Σύντομος οδηγός R

Αλλάξουμε μία επιλογή κάθε φορά και παρατηρούμε τι αλλάξει.

1. Πώς εκτελούμε τα αρχεία

Κάθε ένα από τα τρία αρχεία `.R` είναι αυτοτελές και χρησιμοποιεί μόνο συναρτήσεις και δεδομένα που συνοδεύουν την R. Ανοίξτε το αρχείο ως UTF-8. Στο RStudio, πατήστε `Source` για να εκτελεστεί ολόκληρο. Εναλλακτικά, χρησιμοποιήστε `source("όνομα.R", encoding="UTF-8")`, με το πραγματικό όνομα του αρχείου.

Με `save_plots <- TRUE` δημιουργείται ένα PDF γραφημάτων και πίνακες CSV στον φάκελο `apotelesmata/askisi_1`, `askisi_2` ή `askisi_3`, μέσα στον τρέχοντα φάκελο εργασίας. Η `getwd()` δείχνει αυτόν τον φάκελο. Με `save_plots <- FALSE` τα γραφήματα εμφανίζονται στην οθόνη· στο RStudio χρησιμοποιήστε τα βέλη του παραθύρου `Plots` για τα προηγούμενα γραφήματα. Για ελληνικά στα PDF χρησιμοποιείται Cairo ή, στο macOS, Quartz.

2. Οι βασικές επιλογές βρίσκονται στην αρχή

Επιλογή	Τι αλλάζει
`seed`	Διαφορετική αναπαραγώγιμη κλήρωση. Με ίδιες επιλογές και ίδιο `seed` επαναλαμβάνεται το ίδιο πείραμα.
`n`	Παρατηρήσεις σε *κάθε* προσομοιωμένο δείγμα, ακέραιος $n \geq 2$.
`B`	Πλήθος ανεξάρτητων επαναλήψεων. Δεν είναι το μέγεθος κάθε δείγματος.
`classes`	`NULL`: αυτόματη επιλογή κλάσεων. Π.χ. `6`: έξι ισοπλατείς κλάσεις.
`save_plots`	`TRUE`: PDF. `FALSE`: γραφήματα στην οθόνη.

> **Πρώτο πείραμα.** Κρατήστε $n = 40$ και αλλάξτε μόνο το `seed` από `123` σε `456`. Ξανατρέξτε όλο το αρχείο. Οι θεωρητικές παράμετροι παραμένουν ίδιες, αλλά αλλάζουν οι παρατηρήσεις και οι εκτιμήσεις.
> **Δεύτερο πείραμα.** Δοκιμάστε $n = 10, 40, 160$, κρατώντας ίδιο το B. Συγκρίνετε τη μεταβλητότητα των εκτιμήσεων. Κάθε εκτέλεση παράγει νέα δείγματα· δεν προσθέτει κατ' ανάγκη παρατηρήσεις στο προηγούμενο δείγμα.

3. Επιλογή κλάσεων χωρίς αλλαγή των δεδομένων

Στα συνεχή δεδομένα η αυτόματη επιλογή χρησιμοποιεί τον κανόνα Sturges για λιγότερες από 80 τιμές και τον Freedman--Diaconis για τουλάχιστον 80 τιμές. Επιπλέον περιορίζεται σε έως 40 κλάσεις και, κατά μέσο όρο, τουλάχιστον πέντε τιμές ανά κλάση. Αυτός είναι πρακτικός κανόνας της δραστηριότητας, όχι καθολικά βέλτιστη επιλογή. Δεν απαιτεί να έχουν όλες οι κλάσεις πέντε παρατηρήσεις.

Με ίδιο `seed` δοκιμάστε `classes <- 3`, `6` και `12`. Οι παρατηρήσεις και οι εκτιμήσεις δεν αλλάζουν· αλλάζει η απεικόνισή τους. Πολύ πολλές κλάσεις κάνουν το ιστόγραμμα ασταθές, πολύ λίγες κρύβουν τη μορφή. Στα ακέραια δεδομένα χρησιμοποιούνται σχετικές συχνότητες ανά ακέραιο, όχι συνεχής πυκνότητα.

4. Τι μπορούμε να αλλάξουμε σε κάθε δραστηριότητα

Αρχείο 1 — ροπές. Οι `mu0` και `sigma0` καθορίζουν κοινή θεωρητική μέση τιμή και τυπική απόκλιση για τους τέσσερις πληθυσμούς, με `sigma0 > 0`. Δοκιμάστε `mu0 <- 5`, `sigma0 <- 2`. Η λοξότητα και η κύρτωση δεν αλλάζουν με θετική αλλαγή κλίμακας.

Η `real_variable` επιλέγει `"eruptions"` ή `"waiting"` από το `faithful`. Οι `n_real` και `B_real` ελέγχουν μέγεθος και πλήθος υποδειγμάτων χωρίς επανάθεση, με $2 \le n_{\text{real}} \le 272$. Αν ένα υπόδειγμα είναι σταθερό, η λοξότητα και η κύρτωσή του δεν ορίζονται. Τέτοιες περιπτώσεις καταμετρώνται στον πίνακα.

Αρχείο 2 — διωνυμικό μοντέλο. Αλλάξτε `N0` και `p0`, με `N0` θετικό ακέραιο και $0 < \texttt{p0} < 1$. Το `N0` είναι το πλήθος δοκιμών ανά παρατήρηση, όχι το πλήθος παρατηρήσεων `n`. Δοκιμάστε $p_0 = 0.1, 0.5, 0.9$ και συγκρίνετε τη συχνότητα ύπαρξης εσωτερικής λύσης. Οι περιγραφές των εκτιμήσεων αφορούν μόνο τα δείγματα όπου αυτή υπάρχει· εμφανίζεται και το πλήθος τους.

Η `r0` αφορά μόνο την πρόσθετη αρνητική διωνυμική σύγκριση. Η `block_length` επιλέγει πόσα διαδοχικά έτη των `discoveries` περιλαμβάνει κάθε τμήμα. Με `compare_n <- TRUE`, οι `n_grid` και `B_grid` ελέγχουν μια πρόσθετη σύγκριση μεγεθών δείγματος.

Αρχείο 3 — ομοιόμορφο μοντέλο. Δοκιμάστε `theta <- 3.7`. Η θ είναι θετικός *πραγματικός* αριθμός, όχι υποχρεωτικά ακέραιος. Παρακολουθούμε τις

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

Η `classes` αφορά το αρχικό δείγμα και η `classes_B` τα ιστογράμματα των B εκτιμήσεων. Με `compare_n <- TRUE` εμφανίζεται η σύγκριση MSE για τις τιμές της `n_grid`. Τα σημεία είναι αποτελέσματα προσομοίωσης· οι γραμμές προκύπτουν από τους ακριβείς τύπους.

5. Πώς διαβάζουμε τις καμπύλες

Στα προσομοιωμένα δεδομένα γνωρίζουμε την κατανομή παραγωγής. Η θεωρητική πυκνότητα διακρίνεται από τις προσαρμογές με χρώμα, τύπο γραμμής και υπόμνημα. Στο $N(0,1)$ το μέγιστο της πυκνότητας είναι $1/\sqrt{2\pi} \simeq 0.399$. Δεν απαιτείται να συμπίπτει με την κορυφή μιας στήλης του ιστογράμματος: το ύψος της στήλης είναι σχετική συχνότητα διαιρεμένη με το πλάτος της κλάσης.

Στην ομοιόμορφη, κάθε προσαρμογή τελειώνει στο *δικό της εκτιμημένο άκρο*. Στα ιστογράμματα *εκτιμητριών*, η κατακόρυφη γραμμή στο θ είναι τιμή αναφοράς, όχι πυκνότητα. Στα πραγματικά δεδομένα δεν υπάρχει γνωστή «αληθινή καμπύλη». Η εμπειρική συνάρτηση κατανομής περιγράφει τα δεδομένα· η κανονική ή η ομοιόμορφη είναι υπόθεση προς διερεύνηση.

Για σκέψη. Αν δύο μοντέλα έχουν ίδιες μέσες τιμές και διασπορές, ποια άλλα χαρακτηριστικά και ποια γραφήματα θα χρησιμοποιούσατε για να τα ξεχωρίσετε; Αν δεν υπάρχουν ροπές, σκεφτείτε τη διάμεσο και αποστάσεις μεταξύ ποσοστημορίων.

Πηγές R: `?faithful`, `?discoveries`, `?hist`, `?nclass.FD`.