IV. Tipurile de variabile şi testele statistice

Alegerea metodei de analiză statistică pentru o anumită problemă depinde de comparaţia pe care vrem să o facem şi de tipurile de variabile utilizate. Aşadar, pentru a alege testul potrivit trebuie să ne punem două întrebări: Ce fel de date am colectat? Care este scopul nostru? Aceeaşi analiză o facem şi citind un articol, pentru a vedea dacă testele utilizate de autorii acestuia sunt cele corecte.

Variabilele nominale sunt variabile sub formă de nume sau alte simboluri reprezentând categorii ce nu pot fi ordonate una în raport cu cealaltă, de exemplu numele, grupa sanguină, sexul, rasa, culoarea ochilor, diagnosticul etc. Atunci când o variabilă nominală nu poate lua decât două valori, ea este o variabilă dihotomică (binară, bimodală), cum ar fi sex masculin/feminin, mort/viu, fumător/nefumător, prezent/absent, normal/anormal, care a suferit efectul(*end-point*)/care nu l-a suferit etc. – variabile de tip DA/NU.

Variabilele ordinale sunt variabilele ce sunt clasificate în mai mult de două categorii şi la care există o ordine naturală între categorii (de la valoarea cea mai mică la cea mai mare) - de exemplu evoluţia bolii (agravat, staţionar, ameliorat), stadializări (insuficienţa cardiacă, TNM în cancer), scoruri etc. Chiar dacă unele variabile iau valori numerice, ele sunt considerate ordinale pentru că nu îndeplinesc condiţiile celor cantitative măsurabile (luând exemplul unui scor de calitate a vieţii, sau al unei scale analogice vizuale pentru durere, putem spune că un individ cu scorul 10 are o durere mai mare sau o calitate a vieţii mai bună decât un individ cu scorul 8, dar nu înseamnă că diferenţa dintre ei este aceeaşi cu aceea dintre un individ cu scorul 4 şi unul cu scorul 2; la fel, nu putem spune că un individ cu scorul 8 are o durere de două ori mai puternică sau o calitate a vieţii de două ori mai bună decât un individ având scorul 4; din acelaşi motiv, calcularea mediei nu are nici un sens).

Variabilele cantitative (măsurabile) pot fi *continue* (variabile cu un număr potenţial infinit de valori de-a lungul unui continuum: înălţimea, greutatea, TA, vârsta etc.) sau *discontinue (discrete)*(variabile descrise numai prin unităţi întregi ce nu pot fi măsurate în intervale mai mici decât unitatea: frevenţa cardiacă, numărul de copii etc.).

În privinţa variabilelor cantitative, este important de văzut dacă acestea au o distribuţie normală (simetrică, sub forma clopotului lui Gauss); în cazul acestei distribuţii, media este egală cu mediana şi cu modul, iar 95% dintre valorile pe care le poate lua variabila se află în intervalul media ± două deviaţii standard. Pentru a vedea dacă o variabilă are distribuţie normală, putem folosi orice program statistic şi verificăm cifric dacă media, mediana şi modul sunt foarte apropiate (ideal identice, dar în viaţa reală nu există ideal), iar media minus dublul deviaţiei standard nu trebuie să ia valori negative; sau reprezentăm variabila sub forma unei histograme şi vizual verificăm forma simetrică, de clopot.

Este important să ştim dacă distribuţia variabilei noastre este normală pentru că *numai variabilelor cantitative, cu distribuţie normală li se pot aplica testele statistice parametrice!* De asemenea, pentru a putea aplica teste statistice parametrice trebuie ca nici dispersia (deviaţia standard) celor două grupuri să nu difere foarte mult. Testele parametrice sunt testele care compară mediile şi deviaţiile standard ale grupurilor despre care vrem să dovedim că sunt identice sau dimpotrivă, diferite, ori media şi deviaţia standard nu au nici un sens dacă nu avem o distribuţie normală.

Să presupunem că vrem să vedem dacă TA este diferită la bărbaţii faţă de femeile din Bucureşti; pentru aceasta, ideal ar fi să extragem la întâmplare două eşantioane, unul de femei, altul de bărbaţi din Bucureşti şi să le măsurăm TA. Cum valorile TA vor avea, probabil, o distribuţie normală în cele două grupuri, atunci pentru a le compara putem folosi un test parametric, care este *testul t (Student)*. Dacă vrem să comparăm mai multe grupuri deodată (de exemplu vrem să vedem dacă TA a moldovenilor, ardelenilor sau regăţenilor diferă între ele), folosim analiza varianţei în sens unic (*one way ANOVA*), aplicând *testul F*, care ne va arăta dacă TA este diferită sau nu în cele trei regiuni istorice, fără a ne spune însă care este grupul care diferă de celelalte.

Testele nonparametrice se aplică pentru variabilele cantitative fără distribuţie normală (Figura IV.1) şi pentru variabilele ordinale. Să luăm un exemplu din revista Medicina Internă 2004, 1:57-59 (Dumitraşcu DL et. al), în care se compară stresul la pacienţii cu dispepsie funcţională şi la martori, pe baza unui chestionar. Pentru comparaţia scorurilor de stres s-a folosit testul t. În primul rând, în cazul scorurilor nu trebuie folosite teste parametrice, scorurile nefiind variabile cantitative (vezi explicaţia de la prezentarea variabilelor ordinale). Să ne imaginăm, totuşi, că scorurile de stres sunt variabile cantitative – ca să putem aplica testul t ar trebui ca distribuţia lor să fie normală, ori în tabelul 2 putem vedea scoruri (media şi deviaţia standard) de genul 0,89 şi 0,93; 7,5 şi 5,28; 1,25 şi 0,93; 0,54 şi 0,60; 0,94 şi 1,39 (şi încă altele), din care se observă clar că distribuţia nu este normală (dacă scădem din medie 2 deviaţii standard avem scoruri negative, care nu există în realitate), aşadar pentru comparaţie ar fi trebuit folosit un test nonparametric (*testul Mann-Whitney U*).

Aşadar, pentru variabilele cantitative care nu au o distribuţie normală şi pentru cele ordinale se folosesc testele *nonparametrice*. De exemplu, dacă vrem să demonstrăm că pacienţii cu insuficienţă cardiacă internaţi în spitalul X sunt mai gravi decât cei internaţi în spitalul Z, comparînd clasa NYHA de insuficienţă cardiacă între cele două grupuri.

Atunci când pacienţii sunt împerecheaţi, folosim testele statistice *împerecheate (paired)*, parametrice sau nonparametrice. Singura împerechere perfectă se realizează atunci când împerechem pacientul cu el însuşi, în comparaţiile înainte-după. De exemplu, comparăm TA, sau colesterolul unor participanţi la un studiu înainte de a începe tratamentul şi după o lună de tratament. Variabila (TA, colesterolul) fiind continuă şi cu o distribuţie probabil normală, vom folosi un test parametric, şi anume *testul t împerecheat.* Dacă variabila de comparat nu are o distribuţie normală (de exemplu valoarea creatininei la pacienţii cu insuficienţă renală) sau este o variabilă ordinală (stadializarea tumorii, sau clasa NYHA a insuficienţei cardiace, sau scorul durerii pe o scală analogică vizuală, înainte şi după un tratament), vom folosi un test nonparametric împerecheat, care este *testul Wilcoxon.*

Echivalentul nonparametric al ANOVA (testul F) este testul *Kruskal-Wallis*.

Testele nonparametrice nu ţin cont de valoarea efectivă a variabilei, ci de ordinea lor (*rank tests*) – care este valoarea cea mai mică, care este următoarea şi aşa mai departe...

În cazul variabilelor dihotomice (pentru compararea proporţiilor) se foloseşte *testul* X^2 sau variantele sale *Yates* şi mai ales *testul exact al lui Fisher* (atunci când în tabelul de contingenţă 2x2 avem într-una din căsuţe o valoare aşteptată mai mică de 5). De exemplu atunci când vrem să comparăm proporţia de pacienţi care a făcut infarct în grupul tratat cu statină cu proporţia de pacienţi care a făcut infarct în grupul tratat cu placebo. De remarcat că în studiile terapeutice, atunci când avem de-a face cu efecte surogat studiem variabile cantitative (TA, transaminaze, clasa NYHA, fracţia de ejecţie,

densitatea osoasa etc.), pe când în cazul efectelor serioase avem de-a face cu variabile dihotomice (pacientul a suferit sau nu infarctul de miocard, fracura, decesul etc.).
Când vrem să vedem cum (şi dacă) variază o variabilă cantitativă în funcţie de o altă variabilă cantitativă, aşadar vrem să vedem în ce măsură două variabile cantitative se *corelează*, calculăm *coeficientul de corelaţie al lui Pearson (r)*. De exemplu, putem vedea dacă vârsta se corelează cu VSH (adică VSH creşte odată cu vârsta).
Dacă variabilele cantitative nu au o distribuţie normală, sau sunt ordinale (de exemplu, corelaţia dintre fracţia de ejecţie şi clasa NYHA a insuficienţei cardiace stângi, sau dintre valoarea transaminazelor şi cea a scorului necroinflamator găsit la biopsia hepatică) utilizăm echivalentul nonparametric al coeficientului Pearson, care este *coeficientul de corelaţie Spearman*.
Dacă, în cazul a două variabile care se corelează, putem spune care variabilă o determină pe cealaltă şi/sau vrem să calculăm valoarea unei variabile ştiind-o pe cealaltă, utilizăm *regresia* lineară (de exemplu, ştiind valoarea ALAT, putem prezice scorul necro-inflamator de la biopsie, sau ştiind înălţimea prezicem valoarea VEMS, sau ştiind doza de captopril pe care o administrăm prezicem cu cât va scădea TA).
Variabilele cantitative pot fi transformate oricând în variabile ordinale sau dihotomice (de exemplu valorile colesterolului în quartile, sau în colesterol normal/crescut). În baza noastră de date este indicat să trecem (şi pentru aceasta să culegem) variabilele noastre ca atare, pentru că apoi putem să le transformăm oricând în ordinale sau dihotomice, pe când invers nu vom putea niciodată (de exemplu introducem în baza de date anemie DA/NU şi apoi descoperim că ar fi fost mai bine să avem chiar valorile hemoglobinei!).
Partea cea mai dificilă este alegerea între testele parametrice şi cele neparametrice. Alegem clar un test nonparametric în trei situaţii: 1.efectul este o variabilă ordinală şi populaţia este clar non-Gaussiană (de exemplu notele studenţilor, scorul Apgar, scala vizuală analogică pentru durere etc.); 2.efectul este o variabilă cantitativă şi suntem siguri că nu are o distribuţie gaussiană în populaţie (în acest caz o putem aduce la o distribuţie normală prin transformare: logaritmul, reciproca, rădăcina pătrată – din punct de vedere matematic este corect, mai puţin din punct de vedere biologic); şi 3.efectul este o variabilă cantitativă cu distribuţie gaussiană, dar dispersia (deviaţia standard) este mult diferită între grupurile de comparat.
Deseori alegerea este dificilă. Când avem cazuri puţine, este greu de spus dacă distribuţia este Gaussiană, iar testele speciale pentru verificarea normalităţii (Kolmogorov-Smirnov) au putere mică. De fapt, ceea ce contează este distribuţia la nivelul populaţiei, şi nu la nivelul eşantionului nostru, iar informaţii despre distribuţia valorilor unei variabile în populaţie trebuie căutate în literatură! (Este bine de reţinut că în natură, distribuţiile non-gaussiene sunt frecvente, iar acest fapt este valabil îndeosebi în cazul valorilor biologice).
Când nu ştim dacă distribuţia este normală, alegerea tipului de test depinde de mărimea eşantionului: dacă eşantionul este mare (cel puţin 24/30 de date în fiecare grup), este mai uşor de spus dacă eşantionul provine dintr-o populaţie Gaussiană, dar nu are mare importanţă, putem folosi orice tip de test, rezultatul va fi acelaşi. Problema apare dacă eşantionul este mic, când este greu de spus dacă populaţia este gaussiană, dar tocmai atunci este foarte important: testele nonparametrice nu sunt puternice, iar cele parametrice nu sunt robuste.

Analiza multivariabilă (sau multivariată) este o unealtă statistică prin care se determină contribuţia fiecăruia dintre mai mulţi factori la apariţia unui efect. De exemplu, există o mulţime de factori asociaţi cu apariţia bolii coronariene (fumatul, obezitatea, sedetarismul, diabetul, hipercolesterolemia, hipertensiunea) – numiţi *factori de risc, variabile independente, sau variabile explicative*. Analiza multivariabilă ne permite să determinăm contribuţia independentă a ficăruia dintre aceşti factori de risc la apariţia bolii coronariene (numită *efect* sau *variabilă dependentă).*

În studiile observaţionale, fiind mai mulţi factori de risc, nu ştim care dintre ei este adevărat sau în ce măsură asocierea aparentă dintre un factor de risc şi efect nu este datorată de fapt altora. Să presupunem că în baza noastră de date avem toate informaţiile şi variabilele privind pacienţii, şi vom testa dacă există o asociere între variabilele fumat şi boala coronariană, fără a ţine cont de vreo altă variabilă. Neexistând randomizare (nu putem pune pacienţii, prin tragere la sorţi, să fumeze sau nu), chiar dacă în analiza univariată (bivariată, după alţii) găsim o asociere între fumat şi apariţia coronaropatiei, aceasta reprezintă o dovadă prea slabă pentru a o considera cauzală. Poate că fumătorii fac mai degrabă coronaropatie pentru că sunt mai frecvent bărbaţi şi/sau sunt săraci şi/sau au mai degrabă un stil de viaţă nesănătos în alte privinţe, care sunt adevăraţii factori de risc. Cu alte cuvinte, relaţia dintre fumat şi coronaropatie poate fi *confundată* de aceste alte variabile.

Confuzia apare atunci când o asociere aparentă dintre un factor de risc şi un efect este afectată de relaţia unei a treia variabile cu factorul de risc şi cu efectul. Pentru ca o variabilă să fie un *factor de confuzie*, aceasta trebuie să se asocieze atât cu factorul de risc, cât şi cu efectul.

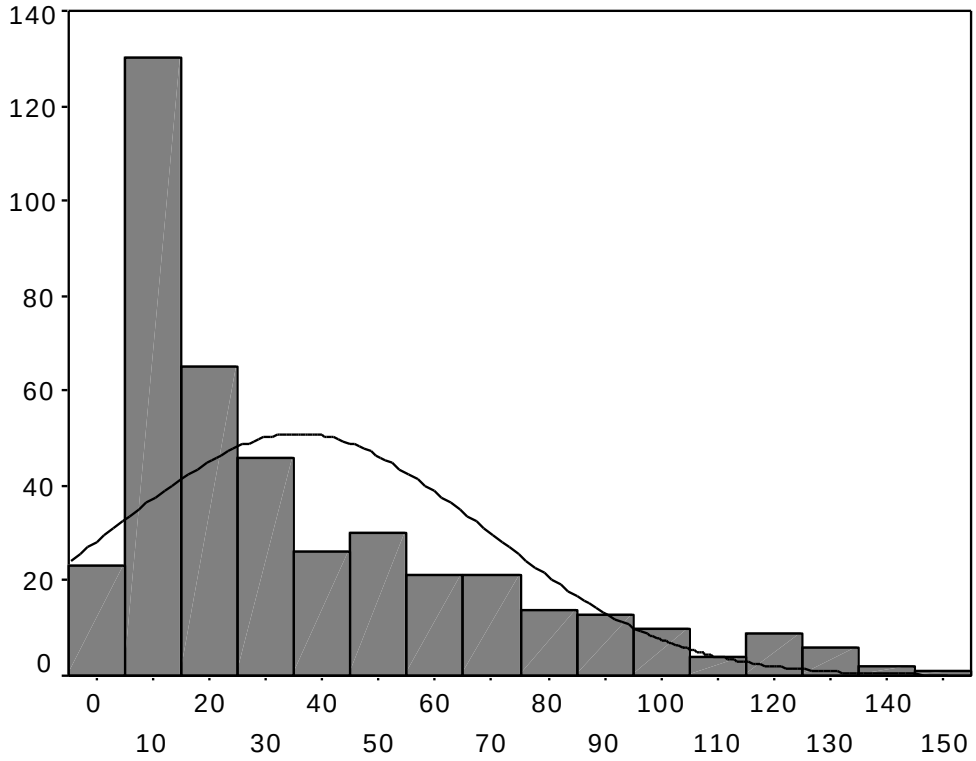
Sexul masculin şi sedentarismul pot fi factori de confuzie, deoarece sunt asociate atât cu fumatul, cât şi cu boala coronariană. Prin analiza multivariabilă, putem demonstra că şi după ajustarea pentru sexul masculin şi sedentarism, fumatul are o relaţie independentă cu boala coronariană. (De altfel acest cuvânt, "ajustat" care apare într-un articol ne spune de fiecare dată că la rezultatul prezentat s-a ajuns printr-o analiză multivariabilă.)

Să presupunem că în analiza univariată, boala coronariană este asociată cu consumul de cafea (riscul relativ=5, cu semnificaţie statistică). Bănuim, însă, că există un factor de confuzie, şi anume fumatul, care este asociat cu consumul de cafea (cei care beau cafea, în general fumează) şi cu efectul (am descoperit asocierea fumat-coronaropatie). Pentru a verifica această ipoteză, facem o analiză multivariată simplă, în care introducem ca variabile independente atât cafeaua, cât şi fumatul, şi vom vedea efectul fiecăreia dintre ele asupra apariţiei coronaropatiei. Dacă riscul de a face boală coronariană se menţine semnificativ statistic la cei care consumă cafea, înseamnă că aceasta reprezintă un factor de risc independent pentru boala coronariană. Dacă însă, în analiza multivariată, riscul dispare, înseamnă că într-adevăr, relaţia cafea-coronaropatie a fost confundată de relaţia adevărată, fumat-coronaropatie. Bineînţeles că în analiza multivariată vom găsi o relaţie semnificativă statistic fumat-coronaropatie.

Deşi teoretic se poate face distincţia între asocierea independentă şi confuzie, o variabilă poate avea în acelaşi timp un efect independent şi să fie un factor de confuzie: de exemplu sărăcia este un factor de confuzie între fumat şi coronaropatie (cei săraci fumează mai mult şi fac mai des boala coronariană), dar sărăcia are de asemenea şi un efect independent asupra apariţiei bolii coronariene (după ajustarea pentru fumat, colesterolemie şi alţi factori de risc, aceasta rămâne totuşi asociată semnificativ cu apariţia bolii).

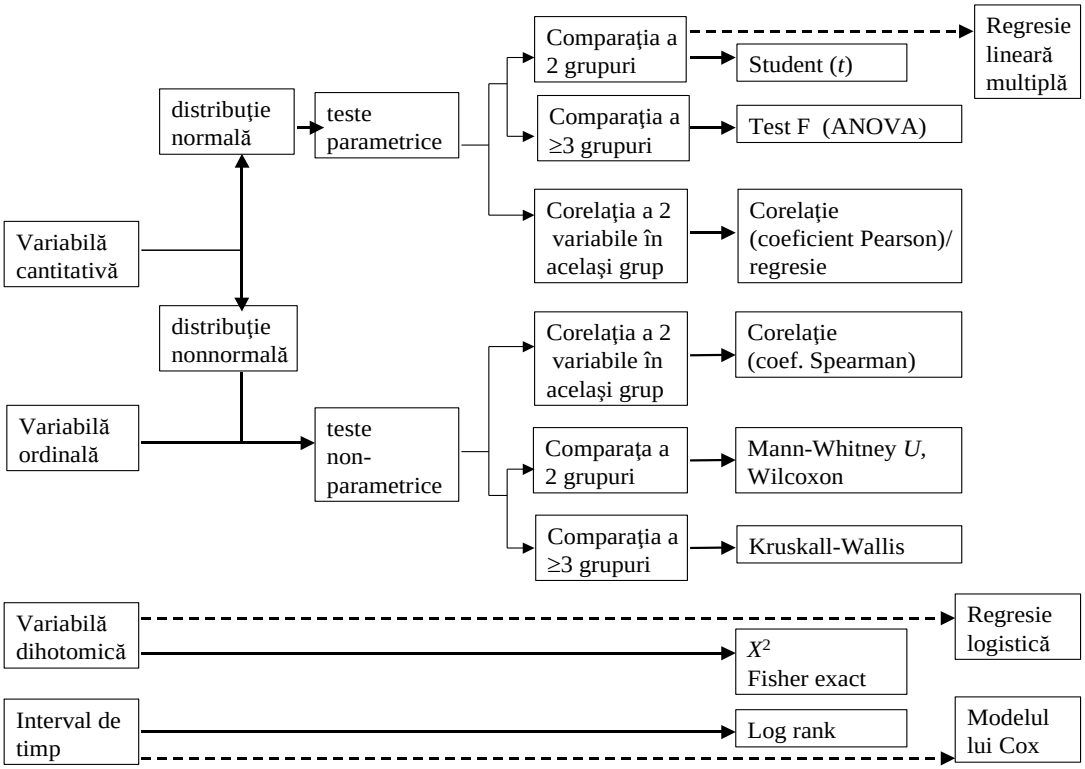
Înafara analizei multivariabile, pentru eliminarea confuziei mai poate fi utilizată *analiza stratificată,* prin care se cercetează rolul unui factor de risc în apariţia unui efect, în timp ce se ţine cealaltă variabilă constantă. Astfel, în exemplul în care cafeaua era asociată în analiza univariată coronaropatiei, putem reface această analiză separat, la fumători şi nefumători, şi dacă asocierea rămâne în picioare în cele două grupuri, înseamnă că efectul cafelei în apariţia coronaropatiei este independent de fumat; dimpotrivă, dacă asocierea cafea-boală dispare, înseamnă că fumatul a fost un factor de confuzie care a determinat apariţia unei false relaţii între cafea şi boala coronariană.
Ne putem folosi de stratificare atunci când există două sau trei variabile potenţiale factori de confuzie; atunci însă când acestea sunt mai multe, stratificarea ar crea zeci de grupuri în care investigatorul ar trebui să determine relaţia dintre variabile, iar numărul de pacienţi din fiecare grup ar fi din ce în ce mai mic, pe măsură ce progresăm cu stratificarea şi s-ar pierde puterea statistică.
Indiferent dacă folosim stratificarea sau analiza multivariată, nu trebuie să uităm că putem stratifica sau ajusta doar pentru variabilele pe care le cunoaştem, ori există mulţi factori de confuzie necunoscuţi şi deci nemăsuraţi, care ne pândesc la toate colţurile!
Tipurile de analiză multivariabilă sunt trei, în funcţie de variabila dependentă (efectul): atunci când variabila dependentă este continuă se utilizează **regresia lineară multiplă**, dacă aceasta este dihotomică se utilizează **regresia logistică**, iar când este reprezentată prin durata de timp până la apariţia unui eveniment (“supravieţuirea”), se foloseşte **analiza hazardului proporţional (modelul lui Cox)**.

Figura IV.1. Exemplu de distribuţie non-normală: distribuţia VSH într-un studiu (histograma) în comparaţie cu distribuţia normală (curba lui Gauss). Se observă asimetria distribuţiei VSH datorită existenţei unor pacienţi cu VSH foarte mare, care trag media spre dreapta, în timp ce mediana nu este influenţată.



VSH (mediana=26, media=35mm/h)

Figura IV.2. Algoritmul utilizării testelor statistice în funcţie de variabile (exemple în text).
(------ = analiză multivariabilă)



C Băicuş. Medicina bazată pe dovezi. Cum înţelegem studiile.	**Ed Medicală, 2007**