Analiza bivariată: cum alegem între un test parametric şi unul nonparametric

După cum am mai spus înainte, alegerea testului adecvat se face, pe de o parte, în funcţie de datele pe care vrem să le colectăm (tipurile de variabile), iar pe de alta în funcţie de scopul nostru.

Alegerea cea mai dificilă este în cazul variabilelor numerice (atunci când datele noastre reprezintă rezultatele unor măsurători), deoarece putem alege între două familii de teste, cele parametrice şi cele nonparametrice. Primele se bizuie pe presupunerea că datele provin dintr-o populaţie cu distribuţie normală, Gaussiană, iar testele cel mai des folosite sunt testul *t* (Student) şi analiza varianţei (ANOVA).

Testele pentru aplicarea cărora nu e nevoie de nici o presupunere asupra distribuţiei sunt numite teste nonparametrice, iar acestea aranjează în ordine valorile variabilei şi apoi face comparaţii între grupuri (vezi capitolul trecut). Testele Wilcoxon, Mann-Whitney şi Kruskall-Wallis sunt nonparametrice (se mai numesc şi *teste care nu depind de distribuţie*).

Uneori, este foarte simplu să alegem între un test parametric şi unul nonparametric. Vom alege, clar, un test parametric atunci când suntem siguri că eşantionul nostru se trage dintr-o populaţie care urmăreşte o distribuţie Gaussiană. Vom alege, dimpotrivă, un test nonparametric, în una din următoarele situaţii:

- Variabila este ordinală, iar populaţia este clar nonGaussiană. Exemple: anul de studiu la facultate, scorul Apgar al nou-născuţilor, un scor analog vizual pentru durere, scale Lickert, scoruri de calitatea vieţii compuse din adunarea mai multor *item*-uri etc.
- Unele valori ale variabilei sunt „înafara scalei”, adică prea mari sau prea mici pentru a putea fi măsurate. Chiar dacă populaţia este Gaussiană, este imposibil să foloseşti teste parametrice pentru astfel de date. În schimb, analiza lor cu teste nonparametrice este foarte simplă: acordăm acestor valori nişte valori arbitrare, fie foarte mici, fie foarte mari, iar cum testele nonparametrice ţine cont numai de ordine, nu şi de valoarea în sine, nu contează dacă nu ştim exact valorile.
- Variabila este cantitativă, numerică, dar ştim sigur că distribuţia nu este Gaussiană (există, în acest caz, posibilitatea de a obţine o distribuţie Gaussiană şi deci a face analiza cu teste parametrice după o transformare a valorilor – logaritm, reciprocă, radical etc).
- Variabila cantitativă are distribuţie Gaussiană, dar dispersia diferă mult între cele două grupuri.

De multe ori, însă, este dificil să-ţi dai seama dacă distribuţia valorilor unei anumite variabile este normală (Gaussiană):

- Dacă eşantionul este mare (cel puţin 100), nu trebuie decât să te uiţi la distribuţia valorilor şi se vede clar dacă este sau nu Gaussiană, pentru a şti ce fel de test să alegi. Dacă, însă, eşantionul este mic, este dificil să-ţi dai seama dacă distribuţia este Gaussiană sau nu prin inspecţie (histogramă), iar testele statistice (Kolmogorov-Smirnov) nu au destulă putere pentru a face diferenţa.
- Ne putem uita şi pe date similare din alte studii. Ceea ce contează este distribuţia valorilor **populaţiei**, nu eşantionului. Pentru a vedea dacă o populaţie este Gaussiană, trebuie să analizăm toate datele disponibile, nu numai datele din studiul la care lucrăm.

Când nu sunt siguri, unii utilizează un test parametric (deoarece nu sunt siguri că prezumţia de normalitate a fost violată), iar alţii aleg un test nonparametric (deoarece nu sunt siguri că distribuţia este Gaussiană). După părerea mea, a doua variantă este mai corectă.

Întrebarea este în ce măsură are importanţă ce fel de test, parametric sau nonparametric, alegem, şi sunt patru posibilităţi:

- Dacă eşantionul este mare (cel puţin 24/30 de date în fiecare grup, cifra diferă între cărţile de statistică), este mai uşor de spus dacă eşantionul provine dintr-o populaţie Gaussiană, dar nu are mare importanţă, putem folosi orice tip de test, deoarece rezultatul este acelaşi (rezultă acelaşi p).
- Dacă eşantionul este mic, distribuţia nu este Gaussiană şi utilizăm un test parametric, p-ul obţinut nu este corect.
- Dacă eşantionul este mic, distribuţia este Gaussiană şi utilizăm un test nonparametric, valorile p sunt mai mari (testele nonparametrice au putere statistică mult mai mică pe eşantioane mici).

Aşadar, în cazul eşantioanelor mari nu sunt probleme. Mai mult, dacă eşantionul este > 100 şi nu există valori extreme, cu influenţă disproporţionată asupra analizei statistice, se pot utiliza teste parametrice fără grijă.

Dilema apare atunci când eşantionul este mic: pe de o parte, este greu de spus dacă provin din populaţii cu distribuţie normală, şi tocmai aici acest lucru este foarte important, pentru că testele nonparametrice nu sunt puternice (pe aceleaşi date, cu unul parametric ai obţine un p mai mic decât cu unul nonparametric), iar cele parametrice nu sunt robuste.