# Τρεις δραστηριότητες στην R

*Διαλέξεις 4–6: δείγμα, ροπές και εκτιμήτριες.*

Για $X_1, \dots, X_n$ γράφουμε

$$\overline{X} = \frac{1}{n}\sum_i X_i, \qquad M_k = \frac{1}{n}\sum_i (X_i - \overline{X})^k, \qquad S^2 = \frac{n}{n-1} M_2.$$

Για θετική διασπορά και πεπερασμένες αντίστοιχες ροπές,

$$\lambda = \frac{\mu_3}{\sigma^3}, \quad \kappa = \frac{\mu_4}{\sigma^4}, \quad \kappa' = \kappa - 3.$$

Όταν $M_2 > 0$, οι δειγματικές εκδοχές είναι $\widehat{\lambda} = M_3/M_2^{3/2}$, $\widehat{\kappa} = M_4/M_2^2$ και $\widehat{\kappa} - 3$.

## 1. Ίδια θέση και μεταβλητότητα, διαφορετική μορφή

**Αρχείο**: `01_ropes_prosomoiomena_pragmatika.R`.

Ξεκινήστε με τις αρχικές επιλογές $n = 40$, $B = 4000$, $\mu_0 = 0$, $\sigma_0 = 1$. Οι τέσσερις πληθυσμοί έχουν τις ακόλουθες ιδιότητες:

| Κατανομή | $\mathbb{E}(X)$ | $\mathbb{V}(X)$ | $\lambda$ | $\kappa'$ |
|---|---|---|---|---|
| $N(0,1)$ | 0 | 1 | 0 | 0 |
| $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$ | 0 | 1 | 0 | $-6/5$ |
| $Z - 1$, $Z \sim \mathrm{Exp}(1)$ | 0 | 1 | 2 | 6 |
| $1 - Z$, $Z \sim \mathrm{Exp}(1)$ | 0 | 1 | $-2$ | 6 |

Εδώ η $\mathrm{Exp}(1)$ έχει πυκνότητα $e^{-z}$ για $z > 0$.

(α) Συγκρίνετε τα τρία κανονικά δείγματα. Γιατί η θεωρητική καμπύλη είναι ίδια, ενώ η προσαρμοσμένη και το ιστόγραμμα αλλάζουν; Παρατηρήστε ότι τα όρια των κλάσεων είναι κοινά.

(β) Υπολογίστε τα χαρακτηριστικά του πρώτου δείγματος κάθε κατανομής. Είναι οι δειγματικές τιμές ίσες με τις θεωρητικές; Μηδενική θεωρητική λοξότητα συνεπάγεται μηδενική δειγματική λοξότητα;

(γ) Από τις επαναλήψεις υπολογίστε μέση τιμή και διασπορά των $\overline{X}, M_2, S^2$. Συγκρίνετε με

$$\mathbb{E}(\overline{X}) = \mu_0, \quad \mathbb{V}(\overline{X}) = \sigma_0^2/n, \quad \mathbb{E}(M_2) = \frac{n-1}{n}\sigma_0^2, \quad \mathbb{E}(S^2) = \sigma_0^2,$$

$$\mathbb{V}(S^2) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma_0^4.$$

(δ) Εξετάστε τα πραγματικά δεδομένα `faithful`. Περιγράψτε τη μορφή τους πριν επιλέξετε μοντέλο. Η κανονική καμπύλη είναι προσαρμογή με μέσο και $M_2$ από τα δεδομένα. Πού διαφέρει από το ιστόγραμμα;

**Για σκέψη.** Τα υποδείγματα από το πραγματικό αρχείο λαμβάνονται χωρίς επανάθεση. Γιατί η τιμή του πλήρους αρχείου δεν είναι γνωστό χαρακτηριστικό του πληθυσμού;

## 2. Διωνυμική κατανομή με δύο άγνωστες παραμέτρους

**Αρχείο:** `02_methodos_ropon_binomial_kai_real_counts.R`.

Για $X_i \sim \text{Bin}(N, p)$, με $N$ θετικό ακέραιο και $0 < p < 1$, οι εξισώσεις ροπών είναι

$$Np = \overline{X}, \qquad Np(1-p) = M_2.$$

Όταν $0 < M_2 < \overline{X}$, οι αλγεβρικές λύσεις με πραγματικό $N$ είναι

$$\overline{p} = 1 - \frac{M_2}{\overline{X}}, \qquad \overline{N} = \frac{\overline{X}^{\,2}}{\overline{X} - M_2}.$$

Η ακέραια διόρθωση του αρχείου επιλέγει $\widehat{N} = \max\{1, \text{round}(\overline{N}), X_{(n)}\}$ και $\widehat{p} = \overline{X}/\widehat{N}$. Διατηρεί την πρώτη εξίσωση, όχι γενικά και τη δεύτερη.

(α) Με $N = 12$, $p = 0.25$ και $n = 40$, ελέγξτε τις εκτιμήσεις από το πρώτο δείγμα. Ποια είναι η διαφορά των $n$ και $N$;

(β) Μελετήστε το δείγμα με εννέα μονάδες και μία τιμή 4. Γιατί η στρογγυλοποίηση της $\overline{N}$ δεν αρκεί; Επαληθεύστε ότι η τελική τιμή καλύπτει όλες τις παρατηρήσεις.

(γ) Σε πόσα από τα $B$ δείγματα υπάρχει εσωτερική λύση; Ποιος είναι ο παρονομαστής της αντίστοιχης σχετικής συχνότητας; Γιατί οι περιγραφές των εκτιμήσεων πρέπει να αναφέρουν ότι αφορούν μόνο αυτά τα δείγματα;

(δ) Στο γράφημα της εμπειρικής κατανομής της $\overline{N}$, ο οριζόντιος άξονας είναι λογαριθμικός και δεν έχουν αφαιρεθεί ακραίες εκτιμήσεις. Τι συμβαίνει όταν $M_2$ πλησιάζει τον $\overline{X}$ από κάτω;

(ε) Συγκρίνετε τη διαφορά $M_2 - \overline{X}$ για διωνυμικό, Poisson και αρνητικό διωνυμικό μοντέλο, με ίδια θεωρητική μέση τιμή. Στο τρίτο μοντέλο δίνεται $\mathbb{V}(X) = \mathbb{E}(X) + \mathbb{E}^2(X)/r_0$. Ποια είναι η θεωρητική μέση τιμή της δειγματικής διαφοράς;

(στ) Στα πραγματικά ετήσια δεδομένα `discoveries`, συγκρίνετε $\overline{x}$ και $m_2$. Εξετάστε τη σειρά καταγραφής και τα διαδοχικά χρονικά τμήματα. Μπορούμε να θεωρήσουμε την ανεξαρτησία και την ισονομία δεδομένες;

**Για σκέψη.** Η ανισότητα $M_2 < \overline{X}$ δεν αποτελεί απόδειξη διωνυμικής κατανομής. Ποια άλλη πληροφορία για τη διαδικασία που παράγει τα πλήθη θα αναζητούσατε;

> **Ανάγνωση των διαγραμμάτων απαρίθμησης.** Οι στήλες δίνουν σχετικές συχνότητες σε ακέραιες τιμές. Τα μαύρα σημεία του προσομοιωμένου παραδείγματος είναι γνωστές θεωρητικές πιθανότητες. Στα πραγματικά δεδομένα τα έγχρωμα σημεία είναι πιθανότητες ενός προσαρμοσμένου μοντέλου Poisson.

## 3. Τρεις εκτιμήτριες ενός άγνωστου άκρου

**Αρχείο:** `03_uniform_sygkrisi_ektimitrion_kai_real_data.R`.

Για τυχαίο δείγμα από $\mathrm{Unif}(0, \theta)$, $\theta > 0$, συγκρίνουμε

$$\widehat{\theta} = X_{(n)}, \qquad \widetilde{\theta} = \frac{n+1}{n} X_{(n)}, \qquad \overline{\theta} = 2\overline{X}.$$

Η $\theta$ μπορεί να είναι οποιοσδήποτε θετικός πραγματικός αριθμός. Η αρχική επιλογή $\theta = 5$ διευκολύνει μόνο την ανάγνωση.

(α) Στο πρώτο δείγμα, διακρίνετε τη θεωρητική ομοιόμορφη πυκνότητα από τις τρεις προσαρμογές. Ποιο είναι το άκρο και ποιο το ύψος κάθε ορθογωνίου; Γιατί δεν πρέπει να εξισώσουμε τα εκτιμημένα άκρα με το πραγματικό $\theta$;

(β) Στα $B$ δείγματα εκτιμήστε τις μέσες τιμές, μεροληψίες, διασπορές και MSE. Συγκρίνετε με τους ακριβείς τύπους:

| | Μεροληψία | Διασπορά | MSE |
|---|---|---|---|
| $\widehat{\theta}$ | $-\theta/(n+1)$ | $\dfrac{n\theta^2}{(n+1)^2(n+2)}$ | $\dfrac{2\theta^2}{(n+1)(n+2)}$ |
| $\widetilde{\theta}$ | $0$ | $\dfrac{\theta^2}{n(n+2)}$ | $\dfrac{\theta^2}{n(n+2)}$ |
| $\overline{\theta}$ | $0$ | $\dfrac{\theta^2}{3n}$ | $\dfrac{\theta^2}{3n}$ |

(γ) Η πυκνότητα του $\widehat{\theta}$ είναι $nt^{n-1}/\theta^n$ για $0 < t < \theta$. Συγκρίνετέ τη με το ιστόγραμμα των εκτιμήσεων. Γιατί αυτή η κατανομή δεν είναι ομοιόμορφη; Τι δηλώνει η κατακόρυφη γραμμή στο $\theta$;

(δ) Αλλάξτε το $n$ και παρατηρήστε τα MSE. Για ποια ακέραια $n \geq 2$ έχει το $\widehat{\theta}$ μικρότερο MSE από το $\overline{\theta}$; Στο $n = 2$ συγκρίνετε και τη διορθωμένη $\widetilde{\theta}$.

(ε) Εφαρμόστε τους ίδιους τύπους στα πραγματικά δεδομένα `faithful`. Εξετάστε το ιστόγραμμα και την εμπειρική συνάρτηση κατανομής. Η αριθμητική ύπαρξη των εκτιμήσεων εγγυάται ότι το μοντέλο είναι κατάλληλο;

**Για σκέψη.** Στα προσομοιωμένα δείγματα γνωρίζουμε το $\theta$ και υπολογίζουμε σφάλματα εκτίμησης. Στα πραγματικά δεδομένα δεν το γνωρίζουμε. Ποια συμπεράσματα μπορούμε ακόμη να αντλήσουμε από τις γραφικές συγκρίσεις και ποια όχι;

Πηγές πραγματικών δεδομένων: τεκμηρίωση R `datasets::faithful` και `datasets::discoveries`. Για τις εξισώσεις ροπών και τις ιδιότητες των εκτιμητριών: σημειώσεις του μαθήματος, Διαλέξεις 4–6.