ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ: *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: *2019-2020*

Εργαστηριακή Άσκηση στην R
21/2/2020
Τίτλος: Προγραμματισμός με Χρήση της R

Άσκηση 1η: Η συνάρτηση μάζας πιθανότητας της κατανομής Poisson είναι η

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,..., \quad \lambda > 0.$$

Επειδή όμως χρειάζεται να υπολογιστεί το παραγοντικό στον παρονομαστή, ο παραπάνω ορισμός δεν είναι ιδιαίτερα εύχρηστος. Για αυτό το λόγο στην πράξη χρησιμοποιούμε αναδρομικό τύπο και συγκεκριμένα, υπολογίζουμε την $P(X = 0) = e^{-\lambda}$ και στη συνέχεια οι υπόλοιπες πιθανότητες προκύπτουν ως

$$P(X = x) = P(X = x-1)\frac{\lambda}{x}, \quad x = 1, 2,...$$

Να γράψετε μια συνάρτηση στην R που να έχει ως παραμέτρους εισόδου την τιμή της παραμέτρου λ και την τιμή x για την οποία θέλετε να υπολογίσετε την πιθανότητα και η οποία να επιστρέφει ως αποτέλεσμα την συνάρτηση μάζας πιθανότητας της κατανομής Poisson στο σημείο x. Αν η τιμή εισόδου x δεν είναι φυσικός αριθμός καθώς και αν η τιμή εισόδου λ δεν είναι μεγαλύτερη του μηδενός η συνάρτησή σας θα πρέπει να επιστρέφει ένα μήνυμα λάθους.

Άσκηση 2η: Α) Να γράψετε μια συνάρτηση στην R η οποία να παίρνει ως παράμετρο εισόδου ένα διάνυσμα δεδομένων **X** (διάστασης n) και κατόπιν να υπολογίζει την ποσότητα

$$G = \frac{2}{n(n-1)}\sum_{i=1}^{n}(2i - n - 1)X_{(i)},$$

όπου $X_{(i)}$ είναι η i διατεταγμένη (κατά αύξουσα τάξη μεγέθους) συνιστώσα του διανύσματος **X**. Η συνάρτηση θα πρέπει να επιστρέφει τόσο την τιμή του G, όσο και τη διάσταση του διανύσματος **X**.

Β) Έστω η τ.μ. X ~ Βήτα (4,5). Με ποιες εντολές της R θα υπολογίζατε:

i) Την πιθανότητα $P(X \leq 0.5)$.
ii) Την τιμή α για την οποία $P(X > \alpha) = 0.2$.
iii) Την τιμή της σ.π.π. της τ.μ. X στο σημείο 0.2.

Άσκηση 3η: Το *bootstrap* είναι μια μέθοδος επαναδειγματοληψίας ιδιαίτερα χρήσιμη για την εκτίμηση του τυπικού σφάλματος ενός εκτιμητή $\hat{\theta}$. Θεωρούμε ότι έχουμε ένα τυχαίο δείγμα αποτελούμενο από N παρατηρήσεις $\mathbf{y} = (y_1, ..., y_N)^T$ και επαναλαμβάνουμε B φορές τα παρακάτω βήματα:

i. Δημιουργούμε ένα bootstrap δείγμα μεγέθους N, με δειγματοληψία με επανάθεση από το αρχικό δείγμα.
ii. Υπολογίζουμε την τιμή του εκτιμητή στο bootstrap δείγμα που δημιουργήσαμε.

Μετά το πέρας της επαναληπτικής αυτής διαδικασίας, έχουμε αποθηκεύσει ένα σύνολο εκτιμητών *bootstrap* $\hat{\boldsymbol{\theta}}_{(boot)} = \left(\hat{\theta}_{(1)}, ..., \hat{\theta}_{(B)}\right)^T$. Τέλος, υπολογίζουμε το τυπικό σφάλμα *bootstrap* του εκτιμητή $\hat{\theta}$ από τον τύπο:

$$se\left(\hat{\theta}_{(boot)}\right) = \left[\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}\right)^2\right]^{1/2},$$

όπου $\bar{\hat{\theta}}$ η μέση τιμή των εκτιμητών *bootstrap*, δηλαδή $\bar{\hat{\theta}} = B^{-1}\sum_{i=1}^{B}\hat{\theta}_{(i)}$.

Να γραφτεί μια συνάρτηση στην R που να υλοποιεί τον αλγόριθμο *bootstrap* αν θεωρήσουμε ότι ο εκτιμητής ενδιαφέροντος είναι ο δειγματικός συντελεστής μεταβλητότητας ($CV = 100 \cdot \left(s/\left|\bar{y}\right|\right)\%$, όπου s είναι η δειγματική τυπική απόκλιση και $\bar{y}$ ο δειγματικός μέσος), με βάση ένα διάνυσμα παρατηρήσεων **y** που θα δίνει ο χρήστης σαν παράμετρο εισόδου καθώς και τον αριθμό B. Η συνάρτηση θα πρέπει να επιστρέφει σε μια λίστα με κατάλληλα ονόματα την τιμή του τυπικού σφάλματος *bootstrap* καθώς και τις τιμές των εκτιμητών *bootstrap*.