Τελικό Πρότζεκτ

Θέλετε να υπολογίσετε το ρυθμό εκθετικής αύξησης στον ευρωπαϊκό πληθυσμό του κορονοϊού. Για το λόγο αυτό, έχω κατεβάσει δεδομένα από το gisaid (μια βάση δεδομένων για τον κορονοϊό) τα έχω προεπεξεργαστη και τα έχω φέρει στην μορφή του αρχείου ms_obs_final.out .
Κάθε γραμμή είναι ένα γονιδίωμα κορονοϊού και κάθε στήλη ειναι μια **πολυμορφική** θέση στο γονιδίωμα του (δηλαδή έχω αφαιρέσει τις στήλες που έχουν μόνο 0 ή μόνο 1). 1 σημαίνει οτι σε σχεση με το γονιδίωμα του κορονοϊού της νυχτερίδας υπάρχει μετάλλαξη, ενώ 0 σημαίνει ότι δεν υπάρχει.

Για να κάνουμε την διαδικασίας εύρεσης του ρυθμού εκθετικής αύξησης, χρησιμοποιούμε την διαδικασία abc, έχω λοιπόν παράξει 10,000 datasets στο αρχειο: ms_sim_final.out **.**

Με τιμές παραμέτρου εκθετικής αύξησης που δίνονται στο αρχείο pars_final.txt .

Τα datasets που παράγονται με τις τιμές αυτές δίδονται στο αρχείο sims.txt και ειναι όπως το πραγματικό dataset, δηλαδή στην μορφή γονιδιωμάτων (σε 0-1 μορφή).

Χρησιμοποιήστε τα ακόλουθα 2 statistics ώστε να υπολογίσετε τον ρυθμό αύξησης στο πραγματικό dataset.

A. $\hat{k} = \frac{\sum\sum_{i<j} k_{ij}}{\binom{n}{2}}$, δηλαδή ο μέσος όρος των διαφορών των αλληλουχιών ανά 2. Για παράδειγμα αν εχουμε τρεις αλληλουχίες:

0 0 1 0 1 0 0 1 0 1 → 0 0 1 0 1 0 0 1 0 1 → 0 0 1 0 1 0 0 1 0 1 → 0 0 1 0 1 0 0 1 0 1
1 0 0 1 0 1 1 1 0 1 → 1 0 0 1 0 1 1 1 0 1 → 1 0 0 1 0 1 1 1 0 1 → 1 0 0 1 0 1 1 1 0 1
1 1 1 1 1 1 1 0 1 0 → 1 1 1 1 1 1 1 0 1 0 → 1 1 1 1 1 1 1 0 1 0 → 1 1 1 1 1 1 1 0 1 0

Τοτε η πρώτη με την δευτερη έχουν 6 διαφορές. Η πρώτη με την τρίτη 8 διαφορές και η δευτερη με την τρίτη 6 διαφορές, άρα $k = \frac{6+8+6}{3} = \frac{20}{3}$

B. $w = \frac{S}{a_1}$, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, όπου S ειναι ο αριθμός των πολυμορφικών θέσεων, δηλαδή των στηλών με 2 καταστάσεις (που έχουν και 0 και 1).
Στο προηγούμενο παράδειγμα $S = 10$ και $a_1 = \frac{1}{1} + \frac{1}{2} = 1.5$ αρα $w = \frac{10}{1.5}$.

Ερωτήσεις:

1. Υπολογίστε το ρυθμό αυξησης χρησιμοποιώντας το πακέτο abc στην R. Γι αυτο χρησιμοποιήστε την posterior κατανομή, της παραμέτρου μετά από την διόρθωση που γίνεται με loclinear. Δώστε την τιμή του mean της posterior κατανομής ως εκτιμητή. Χρησιμοποιήστε hcorr=TRUE, και tol=0.1

2. Χρησιμοποιήστε την συνάρτηση ci() του πακέτου bayestestR με ορισμα method="HDI" και δεδομένα τις τιμές από την posterior κατανομή μετά την διόρθωση loclinear (adj.vaues) ωστε να βρείτε τα credible intervals της posterior (δηλαδή εντός ποιων τιμών πιστευετε οτι βρίσκεται η τιμή με βεβαιότητα 95%). Ένα καλό παράδειγμα για το ci() βρίσκεται εδώ

3. Φτιάξτε ένα διάγραμμα με την prior κατανομή της εκθετικής αύξησης, την posterior πριν την διορθωση και την posterior μετά την διόρθωση (unadj.values και adj.values αντίστοιχα). Βάλτε διαφορετικά χρώματα ώστε να ξεχωρίζουν. Χρησιμοποιήστε την συνάρτηση density ώστε να βρείτε τις πυκνότητες των τριών κατανομών από τα δείγματα και να τις απεικονίσετε γραφικά. Στο διάγραμμα δείξτε με κάθετες γραμμές και χρώμα της αρεσκείας σας (για κάθετες γραμμές στην τιμή X του άξονα-χ χρησιμοποιήστε την abline(v=X) ή αλλιώς ggplot) τις τιμές του mean για την παράμετρος της εκθετικής αύξησης (από το ερώτημα 1) καθώς και την αριστερή και δεξιά τιμή από το credible interval στην ερώτηση 2.

Σημαντικό:
Δώστε σαν τελικό αρχείο τον κώδικα σε R, ο οποίος θα πρέπει να παράγει τα αποτελέσματα ειτε στο τέρμιναλ ειτε να τα γράφει σε αρχείο και να παραγει τα πλοτς σε PDF μορφή. Γράψτε σχολια στον πρόγραμμα. Πχ δώστε την λεζάντα και εξηγήσεις τι είναι η κάθε γραμμή με το κάθε χρώμα για τα πλοτς. Ο R κώδικας θα πρέπει να μπορεί να τρέξει και να αναπαράγει τις εικόνες του PDF.

Θεωρείστε ότι το αρχείο που διαβάζετε βρίσκεται στον ίδιο φάκελο με το R Script. Δεν είναι ανάγκη να δώσετε τα input αρχεία (που σας δίνω εγώ από τα λινκς παραπάνω).