Covid Mutation rates (Bayesian approach)

Felix Barenys Marimon

1/8/2022

1 Abstract

The aim of this study is to analyze the mutation rate of the COVID-19 virus based on the data available in the cov-lineages[1] website and use Bayesian statistical methods in order to be able to estimate the real mutation rate of the virus based on an article[2] that estimates that the real mutation rate is 40% higher than the one seen in data.

[1] https://cov-lineages.org/lineage_list.html

[2] https://www.bath.ac.uk/announcements/mutation-rate-of-covid19-virus-is-at-least-50-per-cent-higher-than-previously-thought/

Contents

2 Data Handling

In this article we will work with the time difference (in days) between days recorded in the website.

```r
covid_rep<-read.csv("/Users/felixbarenysmarimon/Desktop/PROJECT/mutations covid/covid_variants.csv")
covid_rep<-covid_rep[,c(-1,-2,-4,-6,-8,-10,-12,-14,-15,-16)]
covid_rep$Earliest.date<-as.Date(covid_rep$Earliest.date)
covid_rep<-covid_rep[order(covid_rep$Earliest.date),]

#establishing which are the worrying mutations
worring<-c("B.1.351 ","P.1 ","B.1.617.2 ","B.1.1.529 ")
covid_rep$worry<-"No"
covid_rep$worry[which(is.element(covid_rep$Lineage,worring)==TRUE)]<-c("Beta",
                                                                      "Gamma","Delta","Omicron")

#establish alpha for plot color

worring<-c("B.1.351 ","P.1 ","B.1.617.2 ","B.1.1.529 ")
covid_rep$alpha<-0.1
covid_rep$alpha[which(is.element(covid_rep$Lineage,worring)==TRUE)]<-1


#Eliminate Missings:
covid_rep<-covid_rep[-which(is.na(covid_rep$Earliest.date)),]


#difference between mutations in time(days)
covid_rep$diff_time<-c(0,as.numeric(diff.Date(covid_rep$Earliest.date)))
cov<-covid_rep$diff_time
```

3 Data distribution

Here the distribution of the data is shown below:

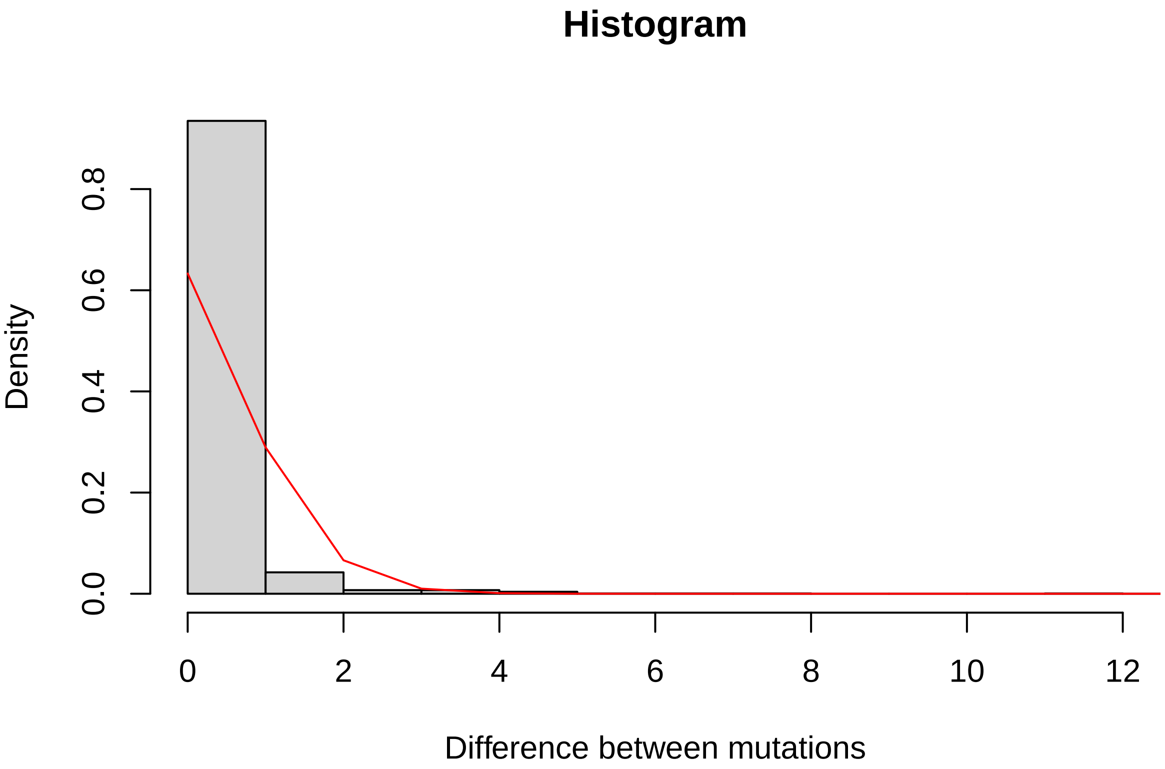
```r
tab<-t(table(cov))
kable(tab,caption="Days between mutations/ nº mutations")
```

Table 1: Days between mutations/ nº mutations

0	1	2	3	4	5	6	7	8	12	18	34	53
1109	327	65	11	11	6	1	1	1	1	1	1	1

If we consider that the mutations occur independently in time, the time between mutations should follow a poisson with lambda 0.457:

```r
hist(covid_rep$diff_time,xlim=c(0,12),breaks=60,
     xlab="Difference between mutations",main="Histogram",
     probability =TRUE)
lines(0:max(cov), dpois(0:max(cov), mean(cov)), col = 'red')
```



4 Bayesian model

First we will compute the likehood function of the Poisson model:

$$L(\lambda; x_1...x_n) = \prod_{i=1}^{n} exp(-\lambda)\frac{1}{x_i!}\lambda^{x_i} = exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} i = n!}$$

and with our data:

$$L(\lambda) = exp(-1536\lambda)\frac{\lambda^{702}}{4.274883e+69} \propto exp(-1536\lambda)\lambda^{702}$$

A good prior distribution from this data could be a Gamma because of its proprieties, which could be centered 0.457 which is a 50% higher than the lambda of the distirbution of the data and a with a standard deviation of 0.01: Now we can estimate the parameters of the gamma distribution having in account that:

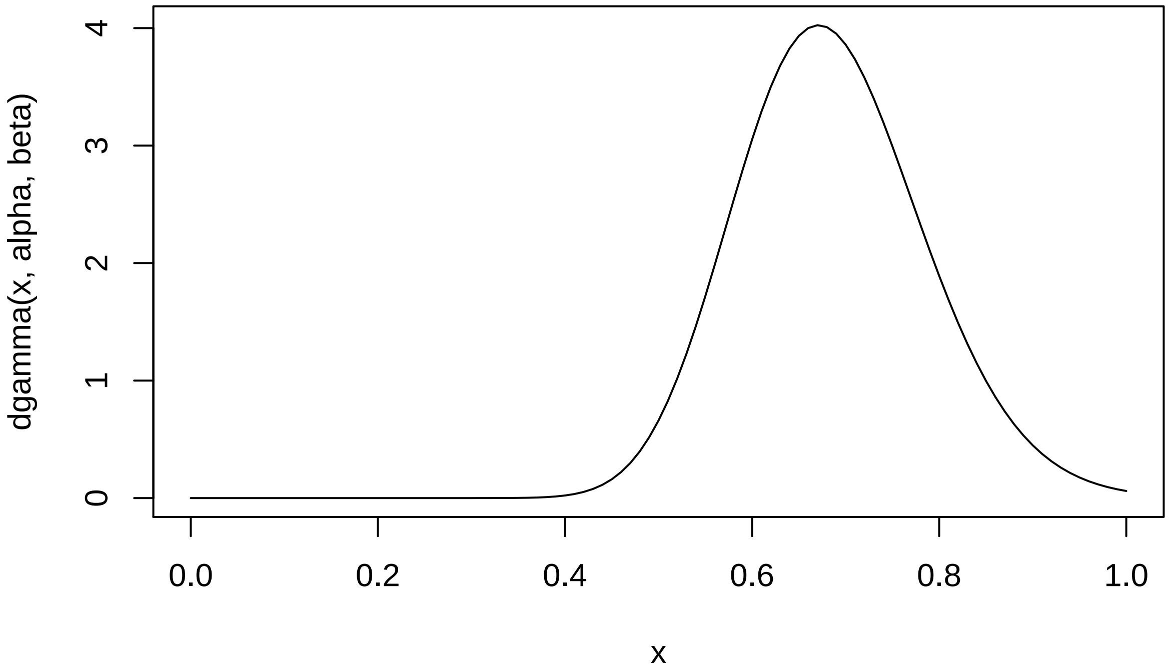
$$\pi(\lambda) = \frac{\beta^{\alpha}(\lambda)^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}$$

with:

$\alpha = \frac{\bar{x}^2}{s^2}$ and $\beta = \frac{\bar{x}}{s^2}$

```
alpha=(1.5*mean(cov))^2/0.01
beta=(1.5*mean(cov))/0.01

curve(dgamma(x,alpha,beta))
```



And then we can calculate the posterior having in account that a posterior from a Poisson and Gamma distributions in equal to:

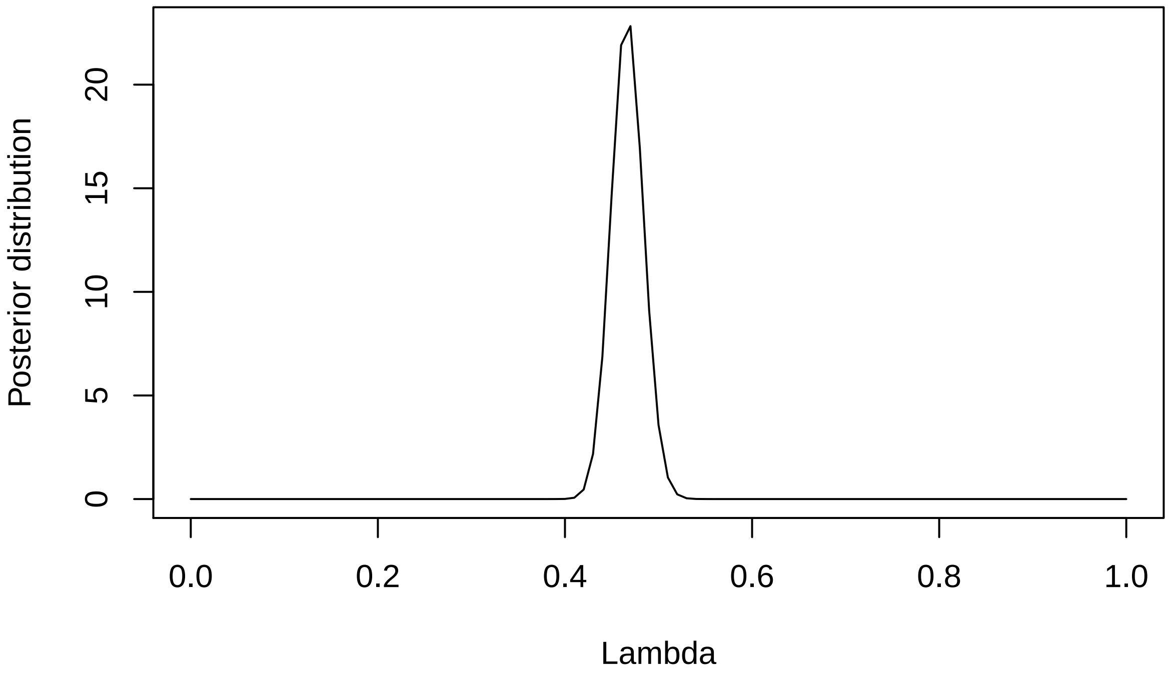
$$f(\lambda/X) \propto L(X/\lambda)\pi(\lambda) \propto \lambda^{\alpha+\sum_{i=1}^{n} x_i - 1} e^{-(\beta+n)\lambda}$$

that in this case is:

$$f(\lambda/X) \propto \lambda^{747} e^{-1604.5\lambda}$$

that is proportional to a Gamma(748,1604.55).

```
curve(dgamma(x,alpha+sum(cov),beta+length(cov)),ylab="Posterior distribution",xlab="Lambda")
```



5 Credibility intervals

With the posterior lambda distribution we can calculate a credibility interval for the lambda:

```
lower<-qgamma(0.025,alpha+sum(cov),beta+length(cov));lower
```

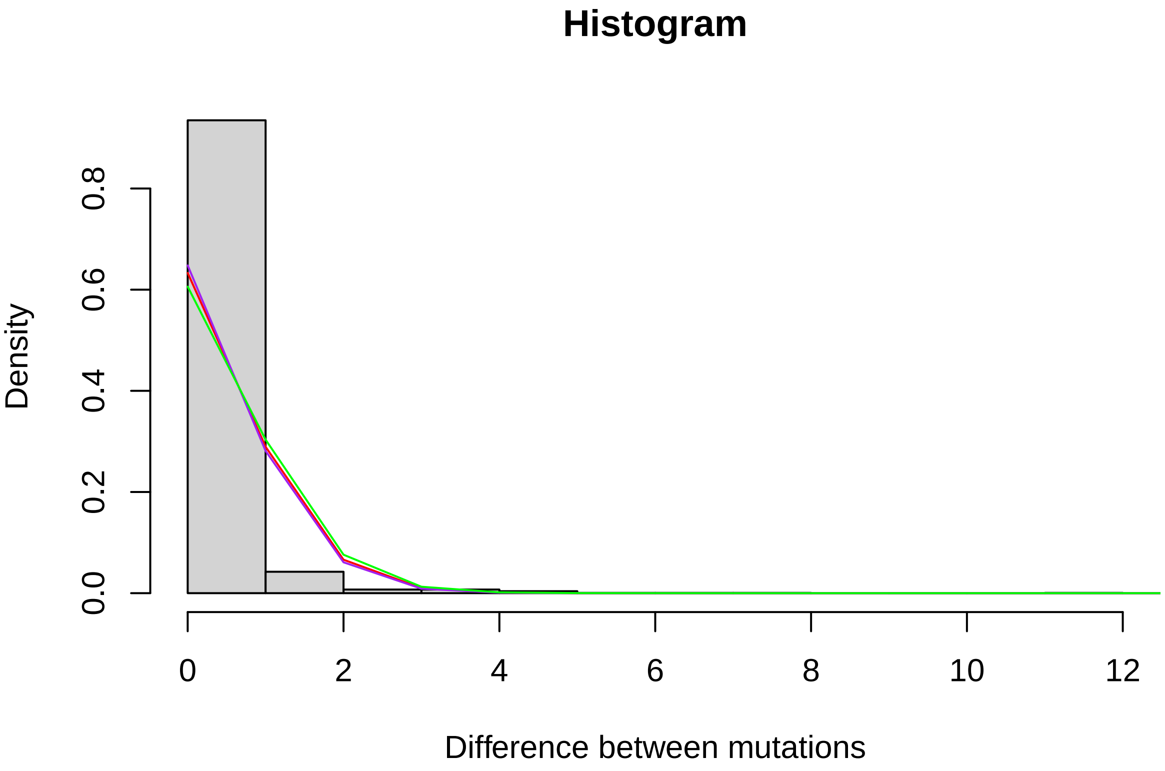
```
## [1] 0.4339589
```

```
upper<-qgamma(0.975,alpha+sum(cov),beta+length(cov));upper
```

```
## [1] 0.5008107
```

And then we can plot the distribution with the credibility intervals:

```
hist(covid_rep$diff_time,xlim=c(0,12),breaks=60,
     xlab="Difference between mutations",main="Histogram",
     probability =TRUE)
lines(0:max(cov), dpois(0:max(cov), mean(cov)), col = 'red')
lines(0:max(cov), dpois(0:max(cov), lower), col = 'purple')
lines(0:max(cov), dpois(0:max(cov), upper), col = 'green')
```



6 Predictive posterior distirbution

The predictive posterior distribution is defined as:

$$p(X_{new}/X) = \int_{\Theta} p(X_{new}/\theta) f(\theta/X) d\theta$$

being Θ the domain of the parameter. With our posterior and the Poisson distribution from the data we have:

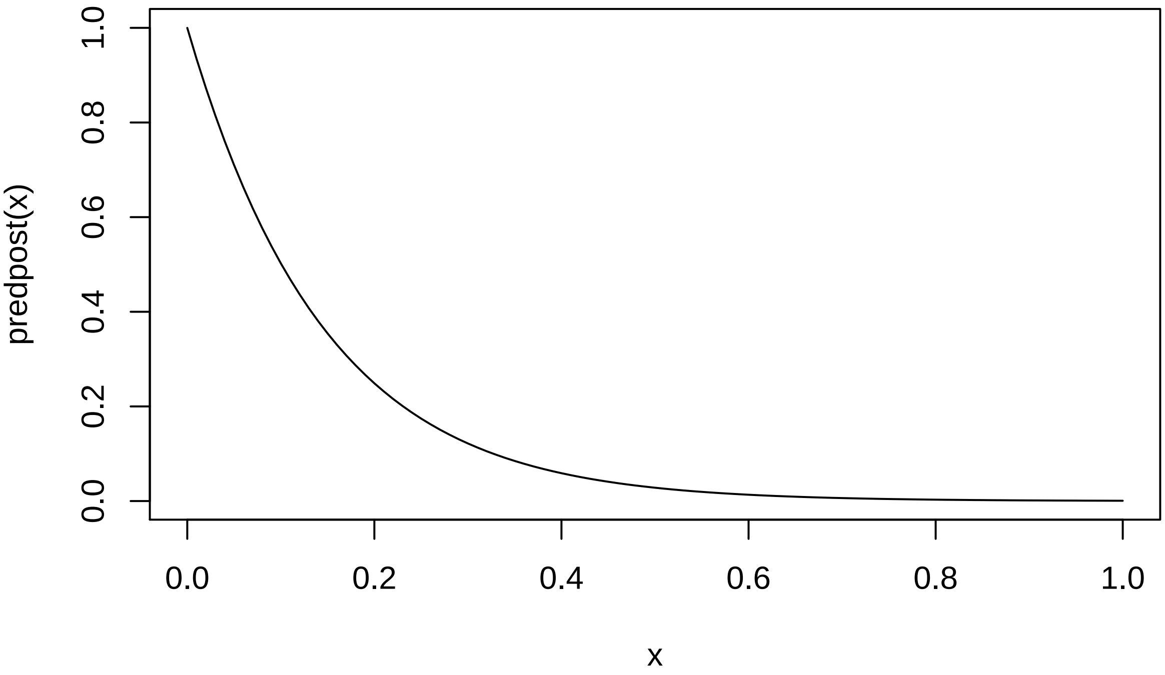
$$p(X_{new}/X) = \int_0^{\infty} e^{-\lambda} \frac{1}{x_{new}!} \lambda^{x_{new}} \frac{1604.5^{748} (\lambda)^{747} e^{-1604.5\lambda}}{\Gamma(748)} d\lambda$$

$$= \frac{1604.5^{748} * \Gamma(x_{new} + 748)}{x_{new}! \Gamma(748) 1605.5^{x_{new}+748}} = \frac{\Gamma(x_{new} + 748)}{x_{new}! \Gamma(748) 1605.5^{x_{new}}}$$

and having in account that $\Gamma(748) \simeq \infty$,

$$p(X_{new}/X) = \frac{1}{x_{new}! 1605.5^{x_{new}}}$$

```
predpost<-function(x){1/(factorial(x)*1605^x)}
curve(predpost(x))
```



7 Conclusions

Applying the Bayesian approach to the data, the real mutation rate (considering the Gamma prior), should be between 0.4339589 and 0.5008107.

More in depth exploration and analysis must be done in order to be able to predict more accurately the mutation nature of the virus.