PROJECT 1
NON PARAMETRIC STATISTIC

Εύα Χαριζάνου

Δεκέμβριος 2021

1 (1-3) Ανάθεση μυστικού κωδικού

Η εργασία ξεκινάει με την δημιουργία συνάρτησης με σκοπό την εύρεση του πορσωπικού μου κωδικού. Συνεπώς, δημιουργώ data frame με την πρώτη στήλη του να είναι τα γράμματα του αγγλικού αλφαβήτου (χρήση έτοιμης συνάρτησης από την R) και τη δεύτερη στήλη του να είναι η αξία κάθε γράμματος. Έπειτα, ορίζω το my vec χωρίς να κάνω χρήση της συνάρτησης strsplit καθώς θεώρησα δεδομένο ότι εισάγονται τα δεδομένα έτσι όπως τα χρειάζομαι. Στη συνέχεια, δημιουργώ τη συνάρτηση calc pass και εισάγοντας τα στοιχεία μου, μου επιστρέφει τον αριθμό 152. Οι δύο επόμενες συναρτήσεις είναι απλή αντιγραφή από τις έτοιμες που μας δόθηκαν στις εκφωνήσεις με αλλαγή συμβόλων από την Python στην R με μία προσοχή στο for loop καθώς αλλάζει το εύρος τιμών από τη μία γλώσσα στην άλλη. Τέλος, αρχικοποιώ τον γεννήτορα των ψευδοτυχαίων αριθμών με τον τύπο που μου έχει δοθεί, ο οποίος λαμβάνει την τιμή 50.

2 (3.1) Τυχαίο δείγμα και γράφημα

Εδώ γίνεται η δημιουργία γραφήματος από τυχαίο δείγμα X 500 παρατηρήσεων που προέρχεται από τη logN(0,1) κατανομή. Γίνεται χρήση των συναρτήσεων geom histogram geom function της ggplot2 βιβλιοθήκης όπου η πρώτη επιστρέφει ιστόγραμμα και η δεύτερη καμπύλη την οποία προσθέτω πάνω στο υπάρχον plot.

3 (3.2) Πραγματοποίηση ελέγχων καλής προσαρμογής

Από το δείγμα που έχουμε δημιουργήσει από τη logN(0,1) κατανομή πραγματοποιούμε ελέγχους Kolmogorov-Smirnov και X^2.

Table 1: P-values

P-values		
Πλήθος παρατηρήσεων	Kolmogorov-Smirnov	X^2
20	0.03513	0.3493
100	0.9971	0.2354
500	0.8139	0.3793

Πριν σχολιάσουμε τις p value, ας αναφερθούμε στο πώς χρησιμοποίησα chi-square test σε συνεχές δείγμα. Η όλη ιδέα βασίζεται στο να χρησιμοποιήσουμε το ιστόγραμμα για να "πακετάρουμε" τις παρατηρήσεις μέσα σε περιοχές τιμών, οπότε και να πάρουμε πίσω διακριτές τιμές, το πλήθος των παρατηρήσεων που "πέφτουν" σε κάθε περιοχή. Θέτω true το simulate p value έτσι ώστε να είμαι σίγουρη ότι το αποτέλεσμα είναι αρκετά ακριβές, ειδικά σε μικρά σύνολα, αφού χρησιμοποείται Monte Carlo. Γενικά, είναι αρκετά αυθαίρετη η ομαδοποίηση ως μέθοδος, δηλαδή δεν υπάρχει συγκεκριμένος τρόπος να ομαδοποιήσεις τις παρατηρήσεις σου, οπότε εγώ επέλεξα ένα d και όρισα ισοπίθανα διαστήματα. Τώρα όσον αφορά τους ελέγχους με υποθέσεις τις:

H_0: το δείγμα προσαρμόζεται καλά στην logN(0,1) κατανομή
H_1: εναλλακτικά

παίρνουμε τα παραπάνω p value και αν θεωρήσουμε ε.σ.σ. $\alpha = 0.05$ δεν απορρίπτουμε την μηδενική υπόθεση σε κανέναν έλεγχο πέρα από τον Kolmogorov-Smirnov στις 20 παρατηρήσεις. Γενικά, επαληθεύεται σε έναν μεγάλο βαθμό ότι το Kolmogorov-Smirnov τεστ χρειάζεται μεγάλο δείγμα για να βελτιωθεί αξιοσημείωτα το p value.

4 (3.3)Συνάρτηση test LogNormal

Εδώ δημιουργώ συνάρτηση με ορίσματα το πλήθος των επαναλήψεων, το μέγεθος των τυχαίων δειγμάτων και το ε.σ.σ, η οποία επιστρέφει το ποσοστό των απορρίψεων στους ελέγχους Kolmogorov-Smirnov και X^2.

Καλώντας την με τα εξής δεδομένα $(10^4, 20, 0.05)$ μου επιστρέφει (0.0517, 0.0425) ενώ με τα $(10^4, 500, 0.05)$ μου επιστρέφει (0.0479, 0.05). Για το Kolmogorov-Smirnov τεστ επαληθεύεται αυτό που γνωρίζουμε από τη θεωρία, ότι απορρίπτουμε τη μηδενική υπόθεση σχεδόν το 5% των φορών ασχέτως του μεγέθους του δείγματος σ'αντίθεση με το chi square τεστ.

5 (3.4) Προσομοίωση δείγματος από την ομοιόμορφη

Η απάντησή μου σε αυτό το ερώτημα βασίζεται στην πρόταση 3.3 της θεωρίας:

Αν η τ.μ. $Q \sim F$ και X συνεχής τ.μ. τότε:
η τ.μ. $U = F(X) \sim Unif(0, 1)$

και οι έλεγχοι που χρησιμοποιώ είναι οι Ks test & t test με p value=0.8088 0.7777 αντίστοιχα (δεν απορ. στο σύνηθες ε.σ.σ.).

6 (3.5) Προσομοίωση δείγματος στην λογαριθμοκανονική

Η απάντησή μου σε αυτό το ερώτημα βασίζεται στην πρόταση 3.2 της θεωρίας:

Αν $U \sim Unif(0, 1)$ τότε η τ.μ. $X = F^{-}(U) \sim F$

Εκτελώντας ελέγχους για το δείγμα με 20 παρατηρήσεις παίρνω:

ks test: p-value=0.007363 (απορ)
chisq test: p-value=0.1116 (δεν απορ)

Ενώ για όλο το δείγμα παίρνω:

ks test: p-value=0.07612(δεν απορ)
chisq test: p-value=0.2493(δεν απορ)