = 100000)  
#prior  
#plot(prior) #looks correct in terms of axes  
#

##### b) For each value of N, simulate a TMRCA. Use same approach as on last week's assignment. Note that you are using N (vs. 2N) because you are dealing with haploid samples (2 haploid samples from 2 individuals, one c-some per individ.)

rate<- 1/(prior)  
mostrecent<- rexp(1e6, rate)  
#######DOES NOT NEED TO BE A LOOP!, R automatically takes each value of "prior" and plugs it in in order to the "mostrecent" vector  
#######Follows that NEITHER DOES NEXT STEP  
#mostrecent  
length(mostrecent)

## [1] 1000000

##### c) For each TMRCA, add a Poisson number of mutations w/ the appropriate mutation rate. rate = mu = 1e-8 per base pair per c-some....for 100kb is 1e-3.... \*2? for 2 c-somes??

ratemu<- 1e-3  
snpnumber<- rpois(1e6, mostrecent\*ratemu\*2)  
#snpnumber  
max(snpnumber)

## [1] 2350

length(snpnumber)

## [1] 1000000

##### d) Set parameters to "choose" which draws from the prior give data that approx. the OBSERVED number of SNPs accept all values of "N" that give b/w 45 and 55 SNPs

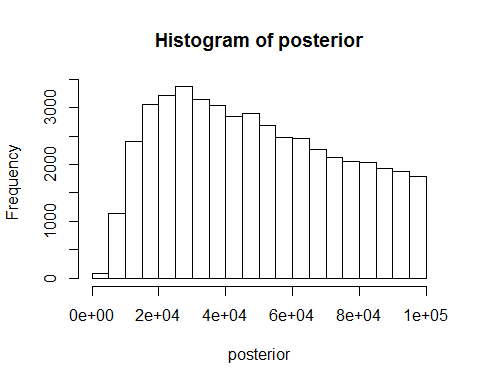
sims<- cbind(prior, snpnumber)  
simslower<- subset(sims, sims[,2]>45)  
simsupper<- subset(simslower, simslower[,2]<55)  
posterior<-simsupper[,1]  
#posterior  
max(posterior)

## [1] 99999.51

min(posterior)

## [1] 2179.713

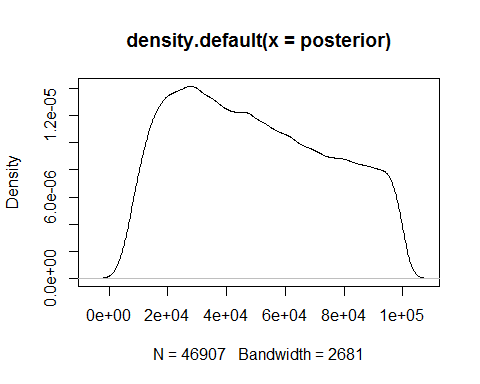
hist(posterior)



density(posterior)

##   
## Call:  
## density.default(x = posterior)  
##   
## Data: posterior (46907 obs.); Bandwidth 'bw' = 2681  
##   
## x y   
## Min. : -5863 Min. :3.560e-10   
## 1st Qu.: 22613 1st Qu.:7.523e-06   
## Median : 51090 Median :9.535e-06   
## Mean : 51090 Mean :8.771e-06   
## 3rd Qu.: 79566 3rd Qu.:1.226e-05   
## Max. :108042 Max. :1.415e-05

plot(density(posterior))



##### e) I should be done!

### EXERCISE TWO

##### Make a density plot of your prior and posterior distributions of N. plot them on the same axes, label which line corresponds to the prior, which corresponds to the posterior

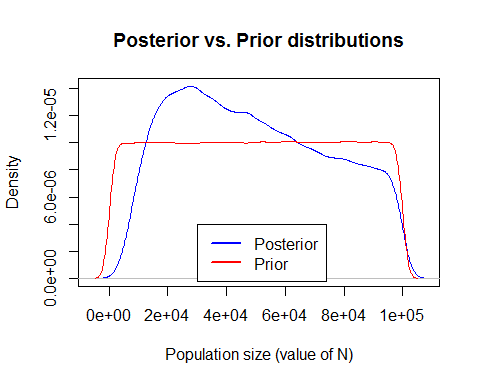
density(posterior)

##   
## Call:  
## density.default(x = posterior)  
##   
## Data: posterior (46907 obs.); Bandwidth 'bw' = 2681  
##   
## x y   
## Min. : -5863 Min. :3.560e-10   
## 1st Qu.: 22613 1st Qu.:7.523e-06   
## Median : 51090 Median :9.535e-06   
## Mean : 51090 Mean :8.771e-06   
## 3rd Qu.: 79566 3rd Qu.:1.226e-05   
## Max. :108042 Max. :1.415e-05

plot(density(posterior), main= "Posterior vs. Prior distributions", xlab= "Population size (value of N)", col= "blue")  
density(prior)

##   
## Call:  
## density.default(x = prior)  
##   
## Data: prior (1000000 obs.); Bandwidth 'bw' = 1637  
##   
## x y   
## Min. : -4811 Min. :1.393e-08   
## 1st Qu.: 22619 1st Qu.:9.975e-06   
## Median : 50050 Median :1.001e-05   
## Mean : 50050 Mean :9.105e-06   
## 3rd Qu.: 77481 3rd Qu.:1.004e-05   
## Max. :104911 Max. :1.011e-05

#?lines  
# "lines() adds a line to an existing plot --- lines(namevector, col = color)  
lines(density(prior), col= "red")  
#?legend  
legend(3e04,4e-06, # places a legend at the appropriate place   
 c("Posterior", "Prior"), # puts text in the legend  
   
 lty=c(1,1), # gives the legend appropriate symbols (lines)  
   
 lwd=c(2.5,2.5),col=c("blue", "red")) # gives the legend lines the correct color and width



### EXERCISE THREE

##### What is the median value of the posterior distribution of N?

median(posterior)

## [1] 46956.71

### EXERCISE FOUR

##### Generate a 95% credible interval for the posterior dist. of N (like a confidence interval). Note, in this framework, there actually is a 95% chance of the true parameter value falling in this region. Hint - use "quantile" function.

#?quantile  
quantile(posterior, c(0.025, 0.975))

## 2.5% 97.5%   
## 9844.008 96665.783

The actual parameter value should, with a 95% chance, fall somewhere between these two values. This range represents 95% of the data output distribution, centered around the mean...so they are the most likely values.

### EXERCISE FIVE

##### How does the posterior dist. differ from the prior distribution? A descriptive answer will suffice.

The posterior distribution is essentially more informative than the prior distribution, as a product of the restrictions placed on its "acceptable" values - which are defined in the ABC simulation process. The range of likely (or even possible) values of N which satisfy the model in the posterior 1) are greatly reduced compared with the prior, and 2) have a more defined, uneven distribution across that range (approximating a normal distribution, more), which indicates that a certain range of N's is more likely to satifsy the model (compared with the prior, which has an even distribution across its range).

### EXERCISE SIX

##### Repeat your ABC analysis, but change the prior dist. of N to be U~[1000,1e6]. What is the mean, median, and 95% credible interval for the posterior? How does this differ from what you computed in questions 3-4 for the original prior? What does this tell you about the effect fo the prior distribution in Bayesian stats?

prior6<- runif(n=1e6, min = 1000, max = 1000000)  
rate6<- 1/(prior6)  
mostrecent6<- rexp(1e6, rate6)  
length(mostrecent6)

## [1] 1000000

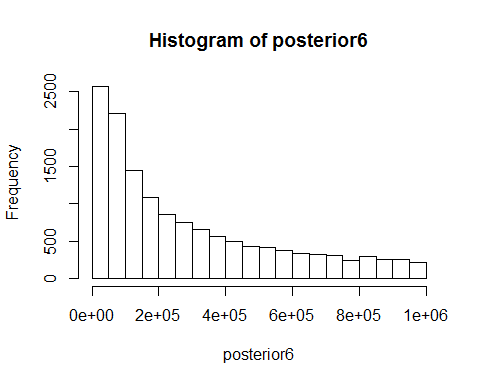
ratemu6<- 1e-3  
snpnumber6<- rpois(1e6, mostrecent6\*ratemu6\*2)  
#snpnumber  
max(snpnumber6)

## [1] 22740

length(snpnumber6)

## [1] 1000000

sims6<- cbind(prior6, snpnumber6)  
simslower6<- subset(sims6, sims6[,2]>45)  
simsupper6<- subset(simslower6, simslower6[,2]<55)  
posterior6<-simsupper6[,1]  
#posterior  
hist(posterior6)



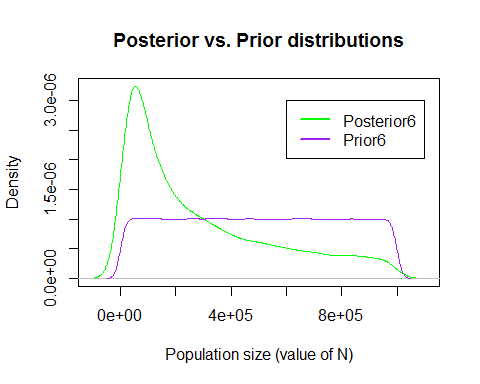
#  
?plot

## starting httpd help server ... done

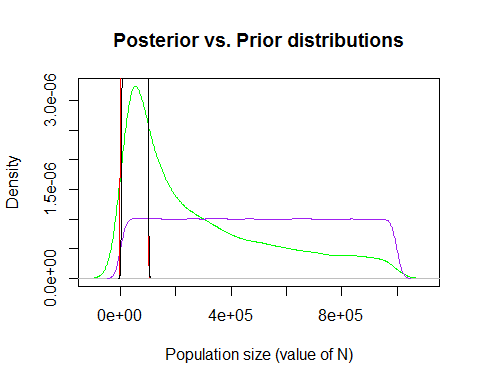
plot(density(posterior6), main= "Posterior vs. Prior distributions", xlab= "Population size (value of N)", col= "green")  
density(prior6)

##   
## Call:  
## density.default(x = prior6)  
##   
## Data: prior6 (1000000 obs.); Bandwidth 'bw' = 1.638e+04  
##   
## x y   
## Min. : -48137 Min. :1.397e-09   
## 1st Qu.: 226181 1st Qu.:9.975e-07   
## Median : 500500 Median :1.001e-06   
## Mean : 500500 Mean :9.105e-07   
## 3rd Qu.: 774818 3rd Qu.:1.004e-06   
## Max. :1049137 Max. :1.011e-06

#?lines  
# "lines() adds a line to an existing plot --- lines(namevector, col = color)  
lines(density(prior6), col= "purple")  
#?legend  
legend(6e05,3e-06, # places a legend at the appropriate place   
 c("Posterior6", "Prior6"), # puts text in the legend  
   
 lty=c(1,1), # gives the legend appropriate symbols (lines)  
   
 lwd=c(2.5,2.5),col=c("green", "purple")) # gives the legend lines the correct color and width



plot(density(posterior6), main= "Posterior vs. Prior distributions", xlab= "Population size (value of N)", col= "green")  
lines(density(prior6), col="purple")  
lines(density(prior), col ="red")  
lines(density(posterior), col ="black")



#graph done to compare both simulations together.....not the most useful thing

For this simulation, the mean value of N is 2.851086310^{5}, the median is 1.840086310^{5}, and the 95% credible interval is the range, 1.500637310^{4}, 9.224596110^{5} , meaning that there is a 95% percent chance that the actual value for N falls within this range. For the previous simulation, the mean was 4.949847410^{4} , the median was 4.695670610^{4}, and the 95% credible interval was the range, 9844.0079419, 9.666578310^{4}

Increasing the prior distribution by an order of magnitude resulted in a posterior distribution that appears more refined and which hones in around a range of values that are perhaps more likely to be the "true N value". The range for the 95% credible interval is smaller for the second simulation. This implies that the posterior stemming from a larger range of N's in the prior improves the degree of likelihood or credibility for the posterior.

### EXERCISE SEVEN

##### If you wanted a more precise estimate of the pop. size (assume there is no sex-biased demographic history so you can easily extrapolate the total pop. size from the Y c-some pop. size and vice versa), would you be better off:

##### a) sequencing a bigger region of the Y chromosome?

##### or

##### b) sequencing the same amount of the *autosomes*?

##### why?