Computer Environmental Data Processing

Lect.dr. Giani GRĂDINARU
Catedra Statistică şi Previziune Economică, ASE Bucureşti

Usually, the individual values gathered present important variations from one unit to the other, a reason for which the data picked in a raw form won't allow knowing the manifestation form of the studied phenomena or the existing connections between the collectivity units. The recorded statistic data can be analyzed only to an extent, which allows their processing. The data processing is a passing stage from concrete primary data to typical values, to the synthetic indicators system corresponding to the studied phenomenon manifestation manner, including the operations through which the data is transformed into information.
Keywords: statistic *data, typical values, synthetic indicators.*

Tehnica de calcul modernă permite posibilităţi multiple de reprezentări grafice prin intermediul unor software-uri mixte (Excel®, Lotus®, QuattroPro®) sau specializate pentru analiza statistică a datelor (Statistica®, SPSS®, SYSTAT®), care pot fi utilizate cu succes pentru analiza datelor de mediu.

Analiza varianţei

Analiza dispersională, denumită şi ANOVA – ANalysis Of VAriance – reprezintă unul din procedeele de prelucrare statistică cele mai pertinente a datelor de observaţie. Metoda a fost pusă la punct de către R.A. Fisher, un matematician angajat în anul 1920 la staţiunea agricolă experimentală Rothamsted, pentru a sprijini activitatea de prelucrare şi interpretare a unui vast material de observaţie acumulat pe parcursul mai multor ani de experienţe agrotehnice. Doar în câţiva ani de lucru a pus la punct o serie de principii şi metode, nu doar de interpretare a rezultatelor, ci metodologii de programare, dirijare a experimentelor şi de interpretare statistică a rezultatelor.

În esenţă, problema pe care a trebuit R. Fisher să o rezolve şi care prin similitudine se poate regăsi într-o sumedenie de alte domenii de activităţi, s-ar putea descrie astfel: să se compare producţiile medii pe anumite suprafeţe de teren a unor soiuri diferite de cereale şi plante tehnice, suprafeţele fiind prelucrate în mod diferit: ca adâncime de arătură, cantităţi şi tipuri de îngrăşăminte aplicate, cantităţile şi periodicitatea udăturilor etc.

Deci, în fond, problema de soluţionat se reduce la a compara mediile caracteristicilor populaţiilor respective, de a testa omogenitatea mediilor. Componenta economică a unor astfel de procedee de experimentare constă în faptul că permite identificarea efectelor semnificative cu un efort experimental minim, deci cu un număr redus de măsuri.

Analiza dispersională permite testarea semnificaţiei relaţiei între două sau mai multe tipuri de clasificări, determinând importanţa factorilor respectivi asupra acestor relaţii. La baza analizei dispersionale se află segregarea dispersiei totale a unei serii de date în dispersiile componente, care pot fi atribuite diferiţilor factori implicaţi. Varianţa care poate fi atribuită unui factor este împărţită la varianţa estimată a eşantionului, acesta fiind o dispersie normală, dată de efectul variaţiilor de eşantionare asupra datelor din eşantion. Dacă în acest raport varianţa atribuită unui factor este mai mare decât varianţa estimată a eşantionului, şi anume într-o mărime superioară celei la care ne-am aştepta ca să provină numai din variaţiile de eşantionare, atunci i se recunoaşte acestui factor calitatea de a exercita o influenţă asupra datelor din eşantion.

Semnificaţia mărimii cu care variaţia unui factor depăşeşte variaţia estimată a eşantionului se determină interpolând valoarea acestui *raport F* în tabelul distribuţiilor de probabilitate F, stabilind astfel dacă valoarea calculată depăşeşte sau nu valoarea corespunzătoare a lui F, la nivelul de semnificaţie $\alpha = 0{,}05$ (în cazul studiilor de mediu) şi pentru fiecare mărime a gradelor de libertate.

Spre deosebire de testul χ^2, se lucrează cu

două serii de grade de libertate, datorită faptului că se iau în considerare raporturile pentru două dispersii independente, date de atributele de pe rândurile/coloanele tabelului de contingenţă şi fiecăreia dintre ele îi corespunde un număr diferit de grade de libertate. Valoarea lui F arată probabilitatea aleasă după care o valoare de F ori mai mare decât cea dată ne aşteptăm să apară ca urmare a variaţiilor de eşantionare întâmplătoare.
Cel mai mare avantaj al analizei dispersionale este dat de capacitatea acesteia de a localiza sursa diferenţelor semnificative la grupările combinate, întocmite după două, trei sau mai multe caracteristici. Astfel, se poate vorbi despre analiza dispersională unifactorială, bifactorială şi multifactorială.

Experimentele statistice

Metodele puse la punct de R. A. Fisher şi generalizările ulterioare s-au consacrat sub denumirea de planificarea sau programarea experimentelor. Această metodă şi-a dovedit utilitatea în obţinerea rapidă, cu un grad ridicat de certitudine şi în condiţii de cheltuieli avantajoase a unor informaţii necesare în fundamentarea activităţilor de mediu.
În general, se consideră că *experimentul* este acea metodă de cercetare prin care variaţia (modificarea) uneia sau mai multor variabile explicative este controlată de către cercetător, măsurându-se apoi efectul acesteia asupra variabilei/variabilelor rezultative.
Cercetătorul poate interveni în proces controlând în mod conştient modificarea uneia sau mai multor variabile presupuse a determina evoluţia variabilei rezultative, controlând variaţia altor variabile independente exogene care nu sunt supuse experimentării şi a căror influenţă se poate interfera cu cea a variabilei/variabilelor explicative, reducând performanţele de fidelitate ale modelului sau măsurând valorile variabilelor pe parcursul derulării experimentului, în vederea estimării efectului controlării efectului variabilei/variabilelor independente asupra variabilei/variabilelor rezultative. Se pleacă de la premisa că variabilele independente explicative, numite şi *factori experimentali* influenţează variabilele dependente iar *tratamentul* experimental se aplică numai lor.
Într-un experiment ideal, organizatorul experimentului are posibilitatea de a controla efectiv - de a menţine la un nivel constant - toate variabilele "din afară". Din acest motiv, în cadrul experimentelor în care nu se poate exercita un control "din afară" se va efectua un "control" de natură statistică, selecţionând aleator unităţile de observare. Acest lucru se poate efectua utilizând una din multiplele scheme de programare a experimentelor.
Unităţile de observare (produse, indivizi, utilaje, elemente de mediu, tratamente medicale etc.), care reprezintă obiectul investigaţiei şi de la care se culeg informaţiile se clasifică în *unităţi experimentale* şi *unităţi de control*. În timp ce primele formează grupul experimental asupra căruia se aplică "tratamentul" experimental, cele din urmă sunt doar observate şi măsurate pentru comparaţie cu unităţile "tratate".
În programarea experimentelor un rol important îl deţine noţiunea de *celulă*, care reprezintă un sistem de nivele ale factorilor şi noţiunea de *bloc*, care reprezintă un şir de celule constituit după un anumit criteriu de omogenitate. Similar studiului regresiei, variabila care influenţează rezultatele experimentale se va nota cu X, cu valorile $x_1,x_2,...,x_n$ iar variabilele rezultat cu Y, având valorile y_1, y_2, ..., y_n.
Datorită dificultăţilor de a produce exact condiţiile de lucru reale, apar erori de experimentare. Pentru a minimiza erorile sistematice se utilizează *randomizarea experimentului*, care presupune dispunerea întâmplătoare a "procedeelor de prelucrare" pe "obiectele experimentului". Randomizarea poate fi considerată ca o precauţie împotriva perturbaţiilor care pot apărea sau nu şi dacă apar pot avea sau nu efecte serioase[1].
Randomizarea implică utilizarea unui mecanism aleator: cercetătorul trebuie să fie sigur că fiecare tratament are o şansă egală de a fi alocat oricărei unităţi experimentale. Dacă într-un experiment un tratament apare de mai

[1] Isaic-Maniu A., Mitruţ C., Voineagu V., "*Statistica pentru managementul afacerilor*", Editura Economică, Bucureşti, 1995.

multe ori, înseamnă că se repetă, deci există o *repetiție*, prin intermediul acesteia reducându-se abaterea standard a mediei tratamentelor şi deci creşte precizia experimentelor.

Analiza în componente principale

Această metodă[2] este utilizată pentru descrierea datelor conținute de un tabel indivizi-caracteristici numerice: "p" caracteristici sunt măsurate pe "n" indivizi. Prin intermediul acestui tip de analiză, un ansamblu de date poate fi redus într-o formă compactă, dar care totuşi poate scoate în relief anumite structuri fundamentale ale datelor respective. Metoda permite evidențierea unor relații semnificative de interdependență, care nu ar putea fi cunoscute numai prin examinarea datelor de intrare. Scopul acestei analizei este de a reduce complexitatea, prin identificarea unui număr mic de factori ale căror caracteristici care stau la baza numeroaselor evaluări ale unui produs, utilaj sau element de mediu.

În cazul în care există doar două caracteristici x^1 şi x^2, datele pot fi prezentate uşor cu ajutorul geometriei plane: fiecare individ e_i va fi un punct de coordonate x_i^1 şi x_i^2 iar simpla vizualizare a alurii norului de puncte permite studierea intensității legăturii dintre x^1 şi x^2 precum şi stabilirea indivizilor sau grupurilor de indivizi care prezintă caracteristici apropiate. Dacă există trei caracteristici, studiul vizual va fi încă posibil dacă se recurge la geometria în spațiu. Dacă numărul caracteristicilor va fi mai mare sau egal cu patru, studiul vizual va deveni imposibil.

Analiza canonică

Analiza canonică a fost propusă în anul 1936 de către H. Hotelling, în lucrarea "Relations between two sets of variables" şi are un rol teoretic foarte important. Ea înglobează majoritatea metodelor de analiză: regresia multiplă, analiza dispersională, analiza corespondențelor, analiza discriminantă, acestea putând fi considerate cazuri particulare ale analizei canonice[3].

Deşi este disponibilă sub forma unor software-uri de specialitate (Statistica® de exemplu), ea nu este utilizată decât foarte puțin datorită dificultăților care apar în interpretarea şi utilizarea rezultatelor. Scopul analizei canonice îl constituie studierea relațiilor liniare existente între două grupe de *caracteristici cantitative* observate pe acelaşi eşantion. Într-o manieră foarte precisă se caută o combinație liniară a caracteristicilor primei grupe şi o combinație liniară a caracteristicilor celei de-a doua grupe, care să fie cât mai puternic corelate.

Analiza tipologică

Metodele de clasificare (sau taxonomie) au drept scop regruparea indivizilor într-un număr restrâns de clase (clustere/aglomerări) omogene. Clasele se obțin cu ajutorul unor algoritmi formalizați şi nu prin intermediul unor metode vizuale care se bazează pe intuiția analistului.

Clasificarea face parte din tehnicile de analiză a datelor care funcționează într-un cadru general, având un număr mic de ipoteze. Aranjarea trebuie făcută astfel încât indivizii care aparțin aceleiaşi clase să fie cât mai *asemănători* între ei prin valorile caracteristicilor lor (adică să fie *similari*), în timp ce indivizii care aparțin unor clase diferite să fie cât mai *diferiți* între ei (adică să fie *disimilari*).

Se disting două mari tipuri de metode de clasificare:

- *Metodele neierarhice*, care au drept rezultat partiționarea indivizilor într-un număr fix de clase.
- *Metodele ierarhice*, care au drept rezultat serii de partiții de clase din ce în ce mai diferite, de tipul celor cu care operează zoologia: specii, clase, familii, ordine etc.

În acest caz, tabelul de date analizate poate fi tabelul distanțelor sau disimilarităților între "n" indivizi sau tabelul coordonatelor indivizilor pe axele "p" (tabelul indivizi x caracteristici numerice sau coordonatele axelor în cazul caracteristicilor calitative).

[2] Bouroche J–M., Saporta G., *"L'analyse des données",* Presses Universitaires de France, Paris, 1980.

[3] Bouroche J–M., Saporta G., *op. Cit.*

Analiza factorială discriminantă

Analiza discriminantă pune în evidență legăturile existente între caracteristicile explicative cantitative şi o caracteristică ce urmează a fi explicată[4]. Metoda permite acest lucru prin intermediul vizualizării pe un plan factorial a caracteristicilor studiate. Totodată sunt prevăzute şi modalitățile caracteristicii explicate pornind de la valorile luate de caracteristicile explicative.

Se consideră un eşantion de indivizi asupra căruia se urmăreşte o caracteristică calitativă având "q" modalități. Fiecare individ va fi reperat printr-o singură modalitate a acestei caracteristici, astfel că s-a definit o parte a eşantionului de indivizi în "q" clase disjuncte. Pe acest eşantion vor fi măsurate cele "p" caracteristici cantitative. Problema la care trebuie să se răspundă următoarea: cele "q" clase diferă în ansamblul de caracteristici cantitative?

Pentru a obține răspunsul, se determină o nouă caracteristică prin intermediul unor combinații liniare ale vechilor caracteristici. Analiza discriminantă conduce la elaborarea unei reguli de decizie cu ajutorul căreia se stabileşte, în funcție de valorile variabilelor explicative, apartenența indivizilor din eşantion la o anumită clasă, pe baza acestor rezultate făcându-se previziuni cu privire la apartenența la clase a altor indivizi.

Sintetizând, se poate spune că analiza discriminantă urmăreşte:

- un scop descriptiv, constând în căutarea unui număr cât mai redus de variabile explicative, care să exprime cel mai bine separarea indivizilor în clase;
- un scop decizional, adică verificarea în ce măsură, un individ oarecare, încă negrupat, se aseamănă cu indivizii dintr-o anumită clasă şi, dacă această asemănare există, de a decide repartizarea sa în clasa respectivă.

Bibliografie

1. Bouroche J–M., Saporta G., *"L'analyse des données",* Presses Universitaires de France, Paris, 1980.
2. Ciucu G., Craiu V., "*Inferență statistică"*, Editura Didactică şi Pedagogică, Bucureşti, 1984.
3. Ciucu G., Craiu V.,*"Introducere în teoria probabilităților şi statistică matematică",* Editura Didactică şi Pedagogică, Bucureşti, 1971.
4. Colibabă D., - Metode statistice avansate de cercetare a pieței, Ed. ASE, Bucureşti, 2000.
5. Grădinaru G. – "Bazele statisticii mediului", Editura ASE, Bucureşti, 2004.
6. Grădinaru G., Colibabă D., Voineagu V. – „*Metode cantitative pentru analiza datelor de mediu*", Editura ASE, Bucureşti, 2003.
7. Isaic-Maniu A., Mitruț C., Voineagu V., "*Statistica pentru managementul afacerilor*", Editura Economică, Bucureşti, 1995.

[4] Bouroche J–M., Saporta G., *op. cit*.