Εργασία 4

Δίνεται αρχείο στην R, διαθέσιμο στις εργασίες – ηλεκτρ. τάξη με τα δεδομένα φοιτητών του χειμερινού εξαμήνου 2017, που αφορούν τη βαθμολογία τους την πρώτη φορά που έδωσαν Πιθανότητες **I**.

(i) ανοίξτε το αρχείο, με την R (αφού πρώτα κάνετε εγκατάσταση του λογισμικού). Θα παρατηρήσετε ότι δίνονται 2 σύνολα δεδομένων. Το πρώτο αφορά συνολικά τα δεδομένα και το δεύτερο μόνο για τα ~~κρ~~ίτσια. Μπορείτε να χρησιμοποιήσετε οποιο λογισμικό θέλετε για να αναλύσετε τα δεδομένα. Υπολογίστε το $\overline{X}$ και το S^2. Κάντε το ιστόγραμμα των σχετικών συχνοτήτων για να έχετε μία ιδέα για τη μορφή της κατανομής των βαθμών των φοιτητών. Σχολιάστε.

(ii) Υποθέστε τώρα ότι το δείγμα αυτό, προέρχεται από $Bin(10, p)$. Εκτιμήστε το p με τη μέθοδο των ροπών, και συγκρίνεται το λόγο $\frac{S^2}{\overline{X}}$ με τον αντίστοιχο θεωρητικό. Σχεδιάστε το γράφημα της εκτιμούμενης διωνυμικής πάνω στο ιστόγραμμα σχετ. συχνοτήτων. Σχολιάστε τα αποτελέσματα.

(iii) Αρχίστε να σκέφτεστε πιο ρεαλιστικά μοντέλα. Χωρίστε πρώτα τους φοιτητές σε αδιάβαστους (Α) και διαβασμένους (Δ). Υποθέστε ότι οι Α γράφουν βαθμούς από 0–2, ενώ οι διαβασμένοι από 3–10. Υποθέστε ότι ένας Α, γράφει βαθμό $X \sim Bin(2, p_a)$, ενώ ένας διαβασμένος, γράφει βαθμό $Y \sim 3 + Bin(7, p_\beta)$. Εκτιμώντας το ποσοστό των διαβασμένων με

$$\pi = \frac{\#\text{ φοιτητών με βαθμό} \geq 3}{\#\text{ όλων των φοιτητών}}$$

, εκτιμήστε τα p_a και p_β με τη μέθοδο των ροπών. Χρησιμοποιήστε επίσης τη μέθοδο μέγιστης πιθανοφάνειας, για να εκτιμήσετε τα π, p_a και p_β, αν και τα 3 είναι άγνωστα. Θεωρούμε ότι ο βαθμός

$$Z_i = \begin{cases} X_i, & \text{με πιθ. } 1-\pi \\ Y_i, & \text{με πιθ. } \pi \end{cases}$$

Συγκρίνετε τα αποτελέσματα.

(iv) Έχετε την πληροφορία από έναν Καθηγητή, ότι κάποιοι φοιτητές που γράφουν από 3.5 – 4.0 βαθμολογούνται με 3 και όχι με 4.
Για να ενσωματώσετε αυτήν την πληροφορία στη μοντελοποίηση σκέφτεστε 2 διαφορετικούς τρόπους:

(α) $$Z_i = \begin{cases} X_i \sim Bin(2, p_\alpha), & \text{με πιθ. } 1-\pi_1-\pi_2 \\ 3, & \text{με πιθ. } \pi_2 \\ Y_i \sim 4+Bin(6, p_\beta), & \text{με πιθ. } \pi_1 \end{cases}$$

(β) $$Z_i = \begin{cases} X_i \sim Bin(3, p_\alpha), & \text{με πιθ. } 1-\pi_1-\pi_2 \\ 3, & \text{με πιθ. } \pi_2 \\ Y_i \sim 3+Bin(7, p_\beta) & \text{με πιθ. } \pi_1. \end{cases}$$

Τα μοντέλα αυτά έχουν πλέον 4 παραμέτρους υπό εκτίμηση. Εκτιμήστε με τη μέθοδο μέγιστης πιθανοφάνειας (ενδεχομένως με αλγόριθμο) για κάποιο από αυτά τα (α) και (β).

(v) Προτείνετε και εσείς ένα μοντέλο, και εκτιμήστε τις παραμέτρους του με τη μέθοδο μέγιστης πιθανοφάνειας.

(vi) Συγκρίνετε όλα τα μοντέλα που προτάθηκαν με το κριτήριο $AIC(m) = -2\log L(\hat{\theta}) + 2s$
s = # παραμέτρων υπό εκτίμηση
και καταλήξτε στο καλύτερο μοντέλο που περιγράφει τα δεδομένα σε αυτά τα πραγματικά δεδομένα (γραφήματα επίσης)

<u>Παρατήρηση</u>: * Για το (v) μπορείτε να προτείνεται όποιο μοντέλο θέλετε, αρκεί να είσαστε σε θέση να εκτιμήσετε τις παραμέτρους του, με την ε.μ.π. (αναλυτικά ή με αλγόριθμο)
* Σε σχέση με το (iv) μην ξεχνάτε επίσης ότι πολλοί φοιτητές που γράφουν από 4.0 – 4.5 βαθμολογούνται επίσης με 5
(μπορεί και μικρότερο)