Final Project HY390.51
June 2023

Deadline: June 25 2023 midnight.

You would like to estimate the rate of exponential growth of the European population of SARS-CoV-2. For this reason you have downloaded data and preprocess them. The final dataset are in the file ms_obs_final.out . (observed dataset).

each line in this file is a genome and each column a polymorphic position in the genome. This means that I have removed all columns that contain only 1 or only 0s. 1 means that there is a mutation compared to the reference genome of the bat SARS virus. 0 means that there is no such a mutation.

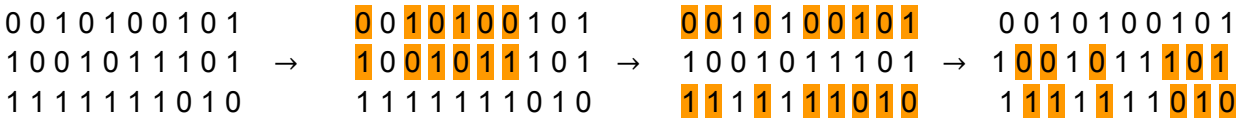
To find the rate of exponential growth, we have produced a dataset of 10,000 simulated datasets, which are stored in the file ms_sim_final.out .

Each dataset has been produced using the corresponding parameter in the pars_final.txt .

Each dataset in the ms_sim_final.out is in the same format as the observed dataset.

Use the following statistics to perform the analysis

A. $k = \frac{\sum_{i<j} k_{ij}}{\binom{n}{2}}$, i.e., the average number of pairwise differences between a pair of sequences. k_{ij} is the number of differences between the sequences i, και j. For example, for 3 sequences:



The first vs second: six differences
The first vs third: 8 differences
The second vs third: 6 differences. Thus, $k = \frac{6+8+6}{3} = \frac{20}{3}$

B. $w = \frac{S}{a_1}$, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, S is the number of polymorphic positions, ie.. $S = 10$ και $a_1 = \frac{1}{1} + \frac{1}{2} = 1.5$ thus $w = \frac{10}{1.5}$.

Γ. Tajima's D which is defined as D = $\frac{k-w}{\sqrt{e_1 S + e_2 S(S-1)}}$, where

$$e_1 = \frac{c_1}{a_1}, e_2 = \frac{c_2}{a_1^2 + a_2}, c_1 = b_1 - \frac{1}{a_1}, c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}, b_1 = \frac{n+1}{3(n-1)}, b_2 = \frac{2(n^2+n+3)}{9n(n-1)}, a_1 = \sum_{i=1}^{n-1} \frac{1}{i}, a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

and *n* is the number of sequences you have.

Questions:

1. Write your own functions to calculate each of the above and eventually the value of *w, k and D.*
2. Thus, for the observed dataset (ms_obs_final.out) and ***each of*** the simulated dataset (ms_sim_final.out), calculate the value of the three statistics. Thus, you will get a single value w_0, k_0, and D_0 for the observed dataset. You will get a vector **w = (**$w_1, w_2, w_3, \ldots, w_{10000}$), **k = (**$k_1, k_2, k_3, \ldots, k_{10000}$), **D = (**$D_1, D_2, D_3, \ldots, D_{10000}$)
3. Normalize each of the vectors, i.e. for each of the vectors **w, k, D, (e.g. for w)** estimate the mean $\bar{w}$(You can use the function mean() in R) and the variance (you can use the function var() in R) and find the normalized values $w_i' = \frac{w_i - \bar{w}}{var(\mathbf{w})}$ (do the same for the **k** and **D** values). Finally using the same $\bar{w}$ and the same var(**w)** transform the w_0. (do the same for the k_0 and D_0 values).
4. Write a function that will calculate the Euclidean distances between the observed and *each of the* simulated datasets. Remember that a Euclidean distance $d = \sqrt{(D_0' - D_1')^2 + (w_0' - w_1')^2 + (k_0' - k_1')^2}$ Let these distances be: *d1, d2, …, d10000.* Now, the important issue is

that for each of these distances, it corresponds a parameter value in pars_final.txt .

5. Find the 500 smallest distances and the keep the indexes.
6. Get the corresponding values from the pars_final.txt.
7. Calculate the mean, the median of these 500 values.
8. Construct the histogram (hint function hist) and the density plot (hint function density) of these 500 values.
9. Explain the final result, i.e., if the rate values in pars_final.txt correspond to the time period of a year, then what does it mean for the population of the SARS-CoV-2?

Important:

1. give the R code in a single R file that will be able to do all the analysis assumeing that that datasets are **in the SAME folder of the source R script. i.e., DO NOT USE ANY PATH TO READ THE FILES**

   OR READ THE FILES DIRECTLY FROM THE WEB LOCATION.

2. Give a final text where you will describe the steps that you followed (similar to the report of the exercise). **It will contain the R code as well as the final plots (question 8) and the estimate of the growth rate and your explanations (question 9).**