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] 2319

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

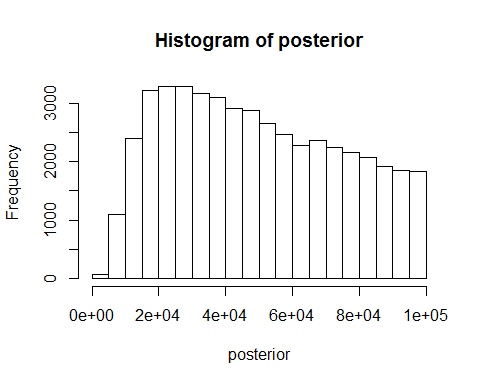
sims<- cbind(prior, snpnumber)  
simslower<- subset(sims, sims[,2]>45)  
simsupper<- subset(simslower, simslower[,2]<55)  
posterior<-simsupper[,1]  
#posterior  
max(posterior)

## [1] 99997.94

min(posterior)

## [1] 2651.561

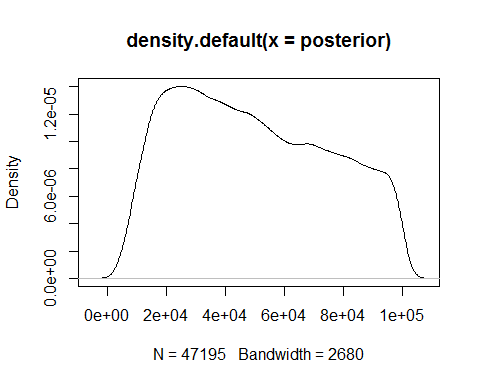
hist(posterior)



density(posterior)

##   
## Call:  
## density.default(x = posterior)  
##   
## Data: posterior (47195 obs.); Bandwidth 'bw' = 2680  
##   
## x y   
## Min. : -5390 Min. :5.660e-10   
## 1st Qu.: 22968 1st Qu.:7.505e-06   
## Median : 51325 Median :9.776e-06   
## Mean : 51325 Mean :8.807e-06   
## 3rd Qu.: 79682 3rd Qu.:1.231e-05   
## Max. :108039 Max. :1.404e-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

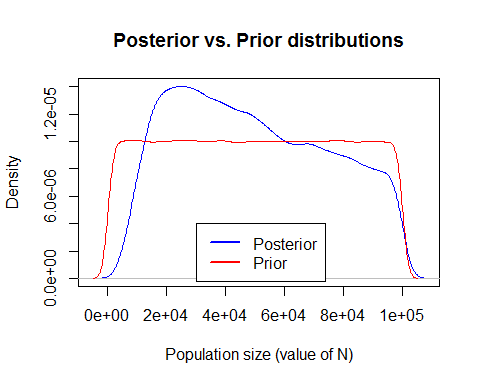
density(posterior)

##   
## Call:  
## density.default(x = posterior)  
##   
## Data: posterior (47195 obs.); Bandwidth 'bw' = 2680  
##   
## x y   
## Min. : -5390 Min. :5.660e-10   
## 1st Qu.: 22968 1st Qu.:7.505e-06   
## Median : 51325 Median :9.776e-06   
## Mean : 51325 Mean :8.807e-06   
## 3rd Qu.: 79682 3rd Qu.:1.231e-05   
## Max. :108039 Max. :1.404e-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' = 1638  
##   
## x y   
## Min. : -4815 Min. :1.417e-08   
## 1st Qu.: 22618 1st Qu.:9.969e-06   
## Median : 50050 Median :1.001e-05   
## Mean : 50050 Mean :9.104e-06   
## 3rd Qu.: 77482 3rd Qu.:1.004e-05   
## Max. :104914 Max. :1.012e-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 a specified place   
 c("Posterior", "Prior"), # defines the text in the legend  
   
 lty=c(1,1), # legend symbols (lines)  
   
 lwd=c(2.5,2.5),col=c("blue", "red")) # gives lines the correct color and weight



### EXERCISE THREE

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

median(posterior)

## [1] 46881.67

### 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%   
## 10035.04 96774.70

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

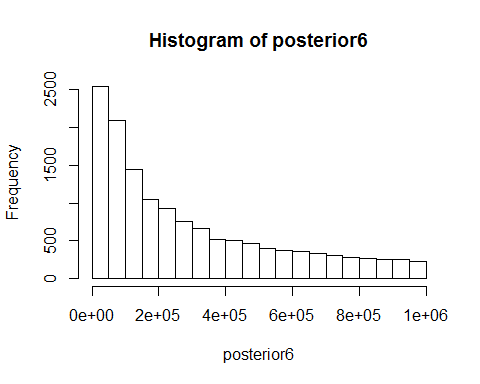
ratemu6<- 1e-3  
snpnumber6<- rpois(1e6, mostrecent6\*ratemu6\*2)  
#snpnumber  
max(snpnumber6)

## [1] 21313

length(snpnumber6)

## [1] 1000000

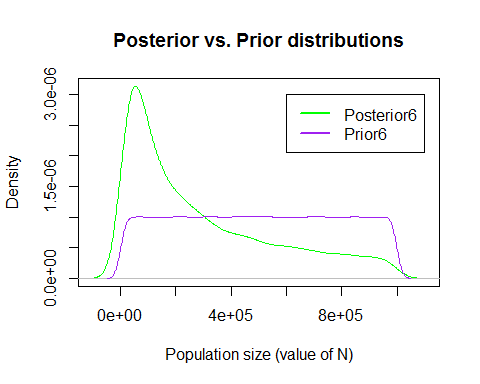
sims6<- cbind(prior6, snpnumber6)  
simslower6<- subset(sims6, sims6[,2]>45)  
simsupper6<- subset(simslower6, simslower6[,2]<55)  
posterior6<-simsupper6[,1]  
#posterior  
hist(posterior6)



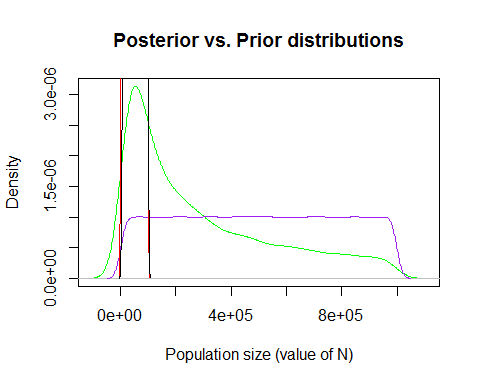
#  
#?plot  
plot(density(posterior6), main= "Posterior vs. Prior distributions", xlab= "Population size (value of N)", col= "green")  
density(prior6)

##   
## Call:  
## density.default(x = prior6)  
##   
## Data: prior6 (1000000 obs.); Bandwidth 'bw' = 1.637e+04  
##   
## x y   
## Min. : -48104 Min. :1.395e-09   
## 1st Qu.: 226198 1st Qu.:9.974e-07   
## Median : 500500 Median :1.001e-06   
## Mean : 500500 Mean :9.105e-07   
## 3rd Qu.: 774802 3rd Qu.:1.004e-06   
## Max. :1049103 Max. :1.011e-06

lines(density(prior6), col= "purple")  
  
legend(6e05,3e-06,   
 c("Posterior6", "Prior6"), # puts text in the legend  
 #try putting "top right" and it will set to right place  
   
 lty=c(1,1),   
   
 lwd=c(2.5,2.5),col=c("green", "purple"))



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 2.882574410^{5}, the median is 1.932101310^{5}, and the 95% credible interval is the range, 1.489750910^{4}, 9.243986810^{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 4.953084110^{4} , the median was 4.688167210^{4}, and the 95% credible interval was the range, 1.003504110^{4}, 9.677470110^{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?

Generally, it seems that the larger the prior distribution, the smaller (or more dense) the posterior distribution. In order to increase the prior for this simulation, you could choose to sequence more of the Y-chromosome, effectively increasing the range of "N" used in the prior. This should result in a more precise estimate as the range for the 95% credible interval in the posterior should decrease.

## Estimation of Population Split Times--------------------------------------------------------------------

### EXERCISE ONE

##### What is the expected additional time we have to wait until the two chromosomes coalesce *once* they make it back into the ancestral population?

Since this is a haploid group, the expected time to coalescence in the ancestral population is just N (vs. 2N) generations, where N is the population size (or number of y-chromosomes). If the ancestral population is of size N= 100,000, then the expected time to coalesce is 100,000 generations.

### EXERCISE TWO

##### What is the expected time until the two chromosomes coalesce with each other? Write this formula using tsplit as split time. (write this as a formula, not a number estimate)

TimeToCoalescence = Ngenerations + tsplit (basically, the time the populations were divergent, plus the time it took for the chromosomes to then coalesce in the ancestral population)

### EXERCISE THREE

-know that that ancestral pop size is N= 100000 (1e5) (haploid) -sequenced a 100kb region (mu for region is 1e-3) -do not know tsplit (time POPULATION split....not the same as TMRCA) -range of tsplit can be 50000 to 1e6 (1 million gen) -do not know divided population size (need estimates) -this influences the Ngenerations? should be Ngenerations = 1e5

##### a) draw 1 million values from the prior distribution of tsplit. tsplit can be any value between 50,000 and 1,000,000, assume that tsplit~U[1000,1000000] .............this is incongruent with the first exercise, so assuming the range min is 50,000.

prior3<-runif(n=1e6, min=50000, max=1e6)

##### b) For each value of tsplit, simulate a TMRCA (again, mimic the first exercise)

rate3<- 1/(prior3)  
tsplit3<- rexp(1e6, rate3)  
length(tsplit3)

## [1] 1000000

TMRCA<- tsplit3 + 1e5 #1e5 is the expected ancestral coalescent time  
length(TMRCA)

## [1] 1000000

##### c) For each TMRCA, add a Poisson number of mutations

ratemu3<- 1e-3  
snpnumber3<- rpois(1e6, TMRCA\*ratemu3\*2)  
  
#snpnumber  
max(snpnumber3)

## [1] 23642

min(snpnumber3)

## [1] 153

length(snpnumber3)

## [1] 1000000

#hist(snpnumber3)

##### d) Choose which draws from the prior generate data that approximate the observed number of SNPs (600) in the actual data.

##### -accept values of tsplit that give between 550 and 650 SNPs

sims3<- cbind(prior3, snpnumber3)  
simslower3<- subset(sims3, sims3[,2]>550)  
simsupper3<- subset(simslower3, simslower3[,2]<650)  
posterior3<-simsupper3[,1]  
#posterior  
max(posterior3)

## [1] 999982.8

min(posterior3)

## [1] 50019.77

#hist(posterior3)

##### e) It should be all done!

### EXERCISE FOUR

##### Make a density plot of both the prior and posterior distributions of tsplit. Plot on the same axes, and label each dist.

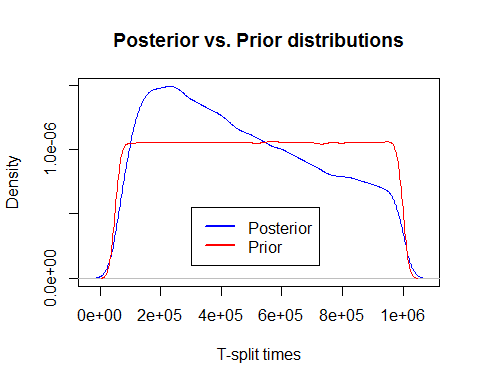
density(posterior3)

##   
## Call:  
## density.default(x = posterior3)  
##   
## Data: posterior3 (62780 obs.); Bandwidth 'bw' = 2.553e+04  
##   
## x y   
## Min. : -26571 Min. :6.070e-10   
## 1st Qu.: 249215 1st Qu.:7.249e-07   
## Median : 525001 Median :9.471e-07   
## Mean : 525001 Mean :9.056e-07   
## 3rd Qu.: 800787 3rd Qu.:1.263e-06   
## Max. :1076573 Max. :1.489e-06

plot(density(posterior3), main= "Posterior vs. Prior distributions", xlab= "T-split times", col= "blue")  
density(prior3)

##   
## Call:  
## density.default(x = prior3)  
##   
## Data: prior3 (1000000 obs.); Bandwidth 'bw' = 1.557e+04  
##   
## x y   
## Min. : 3293 Min. :1.490e-09   
## 1st Qu.: 264148 1st Qu.:1.050e-06   
## Median : 525003 Median :1.053e-06   
## Mean : 525003 Mean :9.574e-07   
## 3rd Qu.: 785858 3rd Qu.:1.056e-06   
## Max. :1046713 Max. :1.060e-06

#?lines  
# "lines() adds a line to an existing plot --- lines(namevector, col = color)  
lines(density(prior3), col= "red")  
#?legend  
legend(3e05,5.5e-07, # legend location   
 c("Posterior", "Prior"), # text in the legend  
   
 lty=c(1,1), # legend symbols (lines)  
   
 lwd=c(2.5,2.5),col=c("blue", "red")) #lines the correct color and width



### EXERCISE FIVE

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

median(posterior3)

## [1] 436178.9

The median value of the posterior is 4.361788810^{5} generations.

### EXERCISE SIX

##### Generate a 95% credible interval for the posterior.

The 95% credible interval for the posterior of tsplit is the range, 8.786487810^{4}, 9.639088910^{5} , meaning that there is a 95% percent chance that the actual value for the split time falls within this range.

### EXERCISE SEVEN

##### How does the posterior dist. differ from the prior dist? The degree to which the posterior differs from the prior relates to the amt. of info. in the data.

As in the first exercise, the posterior distribution differs from the prior in that, while covering the same range of values, the density of values is not equally distributed across that range. Instead, there is more of a normal-distribution-esque curve to the density of values. The curve centers roughly around a median value. In this particular simulation, the degree of difference between the prior and the posterior is not particularly dramatic due to the amount of input information (which is similar to the amount used in the first round of exercise 1).

### EXERCISE EIGHT

##### What if we did not know the true value of N? Repeat the ABC approach to estimate tsplit, but this time, rather than assume N=100,000, draw N from a distribution ~U[1000, 1000000]

priorA<-runif(n=1e6, min=50000, max=1e6) #set A is for the split time  
priorB<-runif(n=1e6, min=1000, max=1e6) #set B is for the ancestral time  
  
  
rateA<- 1/(priorA)  
tsplitA<- rexp(1e6, rateA)  
length(tsplitA)

## [1] 1000000

rateB<- 1/(priorB)  
tsplitB<- rexp(1e6, rateB)  
length(tsplitB)

## [1] 1000000

TMRCAfinal<- tsplitA + tsplitB #tsplitB is the expected ancestral coalescent time  
length(TMRCAfinal)

## [1] 1000000

#add in mutations  
ratemu3<- 1e-3  
snpnumberfinal<- rpois(1e6, TMRCAfinal\*ratemu3\*2)  
  
#snpnumber  
max(snpnumberfinal)

## [1] 25047

min(snpnumberfinal)

## [1] 0

length(snpnumberfinal)

## [1] 1000000

#hist(snpnumber3)  
  
#focusing on prior and posterior of tsplitA (the split time)  
  
simsf<- cbind(priorA, snpnumberfinal)  
simslowerf<- subset(simsf, simsf[,2]>550)  
simsupperf<- subset(simslowerf, simslowerf[,2]<650)  
posteriorf<-simsupperf[,1]  
#posterior  
max(posteriorf)

## [1] 999895.3

min(posteriorf)

## [1] 50008.21

#hist(posteriorf)

### EXERCISE NINE

##### Do your estimates of the median and credible intervale of tsplit differ from the previous simulation?

The 95% credible interval for the posterior of tsplit in this simulation is the range, 6.690226910^{4}, 9.565241110^{5} , and the median value for tsplit in this simulation is 3.9126310^{5}

In the previous simulation, the 95% credible interval for the posterior of tsplit was 8.786487810^{4}, 9.639088910^{5}, and the median value for tsplit was 4.361788810^{5} .