1. Definiti si caracterizati principalele concepte utilizate în analiza datelor (populatie, esantion, observatie, variabile etc.)

Colectivitatea este un ansamblu de entităţi individuale, numite unităţi elemen-tare, unităţi observaţionale, obiecte, indivizi, cazuri etc., care au o existenţă reală, concretă, şi care au una sau mai multe proprietăţi empirice comune, numite *caracteristici* sau *atribute*.

Populaţia statistică este un ansamblu de entităţi informaţionale abstracte, virtuale, numite generic *observaţii*, constând din mulţimea tuturor *valorilor posibile*, efective sau virtuale, pe care le pot lua caracteristicile elementelor unei colec-tivităţi, ansamblu care reprezintă o *generalizare* a valorilor particulare ale respecti-velor caracteristici. O populaţie statistică are natura unei *mulţimi de valori*, de regulă *reale*, ale uneia sau mai multor variabile aleatoare, numite şi dimensiuni ale populaţiei statistice, în funcţie de numărul de caracteristici de interes ale colec-tivităţii pe care populaţia statistică o generalizează.

Eşantionul reprezintă o porţiune informaţională, o parte sau o submulţime a populaţiei statistice, respectiv acea parte care este asociată doar cu acele unităţi ale colectivităţii, de la care sunt colectate în mod efectiv datele de interes

Variabila reprezintă o *abstractizare* a mulţimii de valori posibile pe care le poate înregistra o caracteristică, de-a lungul tuturor unităţilor unei colectivităţi.

Parametrii sunt *mărimi numerice* cu valoare *constantă* şi *necunoscută*, specifice populaţiilor statistice şi modelelor econometrice, care exprimă caracteristici fundamentale şi de interes, a căror valoare nu poate fi obţinută pe calea observării directe.

2. Ce sunt variabilele si cum se clasifica acestea

Variabila reprezintă o *abstractizare* a mulţimii de valori posibile pe care le poate înregistra o caracteristică, de-a lungul tuturor unităţilor unei colectivităţi.

Variabilele se clasifica dupa urmatoarele criterii :
Ca şi caracteristicile unităţilor la care se referă, variabilele pot fi de împărţite, în funcţie de *natura valorilor* pe care acestea le iau, în două mari categorii: *variabile calitative* şi *variabile cantitative*.
Variabilele calitative sunt variabile care *diferă prin tip*, care *se referă la proprietăţi nenumerice* ale unităţilor elementare aparţinând unei colectivităţi şi care *nu pot fi exprimate sub o formă numerică semnificativă*. Valorile variabilelor de tip calitativ se numesc *alternative*, *variante*, *modalităţi* sau *categorii*, motiv pentru care variabilele calitative se mai numesc şi *variabile categoriale*. Exemple de varia-bile calitative : sexul, opţiunea cumpărătorului, opţiunea alegăto-rului, profesia, starea civilă, etc.
Variabilele cantitative sunt variabile care *diferă prin mărime*, care *se referă la proprietăţi numerice* ale unităţilor elementare dintr-o colectivitate şi care sunt exprimate în unităţi numerice de *lungime,* de *frecvenţă,* de *volum,* de *greutate,* de *valoare.* Exemple de variabile cantitative : preţul unui produs, cheltuielile lunare ale unei familii, salariul mediu lunar, venitul naţional, volumul fizic al producţiei etc.

Un alt criteriu de clasificare a variabilelor este cel al *naturii mulţimii* în care acestea pot lua valori. Din acest punct de vedere, variabilele se împart în două categorii: *variabile de tip discret* şi *variabile de tip continuu*.
Variabilele de tip discret sunt variabile *care pot lua valori într-o mulţime finită*, indiferent de natura calitativă sau cantitativă a acestora. Variabilele de tip discret pot să fie atât variabile calitative, cât şi variabile cantitative, cu condiţia ca în cazul ultimelor, numărul de valori posibile să fie finit. Exemple de variabile de tip discret : categoria de venit, nivelul de instruire, vârsta, numărul de salariaţi, numărul de cumpărători, numărul de piese defecte, numărul de firme falimentare, numărul de tranzacţii la bursă.
Variabilele de tip continuu sunt variabile numerice pentru care mulţimea de valori posibile este o mulțime de numere reale, care are puterea *continuului*. Exemple : masa monetară dintr-o economie, prețul unui bun economic, rata inflației, rata șomajului, cursul de schimb al monedei naționale etc.

Un alt criteriu de clasificare a variabilelor este cel reprezentat de *rolul acestora în contextul relaţiilor de cauzalitate* şi, implicit, în cadrul modelelor care descriu relaţii de acest fel. Din acest punct de vedere, variabilele se împart în trei categorii: *variabile endogene, variabile exogene* şi *variabile fictive*.
Variabilele endogene sunt variabile care exprimă fenomene de tip efect sau rezultat, considerate a se forma sub influenţa unor alte fenomene şi apar în cadrul modelelor econometrice ca variabile dependente.
Variabilele exogene sunt varia-bile care simbolizează fenomene de tip *cauze*, care, prin modul lor de manifestare, determină comportamentul unui fenomen de tip *efect*. Variabilele endogene si variabilele exogene pot fi atat de tip calitativ cat si de tip cantitativ.

Variabilele fictive sunt variabile artificiale, care sunt utilizate în con-strucția modelelor econometrice cu scopul de a asigura flexibilizarea modelelor sau cu scopul de a cuantifica influențe de tip sezonier. De regulă, variabilele fictive sunt variabile de tip binar, adică variabile care pot lua două valori posibile. În cazul în care varibilele fictive sunt incluse într-un model cu scopul de a descrie sezo-nalitatea, numărul acestora şi valorile lor posibile sunt determinate de numărul de perioade din intervalul de ciclitate

Variabilele pot fi clasificate şi în funcţie de *tipul scalelor* pe care sunt măsurate valorile acestor variabile. Din acest punct de vedere, există patru tipuri de variabile, respectiv ***variabile nominale*** **sau** ***categoriale*****,** ***variabile ordinale*****,** ***variabile de tip interval*** **şi** ***variabile de tip raport***, tipuri ce corespund scalelor nominală, ordinală, interval sau raport.

3. Ce este scala de masurare si care sunt principalele tipuri de scale de masurare utilizate în analiza datelor

O scala reprezinta un etalon corespunzator care stabileste modul dupa care sunt atribuite valori variabilelor. A defini o scala de masurare este echivalent cu :

- A stabili o multime de valori posibile ale variabilei, multime numita si spatiu de selectie
- A preciza regulile dupa care sunt atribuite simboluri pentru elementele unei relatii date, adic a define o structura asupra spatiului de selectie.

Scalele pot fi:

- Scale de tip non-metric:

Scala Nominala si Ordinala

- Scale de tip metric:

Interval si Raport (NOIR)

4. Definiti si caracterizati scala nominala si scala ordinala. Evidentiati operatiile posibile pe aceste tipuri de scale

2.2.3.2.1 Scala nominală

Scala nominală este o scală ***non-metrică***, pe baza căreia *valorile variabilelor sunt definite prin intermediul simbolurilor nenumerice*. Măsurarea variabilelor pe scala nominală este echivalentă cu procesul de *codificare a variabilelor*. Chiar în cazul în care pentru codificare sunt folosite numere, aceste numere sunt, totuşi, pur convenţionale.

Definiţie: ***Scala nominală*** este o scală *non-metrică*, prin intermediul căreia valorilor posibile ale caracteristicilor măsurate li se atribuie simboluri fără relevanţă numerică, în funcţie de natura acestor valori.

Scala nominală este utilizată pentru a măsura caracteristici ale căror valori sunt de natură *calitativă*, *necuantificabilă*. Valorile pe care pot să le ia caracteristicile de acest tip sunt cunoscute sub numele de ***categorii*** sau ***alternative***. Variabilele măsurate pe scala nominală se numesc ***variabile nominale*** şi sunt variabile a căror formă de exprimare este de tip atributiv şi care pot fi folosite numai pentru stabilirea apartenenţei la o anumită clasă a entităţii descrise prin intermediul variabilei.

O clasă specială a variabilelor de tip nominal o reprezintă ***variabilele binare***, care sunt variabile ce pot să ia doar două valori de tip nenumeric.

Variabilele de tip nominal sunt variabile *discrete* şi pot fi utilizate numai în scopuri de clasificare de tip calitativ, natura nenumerică a acestor variabile făcând imposibilă utilizarea lor pentru comparaţii, ierarhizări sau ordonări.

În cazul măsurării pe scala nominală, valorilor pe care pot să le ia caracteristicile supuse măsurării, respectiv categoriilor sau alternativelor, li se atribuie ***simboluri***, care sunt de natură nenumerică.

Pe scala nominală, două valori diferite ale caracteristicii măsurate sunt evidenţiate prin intermediul a două simboluri diferite. Elementele scalei nominale, "diviziunile" acesteia, sunt reprezentate de *simbolurile* atribuite valorilor caracteristicii studiate, sau, mai exact, de *categoriile* respectivei caracteristici. Scala nominală este reprezentată chiar de mulţimea acestor simboluri. De exemplu, mulţimile:

{"masculin","feminin"},
{"industrie","agricultură","construcţii",...},
{"muncitor","ţăran","intelectual"},

reprezintă scale de tip nominal utilizate pentru a măsura caracteristici cum ar fi sexul, domeniul de activitate, categoria socială, profesia.

Ceea ce este caracteristic scalei nominale este faptul că subiecţii studiaţi *nu pot fi comparaţi* din punct de vedere al valorii pe care o înregistrează la caracteristica măsurată pe această scală. Pe baza valorilor înregistrate pe scara nominală nu se poate afirma care subiect este "mai bine situat" din punct de vedere al caracteristicii studiate sau, cu atât mai puţin, "în ce măsură" un subiect este situat mai bine decât altul.

Tot pe această scală, caracteristicilor li se pot atribui şi numere, numai că aceste numere nu au sensul propriu-zis de *număr*, având practic aceeaşi semnificaţie ca şi simbolurile. Atât simbolurile propriu-zise, cât şi numerele cu rol de simbol, atribuite caracteristicilor pe această scală de măsurare, au numai rol de *clasificare* în anumite grupe a subiecţilor sau de *contorizare* a numărului de subiecţi din fiecare categorie, neputând fi folosite în nici un tip de calcul numeric. Prin intermediul valorilor măsurate pe scala nominală subiecţii se diferenţiază între ei doar din punct de vedere al *apartenenţei la o anumită clasă* sau al *apartenenţei la o anumită categorie*. Aceasta înseamnă că utilizarea scalei nominale pentru măsurarea caracteristicilor măsurabile pe această scală generează *clase* sau *categorii* de subiecţi.

Pentru caracteristicile măsurate pe scala nominală, poate fi calculat un număr limitat de indicatori statistici, care reprezintă, de fapt, *contorizări* ale simbolurilor apărute pe scala nominală. Aceşti indicatori sunt *modulul* şi *frecvenţa*. În cazul caracteristicilor măsurate pe scala nominală poate fi evidenţiată şi *distribuţia de frecvenţă*.

Într-o analiză de date, variabilele nominale pot fi reprezentate de o serie de variabile cum ar fi: *sexul*, *categoria socială*, *tipul familiei*, *profesia*, *marca unui produs* etc.

Unica transformare de tip invariant a scalei nominale este reprezentată de operaţia de ***recodificare***, această operaţie neafectând apartenenţa la o anumită clasă a valorilor măsurate pe acestă scală.

2.2.3.2.2 Scala ordinală

Scala ordinală este o scală ***non-metrică***, similară scalei nominale, adică o scală de codificare cu deosebirea că *pe această scală este posibilă* ***ordonarea valorilor variabilelor***. Această scală este folosită cu precădere pentru *măsurarea preferinţelor consumatorilor.*

Scala ordinală permite clasificarea valorilor unei variabile în funcţie de rangul acestora, însă *diferenţele între ranguri nu sunt relevante şi nu au sens*. Acest tip de scală nu dă posibilitatea stabilirii gradului în care caracteristicile a două entităţi distincte diferă între ele (mai mult, mai puţin).

Definiţie: ***Scala ordinală*** este o scală non-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor li se atribuie numere de ordine sau ranguri, în funcţie de poziţia acestor valori într-o ierarhie.

Variabilele măsurate pe această scală se numesc ***variabile ordinale***, sunt variabile calitative de tip discret şi nu pot fi exprimate sub o formă numerică reală. Ca exemple de variabile ordinale putem menţiona: *categoria de venit* (mic, mediu, mare), *nivelul studiilor* (elementare, medii, superioare), *preferinţa consumatorilor pentru un anumit produs* (foarte mare, mare, mică, foarte mică, deloc), *nivelul calitativ al unui produs sau serviciu* (inferior, mediu, superior), *starea economică* (recesiune, stagnare, expansiune) etc.

Scala ordinală este utilizată în cazul în care caracteristica subiecţilor supuşi analizei determină o diferenţiere a subiecţilor din punct de vedere al *poziţiei pe care fiecare dintre aceştia o ocupă într-o ierarhie, într-o ordonare*, adică în cazul în care caracteristica ia *valori de tip ordinal*. Valorile pe care pot să le ia caracteristicile măsurate pe scala ordinală sunt ***valori ordinale*** sau ***note***, cunoscute şi sub numele de ***ranguri***. Acestor valori li se atribuie fie *numere de ordine,* fie *simboluri* care evidenţiază o anumită ordine a valorilor caracteristicii.

Pe scala ordinală, două valori diferite ale unei caracteristici sunt evidenţiate prin intermediul a două *ranguri* diferite, adică prin intermediul a două poziţii diferite în cadrul ierarhiei. Elementele scalei ordinale, "diviziunile" acesteia, sunt reprezentate de *numerele* sau de *simbolurile* folosite pentru reprezentarea rangurilor, respectiv de poziţiile posibile în respectiva ordonare. Scala nominală este reprezentată chiar de mulţimea acestor numere sau simboluri.

Cu toate că valorile caracteristicilor de tip ordinal nu sunt numere propriu-zise, ele diferenţiază, totuşi, poziţia unui subiect în raport cu un alt subiect, "spun ceva" despre această poziţie. Valorile unei caracteristici măsurate pe scala ordinală permit doar *ordonarea* subiecţilor din punct de vedere al acestei caracteristici, determinând o ierarhizare a subiecţilor sau obiectelor.

Prin intermediul valorilor pe care le pot lua caracteristicile măsurate pe scala ordinală, indivizii se diferenţiază între ei doar din punct de vedere al ***rangului****, al locului pe care îl ocupă în ierarhia generată de scala ordinală*. Aceasta înseamnă că utilizarea scalei ordinale pentru măsurarea caracteristicilor măsurabile pe această scală generează *ierarhii, ordonări* ale subiecţilor.

Măsurarea pe scala ordinală *permite comparaţii* între subiecţi din punct de vedere al caracteristicii măsurate, dar aceste comparaţii se referă numai la modul în care un subiect "este situat" în raport cu altul, fără a se putea spune şi "în ce măsură" subiecţii diferă între ei după caracteristica respectivă. Diferenţele dintre două valori succesive de pe scala ordinală nu pot fi considerate ca fiind egale, ele nedeterminând o distanţare egală între indivizi, astfel încât să se poată afirma, de exemplu, că subiectul situat pe primul loc este "de trei ori mai bun" decât subiectul situat pe locul al treilea.

Pentru caracteristicile măsurate pe scala ordinală, pot fi calculaţi o serie de indicatori statistici cum ar fi: *modulul, mediana, coeficientul de corelaţie a rangurilor, frecvenţa*. De asemenea, pentru caracteristicile de tip ordinal se poate evidenţia şi *distribuţia de frecvenţă*. Este important să se facă, în acest context, precizarea că *media* şi *diferenţele* valorilor variabilelor ordinale *sunt nerelevante*, nu au sens informaţional şi nici sens logic.

Singura transformare invariantă a scalei ordinale este ***translaţia***, adică transformarea care păstrează ordinea valorilor unei variabile. Analitic, acest tip de transformare invariantă a scalei ordinale poate fi definit astfel:

$$y=a+x$$

unde a este o constantă, pozitivă sau negativă, care dă sensul şi mărimea translaţiei valorilor scalei ordinale, valori reprezentate de x.

---

5. Definiti si caracterizati scala ordinala si scala raport. Evidentiati operatiile posibile pe aceste tipuri de scale

Def: **Scala ordinala** este o scala non-metrica, prin intermediul careia valorilor posibile ale caracteristicilor li se atribuie numere de ordine sau ranguri, in functie de pozitia acestor valori intr-o ierarhie.

Caracteristici:

- ✓ Variabilele masurate pe aceasta scala se numesc **variabile ordinale**, sunt variabile calitative de tip discret si nu pot fi exprimate sub o forma numerica reala (exp: *categoria de venit, nivelului studiilor, preferinta consumatorilor pentru un anumit produs,* etc.).
- ✓ Masurarea pe scala ordinala **permite comparatii** intre subiecti din punct de vedere al caracteristicii masurate, dar aceste comparatii se refera numai la modul in care un subiect "este situat" in raport cu altul, fara a se putea spune si "in ce masura" subiectii difera intre ei dupa caracteristica respectiva.

- ✓ Singura transformare invarianta a scalei ordinale este **translatia**, adica transformarea care pastreaza ordinea valorilor unei variabile. Analitic, acest tip de transformare invarianta a scalei ordinale poate fi definit astfel: **y=a+x**
  unde *a* este o constanta, pozitiva sau negativa, care da sensul si marimea translatiei valorilor scalei ordinale, valori reprezentate de *x*.

Operatiunile posibile pe aceasta scala:

Pentru caracteristicile masurate pe scala ordinala, pot fi calculati o serie de indicatori statistici cum ar fi: **modulul, mediana, coeficientul de corelatie a rangurilor, frecventa.** De asemenea, se poate evidentia si **distributia de frecventa**. Este important sa se faca, in acest context, precizarea ca *media* si *diferentele valorilor variabilelor ordinale* sunt nerelevante, nu au sens informational si nici sens logic.

Def: **Scala raport** este o scala metrica, prin intermediul careia valorilor posibile pe care le pot lua caracteristicile masurate li se atribuie numere definite in raport cu o origine prestabilita.

Caracteristici:

- ✓ Originea scalei indica absenta proprietatii, caracteristicii. In plus fata de celelalte scale, pe aceasta scala este definit si **raportul valorilor**, adica se poate compara de cate ori o valoare este mai mare decat alta.
- ✓ Scala raport este invarianta pana la o transformare proportionala pozitiva, adica pana la transformarea: **y=ax**
- ✓ Variabilele masurate pe scala raport se numesc **variabile tip raport** si sunt variabile cantitative (exp: *pretul, venitul, varsta, salariul, profitul, volumul vanzarilor, numarul cumparatorilor,* etc).

Pe aceasta scala sunt permise **toate operatiile definite pentru variabilele numerice**.

6. Care sunt principalele moduri de reprezentare (matriciala) a informatiilor în analiza datelor. Definiti si exemplificati fiecare dintre aceste moduri

Principalele moduri de reprezentare a informatiilor in analiza datelor sunt: **matrici de observatii, matrici de contingenta si matrici de proximitate.**

Matrici de observatii

O matrice de observatii este un tablou rectangular in care *liniile* reprezinta *obiectele* supuse masuratorilor, iar *coloanele* reprezinta *caracteristicile* obiectelor. Elementele tabloului reprezinta valori inregistrate in procesul de masurare pentru caracteristicile obiectelor supuse masuratorilor. Aceste valori mai poarta si numele generic de *scoruri*. Matricile de observatii se mai numesc si matrici de tip **"obiecte×caracteristici".**

Pentru o analiza de date in care numarul obiectelor supuse analizei este *T*, iar numarul de caracteristici ale obiectelor este *n*, matricea de observatii are forma urmatoare:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & & & \\ x_{T1} & x_{T2} & \dots & x_{Tn} \end{pmatrix}$$

- unde un element x_{ij} reprezinta valoarea inregistrata pentru cea de-a j-a caracteristica a obiectului i.
- O linie i a matricii de observatii X defineste un obiect O_i si reprezinta valorile inregistrate de acest obiect la cele n caracteristici pe care le poseda.
- O coloana j a matricii de observatii X reprezinta valorile inregistrate de caracteristica j pe multimea tuturor celor T obiecte supuse analizei.

De regula, in analiza de date, fiecare *linie* a matricii de observatii X este numita *observatie* si fiecare *coloana* a acestei matrici este numita *variabila.*

In multe situatii, nu pot fi obtinute informatii despre toate caracteristicile tuturor obiectelor supuse analizei. In cazul in care datele ce definesc obiectele nu sunt complete, matricea de observatii definita mai sus poarta numele de *matrice de observatii cu valori omise*.

Matrici de contingenta

Sunt tablouri rectangulare de dimensiune m×n, utilizate pentru reprezentarea datelor referitoare la frecventele relative sau absolute inregistrate pe o multime de obiecte de valorile a doua variabile de tip discret, prima variabila, notata cu u, avand m valori posibile, iar cea de-a doua variabila, notata cu v, avand n valori posibile. *Liniile* unei matrici de contingenta reprezinta *valorile* posibile ale *primei* variabile discrete, iar *coloanele* acestei matrici reprezinta *valorile* posibile ale celei *de-a doua* variabile discrete. In analiza datelor, matricile de contingenta se mai numesc si matrici de tip **"modalitati×modalitati".**

Un element x_{ij} reprezinta frecventa, absoluta sau relativa, a obiectelor pentru care prima variabila ia valoarea u_i si cea de-a doua variabila ia valoarea v_j. Acest element arata la cate obiecte cele doua variabile analizate au simultan valorile u_i si v_j.

Matrici de proximitate

Sunt matrici patratice de dimensiune n×n, utilizate pentru reprezentarea datelor cu privire la similaritatea sau nesimilaritatea unor obiecte. Ordinul matricilor de proximitate este determinat de numarul obiectelor supuse studiului. Elementele unei matrici de proximitate reprezinta coeficienti de similaritate, coeficienti de nesimilaritate sau distante. Un element x_{ij} din aceasta matrice masoara gradul de proximitate dintre obiectul i si obiectul j.

Matricile de proximitate se mai numesc si matrici de tip **"obiecte×obiecte"** si sunt utilizate in problemele de clasificare cu ajutorul tehnicilor de tip cluster si in problemele de scalare multidimensionala.

7. Definiti principalii indicatori (unidimensionali) cu ajutorul carora este sintetizata tendinta centrala sau locatia sau pozitia (inclusiv relatii de calcul si proprietati). Aratati ca media este o sinteza optimala pentru o multime de observatii

1.Media

Media este un indicator care caracterizeaza un esantion (o populatie) din punctul de vedere al unei caracteristici studiate

Proprietati::

1. Media este indicator statistic cu cel mai mare grad de aplicabilitate practica.

2. Media se prezinta ca marime cu caracter abstract, în sensul ca valoarea medie - de cele mai multe ori - nu coincide cu niciuna dintre valorile individuale din care s-a calculat

3. Media este nivelul la care ar fi ajuns caracteristica înregistrata, daca, în toate cazurile, toti factorii esentiali si neesentiali ar fi actionat constant.

4. Pentru a asigura un continut real mediei calculate, valorile individuale din care se obtin trebuie sa fie cât mai apropiate, sa existe o omogenitate a colectivitatii. În cazul eterogenitatii colectivitatii, aceasta trebuie separata pe grupe calitative pentru care se calculeaza medii partiale.

5. În analiza statistica se calculeaza mai multe tipuri de medii:

- media aritmetica

- media armonica;

- media patratica;

- media geometrica;

- media cronologica.

Media se calculeaza simplu, adunând toate valorile dintr-un sir de date si împartind totalul la numarul de date:

$$M = \Sigma X / N$$

Unde:

X-sirul de date

N-numarul de date

Media este recomandata în cazul variabilelor numerice care îndeplinesc conditiile parametrice (distributie normala, omogenitate)

2.Mediana

Mediana este acel parametru care prin pozitia sa, se afla în mijlocul seriei de date. Ea reprezinta punctul central al seriei, deoarece la stânga si la dreapta ei se situeaza câte 50% din totalitatea datelor. Mediana coincide cu media în cazul unei distributii teoretice normale si se îndeparteaza mult de aceasta daca distributia este asimetrica

Locul medianei intr-o serie de *n* termini se calculeaxa dupa formula:

$M = \Sigma X / N$

Formula de calcul a medianei este:

$$Me = x_o + h\frac{\sum f/2 - \sum f_p}{\sum f_m}$$

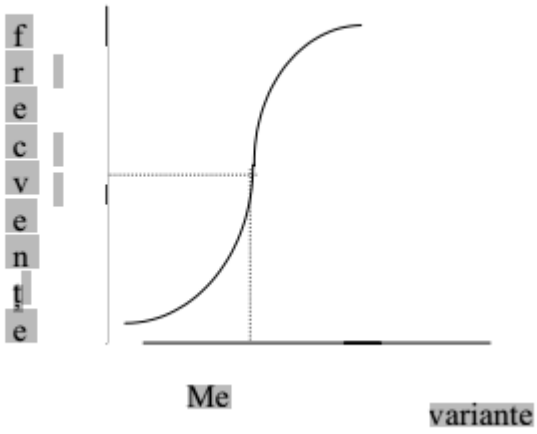
Unde: f_p - reprezintă frecvențele intervalelor precedente intervalului median

x_o - limita inferioară a intervalului median;

h - mărimea intervalului median;

f – frecvența intervalului median.

Reprezentarea grafică a madianei



Mediana se recomanda pentru cazurile în care nu sunt îndeplinite conditiile parametrice (distributii asimetrice, etrogenitate crescuta etc) si în cazul variabilelor de tip ordinal

3.Modulul

Modul (Mo) reprezintă valoarea caracteristicii care are frecvență maximă. Într-o serie de distribuție pe variante valoarea modului se determină direct. Pentru o serie de distribuție pe intervale valoarea modului trebuie calculată. Intervalul modal se consideră intervalul care are frecvența cea mai mare.

Formula de calcul al modului într-o serie de intervale este:

$$Mo = x_o + h\frac{f_2 - f_1}{(f_2 - f_1)+(f_2 - f_3)}$$

Unde: x_o - limita inferioară a intervalului modal

h – mărimea intervalului modal

f_1, f_2, f_3 - frecvențele respective ale intervalelor premodal, modal, postmodal.

Proprietati:

- **nu tine seama decât de masurile cele mai reprezentative;**
- **necesita ordonarea datelor**
- **corespunde unuia sau mai multor elemente ale seriei (în caz de frecvente egale).**

Modul este foarte util în cazul variabilelor de tip categorial (date calitative, nominale), deoarece nu putem calcula ceilalti parametrii centrali.

8. Definiti principalii indicatori (unidimensionali) cu ajutorul carora este sintetizata variabilitatea (inclusiv relatii de calcul si proprietati).

Varianta

Varianta reprezinta suma patratelor abaterilor valorilor individuale în raport cu media ce revine, în medie, pe fiecare valoare individuala, adica pe fiecare observatie efectuata asupra variabilei.

$$s_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(x_{ti} - \bar{x}_i\right)^2.$$

Varianta totala masoara variabilitatea ce caracterizeaza observatiile unei multimi de variabile si se defineste ca suma a variantelor individuale ale variabilelor.

$$V_T = \sum_{i=1}^{n} s_i^2$$

Varianta generalizata corespunzatoare spatiului observatiilor celor doua variabile considerate este data de relatia:

$$V_g = \left(\frac{1}{T-1} \|x^1\| \cdot \|x^2\| \cdot \sin\phi \right)^2$$

9. Definiti varianta simpla, varianta totala si varianta generalizata. Deduceti si interpretati varianta generalizata. Aratati ca varianta generalizata este egala cu determinatul matricii de covarianta

Varianta este *direct proportionala* cu *marimea variatiei* valorilor caracteristicii masurate sau cu *marimea informatiei* care este continuta de observatiile disponibile pentru analiza de date. *Varianta* variabilei , notata cu s_i^2, se determina cu ajutorul formulei urmatoare:

$$s_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} (x_{ti} - \bar{x}_i)^2$$

În mod concret, ***varianta*** reprezinta suma patratelor abaterilor valorilor individuale în raport cu media ce revine, *în medie*, pe fiecare valoare individuala, adica pe fiecare observatie efectuata asupra variabilei.

Varianta totala masoara variabilitatea ce caracterizeaza observatiile unei multimi de variabile Si se defineste ca suma a variantelor individuale ale variabilelor:

$$V_T = \sum_{i=1}^{n} s_i^2$$

O extindere importanta a conceptului de masura a variabilitatii o reprezinta ***varianta generalizata*** care masoara variabilitatea ce caracterizeaza observatiile multimii de variabile, atât din punct de vedere individual, cât si din punct de vedere al simultaneitatii, al interactivitatii informationale ce caracterizeaza variabilele. Pentru a da o interpretare intuitiva variantei generalizate, vom porni de la o constructie geometrica. În acest scop, vom considera ca variabilele $\mathbf{x_1}$ si $\mathbf{x_2}$ reprezinta doi vectori în spatiul observatiilor.

Exista o strânsa legatura între marimea unghiului format de cei doi vectori si corelatia dintre cele doua variabile. Aceasta consta în faptul ca, de fapt, *coeficientul de corelatie este cosinusul unghiului* dintre vectorii ce reprezinta cele doua variabile. Într-adevar, daca unghiul dintre cei doi vectori este zero, adica vectorii se suprapun, legatura perfecta existenta în aceasta situatie este evidentiata atât printr-o valoare a coeficientului de corelatie egala cu unitatea, cât si prin valoarea unitara a cosinusului unghiului respectiv. Invers, daca unghiul dintre vectori este de 90 de grade, adica vectorii sunt ortogonali, inexistenta legaturii specifice acestei situatii este evidentiata prin faptul ca atât coeficientul de corelatie, cât si cosinusul unghiului respective sunt egale cu zero. Cele trei situatii de corelare posibila a doua variabile x_1 si x_2 , ale caror observatii sunt reprezentate prin

intermediul vectorilor x^1 si x^2 , sunt evidentiate în graficele din figura 3.2.

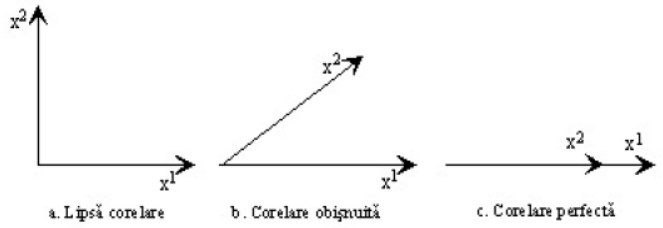


Figura 3.2: Situaţii posibile de corelare a două variabile reprezentate prin intermediul vectorilor $\mathbf{x^1}$ şi $\mathbf{x^2}$

Vom presupune că unghiul format de cei doi vectori este φ şi că cei doi vectori sunt scalaţi prin înmulţirea cu mărimea $1/\sqrt{T-1}$, adică cei doi vectori scalaţi au componentele de forma:

$$z_i^k = \frac{1}{\sqrt{T-1}} x_i^k, \qquad k=1,2;\ i=1,2,...,T.$$

Lungimea unui astfel de vector va fi:

$$\|z\| = \sqrt{\frac{x_1^2}{T-1} + \frac{x_2^2}{T-1} + ... + \frac{x_T^2}{T-1}} = \frac{1}{\sqrt{T-1}} \sqrt{x_1^2 + x_2^2 + ... + x_T^2} = \frac{1}{\sqrt{T-1}} \|x\|,$$

unde x_t reprezintă cea de-a t-a observaţie efectuată asupra variabilei x.

Dacă variabilele x_1 şi x_2 sunt variabile centrate, adică de medie nulă, atunci pătratul lungimii vectorilor z^1 şi z^2 reprezintă chiar varianţele celor două variabile:

$$\|z\|^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_i - \bar{x})^2 = \frac{1}{T-1}\sum_{t=1}^{T}x_i^2 = s_i^2.$$

În cazul *lipsei de corelaţie*, evidenţiată prin ortogonalitatea celor doi vectori, aria paralelogramului este *maximă*. Aceasta corespunde unei situaţii în care redundanţa informaţională aferentă observaţiilor efectuate asupra celor două variabile este *nulă*. În cazul în care *corelaţia este perfectă*, adică cei doi vectori sunt *coliniari*, aria paralelogramului este *minimă*. În această situaţie redundanţa informaţională corespunzătoare observaţiilor efectuate asupra celor două variabile, este *maximă*. În figura 3.3, este reprezentată aria paralelogramului având ca laturi vectorii ce definesc cele două variabile analizate.

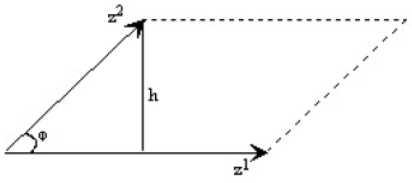


Figura 3.3: Interpretarea redundanţei informaţionale prin intermediul ariei paralelogramului

Definiţie: ***Varianţa generalizată*** corespunzătoare spaţiului observaţiilor celor două variabile considerate este dată de relaţia:

$$V_g = \left(\frac{1}{T-1}\|x^1\|\cdot\|x^2\|\cdot\sin\phi\right)^2$$

Se poate arăta că varianţa generalizată este reprezentată de *determinantul matricii de covarianţă* ce corespunde variabilelor supuse studiului, respectiv:

$$V_g = |S|.$$

10. Definiti principalii indicatori (unidimensionali) cu ajutorul carora sunt sintetizate legaturile (inclusiv relatii de calcul si proprietati)

Principalii indicatori undimensionali cu ajutorul carora sunt sintetizate legaturile dintre variabile sunt:

Covarianţa (s_{yx}) masoara sensul unei legaturi (directa, inversa).
Corelatia (r_{yx}) masoara forta unei legaturi (puternica, medie slaba).

$r_{yx} = s_{yx} / (s_x * s_y) \in [-1, 1]$

În cazul în care covarianţa are valoarea egală cu 1, se consideră că există o *perfectă asociere liniară directă* între cele două variabile, iar în cazul în care covarianţa are valoarea egală cu -1 se consideră că între cele două variabile există o *perfectă asociere liniară indirectă*. De asemenea, dacă valoarea covarianţei este *nulă*, se consideră că nu există asociere de tip liniar între cele două variabile. O consecinţă importantă a acestei ultime proprietăţi este reprezentată de faptul că, în cazul variabilelor standardizate, covarianţele sunt chiar coeficienţi de corelaţie Pearson.

Şi în cazul variabilelor standardizate, coeficienţii de corelaţie de tip Pearson pot fi exprimaţi prin intermediul produsului scalar şi lungimilor vectorilor corespunzători. Astfel, coeficientul de corelaţie dintre variabilele standardizate z şi w este dat de relaţia:

$$r_{zw} = \frac{s_{zw}}{s_z s_w} = \frac{1}{T-1} z^t w.$$

Rezultă că, în cazul variabilelor standardizate, coeficientul de corelaţie dintre două variabile este identic cu covarianţa şi este proporţional cu produsul scalar al vectorilor ce reprezintă observaţiile asupra variabilelor:

$$r_{zw} = \frac{1}{T-1} z^t w.$$

Măsura asocierii de tip liniar poate fi exprimată prin intermediul corelării *variaţiilor simultane* sau ***covariaţiilor*** a două caracteristici pe o mulţime de obiecte sau indivizi. Această măsură evidenţiază cum se corelează, cum se asociază valorile a două caracteristici la nivelul unei mulţimi de indivizi care posedă aceste caracteristici. Mărimea de bază utilizată pentru exprimarea variaţiilor simultane a două caracteristici este reprezentată de indicatorul cunoscut sub numele de ***covarianţă***. Pentru cazul a două variabile $\mathbf{x_i}$ şi $\mathbf{x_j}$, covarianţa acestora se calculează cu ajutorul formulei:

$$s_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j),$$

care, în cazul în care cele două variabile coincid, adică $\mathbf{x_i} = \mathbf{x_j}$, covarianţa *coincide* cu varianţa:

$$s_{ii} = \frac{1}{T-1} \sum_{t=1}^{T} (x_{ti} - \bar{x}_i)(x_{ti} - \bar{x}_i) = s_i^2.$$

Covarianţa este o măsură a variaţiei simultane a două variabile, ea fiind, în valoare absolută, cu atât mai mare cu cât valorile absolute ale variaţiilor celor două variabile în jurul mediei sunt mai apropiate ca magnitudine, evidenţiind o anumită proporţionalitate pe mulţimea subiecţilor studiaţi. Covarianţa este considerată a fi o expresie numerică a gradului de asociere a două caracteristici ca urmare a faptului că, în toate cazurile în care două variabile sunt semnificativ legate între ele, o variaţie într-un sens a uneia dintre ele va determina o variaţie *proporţională* de acelaşi sens (în cazul legăturii *directe*) sau de sens contrar (în cazul legăturii *inverse*) a celeilalte variabile.

În mod similar cu varianţa, şi în cazul exprimării covarianţei apare problema unor unităţi de măsură nefireşti, nenaturale. După modul în care este definită, covarianţa este exprimată în unităţi de măsură care sunt de fapt *produs* al unităţilor de măsură ale caracteristicilor considerate. Ca şi în cazul varianţei, există o dificultate şi mai mare în legătură cu măsura numită covarianţă. Aceasta constă în faptul că ea este o *mărime nescalată*. Deşi, în valoare absolută, covarianţa *are o margine inferioară*, reprezentată de valoarea zero şi care evidenţiază lipsa asocierii de tip liniar, ea nu este limitatată superior, *nu are o margine superioară*:

$$0 \le |s_{ij}| < \infty.$$

Ca urmare a acestei proprietăţi, apar dificultăţi legate de interpretarea magnitudinii covarianţei şi de utilizarea acesteia pentru efectuare de comparaţii.

O măsură scalată a gradului de asociere liniară între două variabile, care elimină unele deficienţe ale covarianţei ca indicator de măsurare a asocierii de tip liniar, o reprezintă ***coeficientul de corelaţie Pearson***. Pentru cazul a T observaţii existente cu privire la două variabile $\mathbf{x_i}$ **şi** $\mathbf{x_j}$, coeficientul de corelaţie Pearson este dat de relaţia:

11. Definiţi si interpretati corelatia si coeficientul de corelatie

Corelatia reprezinta tehnica statistică care măsoară şi descrie gradul de asociere lineară dintre două variabile cantitative continue normal distribuite.

Date

Obs	X	Y
A	1	1
B	1	3
C	3	2
D	4	5
E	6	4

Coeficientul de corelatie Pearson: reprezinta o masura scalata a gradului de asociere liniara între doua variabile, care elimina unele deficiente ale covariantei ca indicator de masurare a asocierii de tip liniar.

$$r_{ij} = \frac{s_{ij}}{s_i s_j} = \frac{\frac{1}{T-1}\sum_{t=1}^{T}(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(x_{ti} - \bar{x}_i)^2}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(x_{tj} - \bar{x}_j)^2}}.$$

Coeficientii de corelatie de tip Pearson pentru variabile centrate pot fi si ei exprimati în aceeasi maniera. Coeficientul de corelatie dintre variabilele centrate v si w este dat de relatia:

$$r_{vw} = \frac{s_{vw}}{s_v s_w} = \frac{\frac{1}{T-1}v^t w}{\frac{1}{\sqrt{T-1}}\|v\| \cdot \frac{1}{\sqrt{T-1}}\|w\|}.$$

Coeficientii de corelatie de tip Pearson pot fi exprimati prin intermediul produsului scalar si lungimilor vectorilor corespunzatori. Astfel, coeficientul de corelatie dintre variabilele standardizate z si w este dat de relatia:

$$r_{zw} = \frac{s_{zw}}{s_z s_w} = \frac{1}{T-1}z^t w.$$

12. Definiti datele de tip profil, de tip cronologic si de tip panel. Exemplificati fiecare dintre cele trei tipuri

Datele de tip profil reprezinta informatii obtinute prin masuratori de natura statica, efectuate asupra caracteristicilor unor unitati ale unei populatii, la acelasi moment de timp.

Exemplu: datele referitoare la salariul individual dintr-o luna al lucratorilor unei firme; datele referitoare la populatia medie a statelor lumii într-un anumit an; datele referitoare la rata inflatiei înregistrata de tarile lumii într-o anumita perioada; sexul cumparatorilor ce cumpara un anumit bun într-o anumita perioada; numarul mediu înregistrat de populatia judetelor unei tari într-un anumit an; volumul anual al vânzarilor unor marci de autoturisme, numarul voturilor înregistrate de partidele înscrise într-o campanie electorala.

Datele de tip serii de timp sau seriile cronologice reprezinta informatii obtinute prin masuratori de natura dinamica, efectuate asupra caracteristicilor unei unitati a unei populatii la momente sau în intervale succesive de timp.

Exemplu: datele reprezentate de seriile de timp se refera la evolutia în timp a starii unui individ, gospodarii, zone geografice, tari etc. Datele de acest tip pot fi date de tip interval sau date de tip moment. Datele de tip interval sunt datele care se refera la caracteristici care sunt marimi de tip stoc, în timp ce datele de tip moment sunt date care se refera la caracteristici care sunt marimi de tip flux. Si în acest caz, datele de tipul seriilor de timp pot fi privite ca reprezentând "sectiuni informationale", însa aceste sectiuni sunt de-a lungul axei timpului, de-a lungul evolutiei, adica sunt sectiuni longitudinale în raport cu axa timpului.

Datele de tip panel reprezinta informatii obtinute prin masuratori mixte, de natura statica si de natura dinamica, efectuate asupra caracteristicilor acelorasi unitati ale unei populatii la momente sau în intervale succesive de timp.

Exemplu: bugetele de familie, în contextul carora se fac înregistrari pe perioade de mai multi ani a veniturilor si cheltuielilor tuturor familiilor care alcatuiesc esantionul respectiv.

13. Definiti datele de tip observational si de tip experimental. Exemplificati fiecare categorie

Datele reprezinta expresii cantitative si calitative ale unor fenomene si procese din realitatea înconjuratoare.
Datele non-experimentale, care se mai numesc si *date observationale*, sunt datele obtinute prin "*observarea*" fenomenelor si proceselor în miscarea lor *naturala*, *libera*, fara impunerea unor restrictii, fara a se exercita un control de un anumit fel asupra fenomenelor si proceselor investigate.
Obtinerea datelor de tip non-experimental reprezinta rezultatul *observarii pasive*, *constatarii*.
Interventia observatorului, a celui care face masuratorile, este de tip *ex-post*, are loc dupa ce desfasurarea fenomenelor si proceselor reale a avut loc.
Datele de tip non-experimental sunt datele specifice domeniului economico-social, domeniu în care organizarea de experimente este fie dificila, fie imposibila.
Ex: observarea atitudinii consumatorilor cand apare un nou tip de produs pe raftul din supermarket.
Datele experimentale sunt datele obtinute prin organizarea unor *experimente de tip controlat*, desfasurate în conditii clare si prestabilite. Contextul obtinerii datelor de tip experimental este restrictionat, prin impunerea unor reguli specifice.
Datele experimentale sunt caracteristice doar unor domenii de cercetare, si anume acelor domenii în care pot fi organizate experimente specifice, necesare obtinerii acestor date. Experimentarea este posibila doar în anumite domenii ale cunoasterii, cum ar fi, de exemplu, domeniul *stiintelor naturale*: fizica, chimie, biologie etc.

Într-o alta modalitate de exprimare, se poate spune ca datele experimentale sunt *date de laborator*, prin "*laborator*" înțelegând aici o serie de conditii speciale, care se refera atât la o serie de restrictii si instrumente specifice de masurare, cât si la modalitatea de desfasurare a unor procese cauzale specifice.

14. Care sunt principalele tipuri de transformari preliminare ale datelor. Interpretati marimile rezultate în urma acestor transformari si mentionati proprietatile acestora

Analiza preliminara este o activitate anterioara, pregatitoare, a analizei propriu-zise a datelor, care are ca scop *initializarea* procesului de analiza. În cadrul acestei etape, informatiile primare disponibile sunt supuse unui proces de prelucrare în cadrul caruia are loc o filtrare a informatiilor din punct de vedere al semnificatiei si utilitatii pe care le au acestea în raport cu scopurile urmarite. Activitatea de analiza preliminara adatelor presupune utilizarea unei game variate de metode si tehnici statistico-matematice în scopul obtinerii unei sugestive caracterizari statistice a acestor informatii.

De obicei, înainte de a fi utilizate, datele brute sunt supuse la doua categorii de operatii preliminare: *operatii de rafinare* si *operatii de transformare*, fiind caracterizate prin:

- ***Centrarea observatiilor***
- ***Standardizarea observatiilor***

15. Definti principalele tipuri de matrici utilizate în analiza datelor (produse-încrucisate, covarianta, corelatie). Evidentiati relatiile de legatura dintre aceste tipuri de matrici

În principiu, datele primare sunt reprezentate în analiza de date sub trei forme matriciale principale: matrici de observatii,
matrici sau tabele de contingenta si matrici sau tabele de proximitate.

O matrice de observatii este un tablou rectangular în care liniile reprezinta obiectele supuse masuratorilor, iar coloanele
reprezinta caracteristicile obiectelor. Elementele tabloului reprezinta valori înregistrate în procesul de masurare pentru
caracteristicile obiectelor supuse masuratorilor. Aceste valori mai poarta si numele generic de scoruri. Matricile de observatii
se mai numesc si matrici de tip "obiecte×caracteristici".

Matrici de contingenta
Sunt tablouri rectangulare de dimensiune m×n, utilizate pentru reprezentarea datelor referitoare la frecventele relative sau
absolute înregistrate pe o multime de obiecte de valorile a doua variabile de tip discret, prima variabila, notati cu u, având m
valori posibile, iar cea de-a doua variabila, notati cu v, având n valori posibile. Liniile unei matrici de contingenta reprezinta

valorile posibile ale primei variabile discrete, iar coloanele acestei matrici reprezinta valorile posibile ale celei de-a doua
variabile discrete. În analiza datelor, matricile de contingenta se mai numesc si matrici de tip "modalitati×modalitati".

Matrici de proximitate
Sunt matrici patratice de dimensiune n×n, utilizate pentru reprezentarea datelor cu privire la similaritatea sau nesimilaritatea
unor obiecte. Ordinul matricilor de proximitate este determinat de numarul obiectelor supuse studiului.Elementele unei matrici de proximitate reprezinta coeficienti de similaritate, coeficienti de nesimilaritate sau distante

16. Ce este analiza componentelor principale. Evidentiati cinci categorii de probleme care pot fi solutionate cu ajutorul tehnicilor de analiza a componentelor principale

Analiza componentelor principale este o tehnica de analiza multidimensionala care are ca scop
descompunerea variabilitatii totale din spatiul cauzal initial sub forma unui numar redus de componente si fara ca aceasta
descompunere sa contina redundante informationale.

Analiza componentelor principale este o tehnica de analiza multidimensionala care are ca scop reducerea
dimensionalitatii spatiului cauzal initial, în conditiile unei pierderi informationale minime.

Analiza componentelor principale poate rezolva urmatoarele categorii de probleme:
- eliminarea redundantelor informationale;
- reducerea dimensionalitatii;
- compresia si restaurarea datelor;
- simplificarea modelelor matematice;
- selectarea variabilelor de influenta;

17. Interpretati logica analizei componentelor principale (inclusiv din punct de vedere geometric)

Cele mai interesante și mai utile aspecte ale analizei componentelor principale sunt în primul rând legate, nu de aparatul matematic pe care această analiză se bazează, ci de multiplele și nuanțatele interpretări posibile pe care aceasta le oferă.

Pentru a da o ilustrare intuitivă clară, bazată pe o interpretare geometrică simplificată, raționamentului primar care stă la baza analizei componentelor principale, vom dedica această parte, în exclusivitate, interpretărilor și exemplificărilor numerice.
În acest sens, vom considera contextul numeric oferit de exemplul următor, context care va servi ca referință pentru multe din interpretările și exemplificările ulterioare.

Exemplu:

Vom considera cazul unui număr de 10 obiecte sau observații, referitoare la două variabile, $\mathbf{X_1}$ și $\mathbf{X_2}$. Tabelul următor conține observațiile inițiale disponibile pentru cele două variabile, precum și valorile centrate ce corespund acestor observații.

Valorile observațiilor inițiale și centrate

Observația	**Valori inițiale**		**Valori centrale**	
	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_1^c}$	$\mathbf{X_2^c}$
O_1	7,0	10,0	0,6	-0,5
O_2	5,0	11,0	-1,4	0,5
O_3	10,0	15,0	3,6	4,5
O_4	2,0	5,0	-4,4	-5,5
O_5	5,0	10,0	-1,4	-0,5
O_6	6,0	13,0	-0,4	2,5
O_7	7,0	12,0	0,6	1,5
O_8	9,0	11,0	2,6	0,5
O_9	7,0	8,0	0,6	-2,5
O_{10}	6,0	10,0	-0,4	-0,5
Media	**6,4**	**10,5**	**0**	**0**
Varianța	**4,933**	**7,389**	**4,933**	**7,389**

Varianța individuală pentru fiecare din cele două variabile este 4,933, respectiv 7,389, iar varianța totală, corespunzătoare celor două variabile, $\mathbf{X_1}$ și $\mathbf{X_2}$ este 12,322:

$$\mathbf{S_{11} = 4,933;} \qquad \mathbf{S_{22} = 7,389;} \qquad \mathbf{V_T = 12,322.}$$

În aceste condiţii, se poate spune că *rolul informaţional* al celor două variabile este aproximativ *acelaşi*, că cele două variabile au aproximativ aceeaşi contribuţie la formarea variabilităţii totale ce caracterizează spaţiul cauzal iniţial. Prima variabilă are o contribuţie la formarea varianţei totale de 46,45%, iar cea de-a doua variabilă contribuie cu 53,55% la formarea varianţei totale:

$$\mathbf{\frac{S11}{VT} = 46,45\%;} \qquad \mathbf{\frac{S22}{VT} = 53,55\%.}$$

Pentru observaţiile din tabelul anterior, matricea produselor încrucişate, matricea de covarianţă şi matricea de corelaţie, corespunzătoare celor două variabile $\mathbf{X_1}$ și $\mathbf{X_2}$, sunt următoarele:

$$\mathbf{C} = \begin{pmatrix} \mathbf{454,0} & \mathbf{712,0} \\ \mathbf{712,0} & \mathbf{1169,0} \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} \mathbf{4,933} & \mathbf{4,444} \\ \mathbf{4,444} & \mathbf{7,389} \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} \mathbf{1,000} & \mathbf{0,736} \\ \mathbf{0,736} & \mathbf{1,000} \end{pmatrix}$$

În cazul observaţiilor centrale, matricea produselor încrucişate, matricea de covarianţă şi matricea de corelaţie sunt următoarele:

$$\mathbf{C} = \begin{pmatrix} 44,4 & 40,0 \\ 40,0 & 66,5 \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} 4,933 & 4,444 \\ 4,444 & 7,389 \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} 1,000 & 0,736 \\ 0,736 & 1,000 \end{pmatrix}$$

După cum se poate observa, în urma operației de centrare se modifică doar matricea produselor încrucișate, matricea de covarianță și matricea de corelație rămânând neschimbate. Matricea de corelație evidențiază faptul că cele două variabile *sunt corelate*, la nivelul unui coeficient de corelație de 0,736, adică:

$$\mathbf{r_{12} = r_{21} = 0,736}$$

Având în vedere intensitatea relativ ridicată a legăturii dintre cele două variabile originale, este de aşteptat ca aceste variabile să poată fi sintetiyate prin intermediul unei singure componente principale, în condiţiile unei pierderi informaţionale minime.

18. Definiti componentele principale si mentionati proprietatile acestora

Componentele principale sunt variabile vectoriale abstracte, definite sub forma unor combinații liniare de variabilele originale.
Proprietățile componentelor principale sunt:

- Sunt *necorelate* două câte două și suma pătratelor coeficienților care definesc combinația liniară ce corespunde unei componente principale este egală cu unitatea;
- Prima componentă principală este o *combinație liniară normalizată a cărei varianță este maximă*, cea de-a doua componentă principală este o combinație liniară necorelată cu prima componentă principală și care are o varianță cât mai mare posibilă, însă mai mică decât cea a primei componente etc.

19. Formulati modelul matematic al analizei componentelor principale, definiti si interpretati marimile definitorii ale acestuia

Analiza componentelor principale este o tehnica de analiza multidimensionala care are ca scop descompunerea variabilitatii totale din spatial cauzal initial sub forma unui numar redus de component si /reducerea dimensionalitatii spatiului cauzal initial, in conditiile unei pierdei informationale minime.

- Simplificarea structurii dependentei cauzale. Structura dependentei este reprezentata de multimea variabilelor cauzale supuse analizei. Prin simplificarea spatiului cauzal se intelege reducerea dimensionalitatii acestuia, astfel incat sa se obtina un spatiu cauzal de dimensiune mai mica si care sa permita o reprezentare mai simpla si mai sugestiva a obiectelor.
- Reducerea dimensionalitatii. La baza analizei componentelor principale sta ideea ca reprezentarea unitatilor în sistemul initial de coordonate, adica în sistemul pe ale carui axe sunt masurate caracteristicile originale ale unitatilor, nu este totdeauna cea mai potrivita, considerânduse ca poate exista o alta modalitate de reprezentare mai relevanta, mai eficienta din punct de vedere informational. Aceasta modalitate de reprezentare, mai avantajoasa din punct de vedereinformational, poate fi obtinuta considerând un *nou spatiu de reprezentare*, spatiu care defineste prin axele sale, în mod implicit, *noi caracteristici* ale obiectelor. Coordonatele obiectelor în acest nou spatiu sunt valorile înregistrate de obiecte la aceste noi caracteristici. În contextul simbolizarii cu ajutorul variabilelor, noile caracteristici sunt numite componente principale, iar valorileînregistrate de obiecte la aceste noi caracteristici sunt numite scoruri

20. Ilustrati modul de deducere a componentelor principale

În scopul formul|rii modelului matematic care st| la baza analizei componentelor principale, vom considera c| spaÛiul cauzal iniÛial supus investig|rii este determinat de un num|r de n variabile explicative notate x1,x2,...,xn . Aceste variabile simbolizeaz| caracteristici ale obiectelor supuse analizei, ceea ce înseamn| c| fiecare obiect este presupus a fi caracterizat de n variabile. Activitatea de determinare a componentelor principale poate fi descris| prin intermediul unei transform|ri de tipul urm|tor:

$\Psi: \vartheta^n -> \vartheta^k$

unde ϑ^n , ϑ^k sunt dou| spaÛii vectoriale reale, iar dimensiunea celui de-al doilea spaÛiu este mult mai mic| decât dimensiunea primului spaÛiu, respectiv k<n.

Prin intermediul transform|rii Ψ , un anumit obiect x, aparÛinând spaÛiului n-dimensional ϑ^n , este transformat într-un obiect w, aparÛinând spaÛiului k-dimensional ϑ^k . Transformarea vizeaz| atât modificarea coordonatelor obiectului, cât Õi reducerea num|rului acestor coordinate.

Rezolvarea problemei const| în determinarea matricii A, astfel încât un obiect w s| constituie o reprezentare cât mai bun| pentru obiectul x.

21. Definiti si justificati 3 dintre proprietatile componentelor principale

Analiza în componente principale

Iniţiată de Pearson (1901) şi dezvoltată de Hotelling (1933).

Tabloul de plecare R este oarecare: r_{ij} semnifică, în mod uzual, a *i*-a observaţie a unei variabile *j*. Variabilele pot fi eterogene în privinţa mediilor lor (de ex. unităţi de măsură diferite, ordine de mărime diferite etc.). Pentru a anula efectul eterogenităţii se efectuează transformarea

$x_{ij} = \frac{r_{ij} - \overline{r_{*j}}}{\sqrt{n}}$, unde $\overline{r_{*j}} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$ este media variabilei a j-a.

Analiza generală se va aplica tabloului X astfel obţinut, matricea X'X este matricea de covarianţă a variabilelor iniţiale.

Analiza în componente principale normate

Dacă variabilele sunt eterogene şi în dispersie, se vor norma valorile prin

$x_{ij} = \frac{r_{ij} - \overline{r_{*j}}}{s_j \sqrt{n}}$, unde s_j este abaterea standard pentru a j-a variabilă.

Analiza generală se va aplica tabloului X, cu observaţia că că matricea X'X implicată în calcule este tocmai matricea de corelaţie a variabilelor iniţiale.

Analiza în componente principale (normate) ACP/ACPN

Numele metodei provine din aceea că factorii (obţinuţi prin analiza generală) sunt numiţi şi componente principale.

Deşi pentru identificarea factorilor se aplică metoda generală asupra matricei de covarianţă (corelaţie) a variabilelor implicate, în continuare se prezintă şi o metodă alternativă, care poate oferi o viziune mai intuitivă asupra calculelor efectuate.

Se doreşte reducerea numărului de variabile dar cu păstrarea a cât mai mult (în limita posibilităţilor) din varianţa datelor iniţiale.

Pentru aceasta se introduce o nouă variabilă, *Z*, ca o combinaţie liniară a variabilelor iniţiale:

$$Z = a_1 x_1 + a_2 x_2 + \ldots + a_p x_p$$

unde $a_1,...,a_p$ sunt ponderi asociate variabilelor inițiale.

Observație. Ecuația precedentă este doar aparent similară unei ecuații de regresie, deoarece nu se cunosc valori observate pentru variabila *Z*, nu există termen liber şi nici erori (reziduuri).

Analiza în componente principale determină acele ponderi a_i care maximizează varianța variabilei *Z*. Cum varianța poate tinde la infinit pentru valori ale ponderilor convenabil alese, metoda determină doar ponderile supuse restricției că vectorul **a** este normalizat, adică $\sum_{i=1}^{p} a_i^2 = 1$. O dată calculate ponderile **a**, variabila *Z* este numită **prima componentă principală**.

Notând cu **C** matricea de covarianță (corelație) a variabilelor **X**, de fapt prin transformarea datelor din analiza în componente principale $\mathbf{C} = \mathbf{X}'\mathbf{X}$, rezultă că dispersia lui *Z* este $\mathbf{a}'\mathbf{C}\mathbf{a}$. Se doreşte maximizarea varianței lui Z cu restricția $\mathbf{a}'\mathbf{a} = 1$.Se ajunge astfel la problema generală:

$$\max \mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} \text{ cu restricția } \mathbf{a}'\mathbf{a} = 1$$

Prin metoda multiplicatorilor lui Lagrange se va căuta maximul funcției

$$F(\mathbf{a}) = \mathbf{a}'\mathbf{C}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1)$$

de unde rezultă, ca în metoda generală, că **a** este vector propriu al matricei **C** corespunzător valorii proprii λ şi $\mathbf{a}'\mathbf{C}\mathbf{a} = \lambda$. Deoarece $\mathrm{Var}(Z) = \mathbf{a}'\mathbf{C}\mathbf{a}$ rezultă $\mathrm{Var}(Z) = \lambda$, adică **a** este vectorul propriu care corespunde celei mai mari valori proprii λ.

A doua componentă principală este definită drept combinația liniară a variabilelor **X** cu următoarea cea mai mare varianță:

$$Z_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{p2}x_p$$

Se ajunge astfel la a doua valoare proprie ca mărime etc. De remarcat că a_{ij} reprezintă ponderea variabilei *i* în componenta principală cu numărul *j*.

O consecință a faptului că varianțele componentelor principale sunt valorile proprii iar ponderile (coeficienții combinațiilor liniare) sunt vectorii proprii este aceea că factorii obținuți (componentele principale) sunt necorelate între ele.

Astfel, din exprimarea matriceală $\mathbf{z} = \mathbf{A}\mathbf{x}$ a componentelor principale şi din faptul că matricea vectorilor proprii este ortogonală, $\mathbf{A}'\mathbf{A} = \mathbf{I}$, rezultă

$$A'z = A'Ax = Ix = x,$$

adică şi variabilele inițiale pot fi exprimate drept combinații liniare între componentele principale. Notând cu $\mathbf{C}_{zz}$ matricea de covariațe a componentelor principale, relația anterioară produce

$$\mathbf{C} = \mathbf{A}'\mathbf{C}_{zz}\mathbf{A}.$$

de unde, utilizând rezultatul cunoscut

$$\mathbf{C} = \mathbf{A}'\mathbf{\Lambda}\mathbf{A},$$

unde $\mathbf{\Lambda}$ este matricea diagonală a valorilor proprii, rezultă că $\mathbf{C}_{zz}$ este o matrice diagonală, adică toate componentele principale sunt necorelate între ele. Se observă astfel că prin trecerea la componentele principale se elimină redundanța din date.

22. Interpretati vectorii si valorile proprii ale matricii de covarianta

Matricea de covarianţă constituie una dintre cele mai frecvent utilizate matrici în analiza datelor, majoritatea tehnicilor de analiză a datelor presupunând calculul acestei matrici. Pentru situaţia în care numărul de variabile analizate este egal cu n, covarianţele dintre orice două variabile pot fi aranjate sub forma unei matrici pătrate şi simetrice, de dimensiune n×n, numită ***matrice de covarianţă***:

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1n} \\ s_{21} & s_2^2 & \dots & s_{2n} \\ \dots & & & \\ s_{n1} & s_{n2} & \dots & s_n^2 \end{pmatrix}, \text{ unde } s_i = \frac{1}{T-1}\|x^i - \bar{x}^i\|^2 \quad \text{şi} \quad s_{ij} = \frac{1}{T-1}(x^i - \bar{x}^i)^t \cdot (x^j - \bar{x}^j).$$

În condiţiile notaţiilor anterioare, matricea de covarianţă pentru variabilele originale poate fi scrisă cu ajutorul matricii produselor încrucişate pentru cazul variabilelor centrate, sub forma:

$$S = \frac{1}{T-1}\cdot C_c = \begin{pmatrix} \frac{1}{T-1}\|x^1 - \bar{x}^1\|^2 & \frac{1}{T-1}(x^1 - \bar{x}^1)^t \cdot (x^2 - \bar{x}^2) & \dots & \frac{1}{T-1}(x^1 - \bar{x}^1)^t (x^n - \bar{x}^n) \\ \frac{1}{T-1}(x^2 - \bar{x}^2)^t (x^1 - \bar{x}^1) & \frac{1}{T-1}\|x^2 - \bar{x}^2\|^2 & \dots & \frac{1}{T-1}(x^2 - \bar{x}^2)^t \cdot (x^n - \bar{x}^n) \\ & \dots & & \\ \frac{1}{T-1}(x^n - \bar{x}^n)^t \cdot (x^1 - \bar{x}^1) & \frac{1}{T-1}(x^n - \bar{x}^n)^t \cdot (x^2 - \bar{x}^2) & \dots & \frac{1}{T-1}\|x^n - \bar{x}^n\|^2 \end{pmatrix}.$$

23. Ce sunt scorurile principale si cum se determina acestea. De ce este necesara determinarea scorurilor principale

Scorurile principale → se mai numesc scoruri ale componentelor principale si reprezinta coordonatele obiectelor in noul spatiu, adica proiectiile obiectelor pe axele acestuia, sunt evaluari obiectelor in raport cu noile variabile. In analiza componentelor principale coordonatele obiectelor in spatial redus se mai numesc si scoruri principale ale obiectelor
Sunt determinate ca urmare a rotatie axelor cu un numar oarecare de grade are loc, in mod implicit, o modificare a coordonatelor initiale ale obiectelorle

Determinarea scorurilor principale e necesara intrucat acestea pot fi folosite in analiza ca substitute al observatiilor originale, simplificand in acest fel baza informationala initiala. In legatura cu aceasta problema, consideram ca este extreme de important sa facem precizarea ca scorurile principale sunt mai potrivite pentru a fi folosite in analize deoarece sunt mai putin afectate de erori, in comparative cu masuratorile originale. Faptul ca scoruri;e principale sunt mai robuste in raport cu perturbatiile introduse de erori, le face sa devina mai importante din punct de vedere informational, decat observatiile originale.

24. Ce este matricea factor (matricea de corelatie intre variabilele originale si componentele principale). Cum se calculeaza si cum se interpreteaza elementele sale

Matricea factor este o matrice uimportanta, utilizata in contextual analizei componentelor principale, ale carei elemente ofera premise pentru interpretari interesante. Legatura dintre vectorul variabilelor originale x si vectorul w al componentelor principale e data de relatia :
x=A W

O matrice importantă utilizată în contextul analizei componentelor principale, ale cărei elemente oferă premize pentru interpretări interesante, este matricea factor, pe care o vom defini în continuare.

În acest scop, vom presupune că cele n componente principale sunt reprezentate prin intermediul vectorului w, iar matricea de covarianţă a componentelor principale este matricea diagonală Λ. De asemenea, vom considera legătura dintre vectorul variabilelor originale şi vectorul componentelor principale ca fiind dată de relaţia:

$$\mathbf{x} = \mathbf{A}\cdot\mathbf{w},$$

unde A este matricea vectorilor proprii ai matricii de covarianţă Σ. Atunci matricea de covarianţă dintre vectorul x al variabilelor originale şi vectorul w al componentelor principale poate fi definită sub forma:

$$\mathrm{Cov}(\mathbf{x},\mathbf{w}) = \mathrm{E}(\mathbf{x}\cdot\mathbf{w}^t) = \mathrm{E}(\mathbf{A}\mathbf{w}\cdot\mathbf{w}^t) = \mathbf{A}\cdot\mathrm{E}(\mathbf{w}\mathbf{w}^t) = \mathbf{A}\cdot\Lambda\ ,$$

matricea de covarianţă a componentelor principale Λ fiind matricea diagonală formată din valorile proprii ale matricii de covarianţă Σ. Pe baza acestui rezultat, matricea de corelaţie dintre vectorii n-dimensionali $\mathbf{x}$ şi w poate fi definită sub forma:

$$\mathrm{Corr}(\mathbf{x},\mathbf{w}) = [\mathrm{Var}(\mathbf{x})]^{-\frac{1}{2}}\cdot\mathrm{Cov}(\mathbf{x},\mathbf{w})\cdot[\mathrm{Var}(\mathbf{w})]^{-\frac{1}{2}},$$

Elementele matricei factor se numesc intensitati ale factorilor ai au o interpretare interesanta din punct de vedere al legaturii dintre variabilele originale si componentele principale. Astfel elemental care se regaseste la intersectia liniei I cu coloana j in matricea factor, reprezinta coeficientul de corelatie dintre cea de-a I variabila standardizata cu cea de-a j component principala.

25. Detaliati modul în care pot fi interpretate componentele principale în termeni cu semnificatie concreta. Exemplificati

26. Criterii de alegere a numarului de componente principale

În mod concret şi într-o viziune simplificată, tehnica pe care se bazează analiza componentelor principale constă în *calculul proiecţiilor fiecărui punct din spaţiul iniţial*, determinat de variabilele originale supuse analizei, *pe axele unui nou spaţiu, a cărui dimensiune este semnificativ mai redusă*. În sens riguros, dar totuşi foarte general, analiza componentelor principale poate fi definită sub următoarea formă:

Definiţie: ***Analiza componentelor principale*** este o metodă de analiză multidimensională care are ca scop determinarea unor noi variabile, numite ***componente principale*** şi exprimate sub forma *combinaţiilor liniare de variabilele originale*, astfel încât aceste variabile noi să fie caracterizate de o variabilitate maximă.

În mod firesc, numărul de combinaţii liniare posibil a fi formate cu variabilele originale este extrem de mare. Deoarece, din punct de vedere al principiilor pe care se bazează activitatea de analiză a datelor, prezintă interes deosebit numai acele combinaţii liniare semnificative din punct de vedere informaţional, caracterizate de o mare variabilitate, este necesară o triere, o selectare a acestor combinaţii liniare. Efectuarea acestei selectări presupune definirea unui *criteriu* care să stea la baza deciziei de reţinere sau de eliminare a unei anumite combinaţii liniare.

În cadrul analizei componentelor principale acest criteriu este bazat pe magnitudinea varianţei fiecărei combinaţii liniare şi poate fi formulat astfel: *se elimină combinaţiile liniare cu varianţă mică, nesemnificativă şi se reţin pentru studiu acele combinaţii liniare care au o varianţă maximă*.

Reţinerea în analiză doar a acelor combinaţii liniare care au varianţă maximă are ca scop final realizarea unui eventual context în care variabilele originale să poată fi înlocuite cu un număr mult mai mic de astfel de combinaţii liniare, în condiţiile în care prin intermediul combinaţiilor liniare reţinute se preia o parte cât mai mare din variabilitatea conţinută în observaţiile variabilelor originale.

27. Ce este analiza factoriala si ce tipuri de probleme pot fi rezolvate cu ajutorul acesteia

Analiza factoriala este o analiza multivariata, care are ca scop sa explice corelatiile manifestate intre o serie de variabile, numite indicatori sau teste, prin intermediul unui numar mai mic de factori ordonati si necorelati, numiti factori comuni.

Analiza factoriala este folosita, in principal, in rezolvarea problemelor al caror scop este legat de:

- Studierea nivelelelor diferite de manifestare a interdependentelor dintre variabilele explicative, in special atunci cand numarul acestora este foarte mare.
- Detectarea unei structuri simplificate si clare a relatiilor de interdependent existenete intre variabilele explicative
- Obtinereaunei “cluster-izari”, unei clasificari a variabilelor explicative prin intermediul unor entitati numite factori, astfel incat variabilele apartinand unui anumit factor sa fie puternic intercorelate.
- Obtinerea unor informatii specifice, sub forma asa-numitilor factori, pe baza carora sa se poata face o interpretare sintetica a relatiilor de cauzalitate
- Verificarea unor ipoteze cu privire la existent unei structure factoriale particulare sau cu privire la existent unui anumit numar de factor comuni.
- Sintetizarea potentialului causal comun al mai multor variabile explicative sub forma unui numar cat mai redus de factori.

28. Structura generala a modelului de analiza factoriala

În formularea sa cea mai generala, formulare extrem de necesara pentru precizari cu caracter terminologic Õi pentru formularera unor ipoteze de natur| teoretic|, modelul analizei factoriale are la baz| dou| ipoteze fundamentale. Prima ipotez| se refer| la presupunerea c| c| nivelul sau valorile unui ansamblu de variabile aleatoare X1,X2,....,Xn se formeaz| ca rezultat exclusiv al influentei a trei categorii de factori:
• o multime formata din p factori comuni, f1,f2,...,fp , a caror influenta se consider| a se exercita asupra fiec|reia dintre cele n variabile considerate;
• o mulÛime formata din n factori unici, u1,u2,...un, a caror influenta se consider| a se exercita în mod individual, fiecare factor unic influentând una si numai una dintre variabilele considerate;
• o mulÛime de n factori reziduali e1,e2,....,en, a caror influenta se consider| a fi exercitat| tot în mod individual, fiecare factor rezidual influenÛând câte o singur| variabil|.
Din punct de vedere statistic, se consider| c| influenÛele semnificative, care trebuie reÛinute în analiz|, sunt cele exercitate de factorii comuni Õi unici, în timp ce influenÛele factorilor reziduali, se consider| a avea caracter accidental, nesemnificativ. La nivelul fiec|rei variabile, influenÛa factorului rezidual corespunz|tor poate fi considerat| a fi neglijabil| Õi este asimilabil| erorilor de m|surare. Din acest motiv, factorii reziduali se mai numesc Õi erori. În ceea ce priveÕte factorii comuni, exist| posibilitatea ca în cazul anumitor variabile influenÛa lor asupra acestor variabile s| fie neglijabil| sau chiar nul|, ceea ce înseamn| c| factorii respectivi pot fi eliminaÛi din lista factorilor pentru variabila respectiv|. În aceste condiÛii, este posibil ca schema de influenÛ| pentru anumite variabile s| conÛin| mai mulÛi factori comuni, iar pentru alte variabile mai puÛini. Num|rul de factori comuni cu influenÛ| semnificativ| asupra variabilei indicator determin| complexitatea variabilei indicator respective.
Faptul c| influenÛele considerate sunt structurate pe cele trei categorii de factori, determin| o anumit| structur| a modelului factorial general.
CoeficienÛii factorilor sunt cunoscuÛi sub numele de intensit|Ûiale factorilor. Prin magnitudinea sa coeficientul m|soar| intensitatea influenÛei exercitate de factorul corespunz|tor asupra nivelului variabilei indicator, iar prin semnul s|u m|soar| sensul influenÛei exercitate. DefiniÛie: Se numeÕte intensitate a unui factor comun fj în raport cu o variabil| indicator x1 marimea aj care arat| cu câte unit|Ûi se modific| nivelul variabilei indicator x1 , atunci când nivelul factorului fj creÕte cu o unitate.
Cea de-a doua ipotez| pe care se fundamenteaz| analiza factorial| este aceea c| în conÛinutul informaÛional al variabilelor aleatoare x1,x2,...,xn se reg|sesc informaÛii cu privire la factorii comuni Õi unici, ceea ce înseamn| c| ele pot fi folosite ca indicatori ai acestor factori, ca semnale informaÛionale generate de aceÕti factori.
Având în vedere c| la nivelul unei variabile indicator, nu se poate face, sub nici o form|, o distincÛie clar| între factorul unic Õi factorul rezidual, din motive legate de simplificare Õi de crearea posibilit|Ûilor de soluÛionare efectiv| a problemei de analiz| factorial|, factorul rezidual este neglijat sau, ceea ce înseamn| acelaÕi lucru, este unificat cu factorul unic.
În raport cu aceast| ultim| form| a modelului factorial se defineÕte conceptul de configuraÛie factorial|, concept care este folosit Õi într-un sens mai larg, cu referire la întregul set de ecuaÛii care defineÕte modelul.

29. Definiti si interpretati descompunerea variabilitatii în contextul analizei factoriale

În mod similar cu analiza componentelor principale, analiza factoriala îsi propune sa reexprime variabilitatea continuta în spatiul cauzal initial, într-o maniera diferentiata, în functie de rolul pe care îl au în formarea acesteia factorii comuni, pe de o parte, si factorii unici, pe de alta parte. Prin utilizarea tehnicile de analiza multidimensionala care au ca scop reducerea dimensionalitatii, variabilitatea spatiului cauzal n-dimensional, determinat de multimea de variabile indicator , este conservata într-o proportie, mai mare sau mai mica, prin intermediul variabilitatii induse de un numar mai redus de factori abstracti , , care sunt factorii comuni. Împreuna cu factorul unic, acesti factori determina un spatiu (p+1)-dimensional numit spatiul test sau spatiul factor. Variabilitatea ce caracterizeaza celor doua spatii implicate în analiza, spatiul original si spatiul test, este masurata prin intermediul variantei sau dispersiei.

Descompunerea variabilit|Ûii spaÛiului iniÛial

În mod similar cu analiza componentelor principale, analiza factorial| îÕi propune s| reexprime variabilitatea conÛinut| în spaÛiul cauzal iniÛial, într-o manier| diferenÛiat|, în funcÛie de rolul pe care îl au în formarea acesteia factorii comuni, pe de o parte, Õi factorii unici, pe de alt| parte.

În cadrul acestui paragraf, vom trata modul în care varianÛa unei variabile aleatoare poate fi descompus| în componente relevante din punct de vedere al interpret|rilor interdependenÛelor cauzale.

SpaÛiul factor Õi exprimarea conÛinutului sau informaÛional

Prin utilizarea tehnicile de analiza multidimensionala care au ca scop reducerea dimensionalitatii, variabilitatea sptiului cauzal n-dimensional, determinat de mulÛimea de variabile indicator $x_1, x_2, \ldots x_n$ este conservat| într-o proporÛie, mai mare sau mai mic|, prin intermediul variabilit|Ûii induse de un num|r mai redus de factori abstracÛi ,f1, f2,f3,…f4, (p<n) care sunt factorii comuni. Împreun| cu factorul unic, aceÕti factori determin| un spaÛiu (p+1)-dimensional numit spaÛiul test sau spaÛiul factor.

DefiniÛie: SpaÛiul test sau spaÛiul factor este un spaÛiu real, de dimensiune (p+1), ale c|rui axe sunt ortogonale dou| câte dou| Õi sunt reprezentate de factorii comuni **f1, f2,.. f4p** Õi de factorul unic **u** .

30. Ce sunt scorurile factor, cum se calculeaza si cum se interpreteaza acestea

O anumita observatie, corespunzatoare unui factor dat, este determinata sub forma unui scor corespunzator respectivului factor, scor format pe baza contributiei variabilelor originale. Exprimarea generica a scorurilor pentru un anumit factor în functie de variabilele originale este data de urmatoarea relatie:

$F_i=b_{i1}x_1+b_{i2}x_2+\ldots\ldots b_{in}x_n \qquad i=1,2,..p$

, und ij e b reprezinta coeficientii scorurilor factor si sunt elemente ale transpusei matricii factor F. Sub forma matriciala aceasta relatie poate fi scrisa astfel:

$f=F^tx$

În mod practic, exprimarea celor T observatii efectuate asupra variabilelor originale sub forma scorurilor factor, respectiv calculul concret al scorurilor factor, se bazeaza pe urmatoarele relatii:
$Z_{kj}=\sum b_{ki}x_{ji}$ K=1,2,….p J=1,2,…T

31. Metode de estimarea modelului factorial

Utilizarea analizei factoriale pentru dezvoltarea unor probleme specifice presupune și determinarea numărului de factori comuni ce vor fi reținuți în model.Există o serie de criterii care pot să orienteze utilizatorul atunci când ia o astfel de decizie.

1. **Criteriul procentului de acoperire:**

În general, alegerea numărului de factori care să fie incluși în modelul factorial depinde de proporția din variabilitatea comună conținută ân spațiul cauzal inițial pe care utilizatorul dorește s-o exprime prin intermediul unei siccesiuni de factori comuni. O estimație aproximativă a acestei proporții poate fi obținută cu ajutorul formulei :

$$p_k = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i$$

pk = Σ{i=1,k}λi/Σ{i=1,n}λi, unde k este numărul de factori reținuți, n este numărul variabilelor originale, iar λ este valarea proprie în raport cu care este definit factorul comun i.

2. **Criteriul lui Kaiser:**

Pate fi folosit atunci când analiza factorială este efectuată pe o matrice de corelație, adică atunci când se presupune că variabilele originale sunt standardizate. În conformitate cu acest criteriu, numărul de factori necesari a fi incluși într-un model de analiză factorială este egal cu numărul de valori proprii mai mari sau egale cu 1.

3. **Criteriul „granulozității”:**

După acest criteriu, numărul de factori ce vor reținuți în modelul de analiză factorială se stabilește pe baza unei analize grafice a valorilor proprii. Graficul se construiește luând în abscisă numărul de ordine al valorilor proprii, iar în ordonată valorile acestor valori proprii.
Valorile proprii fiind ordonate după magnitudinea lor descrescătoare, graficul are forma aproximativă a unei curbe de tipul exponențialei negative.
Numărul de factori ce se vor reține în model este determinat de punctul de pe grafic în dreapta căruia panta curbei devine neglijabilă, numărul de ordine al valorii proprii corespunzătoare acestui punct determinând număril de factori ce se vor reține.

32. Definiti recunoasterea formelor si exemplificati câteva dintre aplicatiile acesteia în domeniul economico-financiar.

Totalitatea tehnicilor de clasificare , adica de discriminare si de clusterizare, este cunoscuta si sub numele generic de teoria recunoasterii formelor.
Teoria recunoasterii formelor poate fi definita ca reprezentand totalitatea normelor, principiilor, metodelor si instrumentelor de analiza si decizie utilizate in scopul de a identifica apartenenta unor forme sau obiecte (unitati, fenomene, evenimente, actiuni, procese etc.) la anumite clase cu individualitate bine determinata.
In domeniul economico-social teoria recunoasterii formelor isi gaseste o larga utilizare mai ales în **procesul de analiza a datelor** si in activitatea de **predictie**. Problema clasificarii unei multimi de obiecte este o problema standard, frecvent întâlnita in investigarea socio-economica, iar abordarea ei presupune utilizarea metodelor si tehnicilor specifice teoriei recunoasterii formelor.

Tehnicile de recunoastere a formelor pot fi utilizate în domeniul economico-social pentru rezolvarea unor probleme cum ar fi: analiza datelor cu grad ridicat de eterogenitate, fundamentarea criteriilor de alegere a proiectelor de dezvoltare, clasificarea deciziilor în functie de impactul acestora asupra diverselor compartimente ale vietii economico sociale, detectarea unor perioade cu caracter specific din evolutia unor sisteme economice, stabilirea politicilor de creditare în domeniul financiar-bancar, evaluarea eficientei activitatilor de promovare a unor produse, determinarea perioadelor cele mai potrivite pentru vânzarea anumitor sortimente de marfuri, identificarea celor mai profitabile domenii de afaceri, clasificarea Si ierahizarea unor entitati economico-sociale etc.

33. Definiti principalele concepte ale recunoasterii formelor

Dintre numeroasele conceptele utilizate în teoria recunoasterii formelor, trei pot fi considerate ca fiind fundamentale si definitorii pentru esenta si scopurile teoriei recunoasterii formelor: *forma, clasa* si *clasificatorul*.
Forma sau obiectul este o entitate informationala individuala, caracterizata prin intermediul unui vector ndimensional, ale carui componente definesc valorile caracteristicilor acesteia, si care face obiectul procesului de clasificare sau de predictie.

Clasa, grupa sau *clusterul* reprezinta o entitate informationala *distincta* si cu *semnificatie concreta,* formata din totalitatea obiectelor ale caror caracteristici sunt identice sau difera foarte putin si care sunt semnificativ diferite de caracteristicile obiectelor din alte clase sau grupe.

Clasificatorul sau *criteriul de clasificare* reprezinta regula sau multimea de reguli pe baza carora obiectele care apartin multimii analizate sunt afectate sau atribuite unor clase sau grupe bine definite.

34. Formulati problema generala a clasificarii

Sub cea mai generala forma a sa, problema de clasificare poate fi formulata în termenii *teoriei deciziei*, iar metodele de clasificare pot fi definite sub forma unor *instrumente decizionale* specifice.

Vom descrie in continuare modul in care problema de clasificare poate fi definita ca o problema decizionala. In acest scop, vom presupune existenta unei populatii de forme sau de obiecte, notata cu **Ω** si definita sub forma:

$$\Omega = \{O_1, O_2, \ldots\ldots, O_M\}$$

unde **M** reprezinta numarul de unitati ale populatiei analizate.

Fiecare obiect care alcatuieste populatia **Ω** este definit prin intermediul unui numar de N caracteristici, pe care le vom nota cu $\chi_1, \chi_2, \ldots\ldots, \chi_n,$ si care se numesc variabile explicative. In acest fel, un obiect din populatia **Ω** poate fi reprezentat sub forma unui vector N-dimensional de forma:

$$\chi = (\chi_1, \chi_2, \ldots\ldots, \chi_n,)^t$$

35. Definiti sistemele de recunoastere controlata si necontrolata

Sistemele de recunoastere necontrolata - sunt sistemele în cadrul carora nu se dispune de informatii initiale referitoare la numarul de clase si la apartenenta formelor la anumite clase, construirea claselor facându-se progresiv, pe masura cresterii numarului de forme analizate, iar numarul de clase posibile fiind stabilit doar în faza finala a procesului de recunoastere. Caracteristica principala a sistemelor de recunoastere necontrolata a formelor consta în faptul ca nu se cunoaste apartenenta obiectelor analizate la o clasa sau alta. Aceasta înseamna ca, în mod implicit, nu se cunoaste cu precizie nici numarul de clase. În legatura cu aceasta ultima afirmatie, consideram ca este necesar sa facem urmatoarea precizare importanta: o serie de algoritmi de clasificare necontrolat, cum ar fi de exemplu algoritmii de partitionare, presupun fixarea apriorica a numarului de clase în care vor fi împartite obiectele analizate. Aceasta nu înseamna însa ca este cunoscut, în mod real, si numarul de clase, ci doar ca se face o presupunere cu privire la acest numar.

Sistemele de recunoastere controlata - sunt acele sisteme în cadrul carora se presupune existenta apriorica a unui numar dat de clase si a unui set de forme, numite prototipuri sau

referinte, a caror apartenenta la aceste clase este cunoscuta. Acest set de forme este reprezentat de esantionul de obiecte extrase din populatia supusa studiului, esantion cunoscut si sub numele de set de formare sau set de învatare. Sistemul de recunoastere controlata a formelor reprezinta totalitatea activitatilor si procedurilor care au ca scop deducerea unor criterii de partajare a unei populatii de entitati informationale (obiecte sau variabile), sub forma unui numar cunoscut de clase, pe baza cunoasterii caracteristicilor si a apartenentei elementelor unui esantion provenit din respectiva populatie.

36. Ce este analiza cluster, care sunt conceptele fundamentale ale acesteia si care sunt domeniile utilizarii ei

Analiza cluster - poate fi definita ca reprezentând o multime de principii, metode si algoritmi de clasificare, având ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.

Concepte fundamentale

Termenul de analiza cluster a fost utilizat pentru prima oara în anul 1939, de catre R. C. Tyron, în lucrarea “Cluster Analysis”. Acest termen este folosit în prezent ca nume generic pentru o multime variata de proceduri si algoritmi de clasificare de tip necontrolat.

Prin intermediul analizei cluster fiecare obiect din multimea analizata este atribuit unei singure clase, iar multimea claselor este o multime discreta si neordonabila. Clasele rezultate în urma utilizarii analizei cluster au o semnificatie concreta si generalizatoare, pe baza careia pot fi efectuate o serie de interpretari si pot fi formulate o serie de concluzii importante pentru procesul de cunoastere. Clasele sau grupele sub forma carora se structureaza multimile de obiecte se mai numesc si clustere. Un cluster este o submultime formata din obiecte similare, adica din obiecte care sunt suficient de asemanatoare între ele din punct de vedere al caracteristicilor care le definesc. Clusterul poate fi privit si ca reprezentând o regiune a unui spatiu multidimensional, caracterizata printr-o densitate relative mare de puncte sau de obiecte. De exemplu, în cazul aplicatiilor informatice, clusterul poate sa fie reprezentat de o submultime de documente de acelasi tip sau cu continut asemanator. Aceste documente pot fi programe sursa, pagini WEB, fisiere de tip text, fisiere HTML etc. Un astfel de document poate fi privit ca un punct dintr-un spatiu multidimensional, în care fiecare dimensiune a spatiului este asociata cu un anumit cuvânt.

Coordonatele care definesc pozitia unui document în acest spatiu sunt reprezentate de frecventele cu care apar diferitele cuvinte în cadrul documentului.

Domeniile utilizarii analizei cluster

Desi folosirea tehnicilor de analiza cluster nu este specifica doar pentru anumite domenii de activitate, totusi, utilizarea cea mai frecventa a acestora este întâlnita în domeniul marketingului, în investigatiile de natura psihosociala sau în evaluarile economico-sociale la nivel teritorial.

În domeniul marketingului, se detaseaza aplicatiile tehnicilor de analiza cluster în studierea comportamentului consumatorilor. Aceste aplicatii vizeaza evaluarea sanselor pe care poate sa le aiba lansarea unui produs nou, identificarea unor noi piete, modalitatile de segmentare a pietii sau identificarea pozitionarii pe piata a produselor diferitilor producatori. Posibilitatea de a deduce tipologii specifice pe multimea clientilor unei firme este deosebit de importanta pentru fundamentarea si stabilirea politicilor comerciale ale firmei.

În cazul determinarii pozitionarii pe piata a diferitelor marci ale unui produs, analiza cluster este folosita pentru a clasifica marcile de fabricatie, în functie de similitudinea sau disimilitudinea perceptiilor pe care le manifesta consumatorii fata de aceste marci. Pe baza modului în care se clasifica marcile si a caracteristicilor consumatorilor care îsi manifesta preferintele, un producator poate identifica marcile concurente si trasaturile specifice ale categoriilor de consumatori care prefera produsul acestui producator. De exemplu, marcile aflate în aceeasi clasa cu marca unui producator sunt marci concurente, deoarece ele se adreseaza aceluiasi segment de consumatori.

37. Definiti scopurile analizei cluster si descrieti tipul informatiilor utilizate în analiza cluster

Analiza cluster are ca ***scop*** cautarea si identificarea de clase, grupe sau clustere în cadrul unor multimi de obiecte sau forme, astfel încât elementele care apar in aceleiasi clase sa fie cât mai asemanatoare, iar elementele care apartin la clase diferite sa fie cât mai deosebite între ele. Altfel spus, analiza cluster este o modalitate de examinare a similaritatilor si disimilaritatilor dintre obiectele apartinând unei anumite multimi, în scopul gruparii acestor obiecte sub forma unor clase distincte între ele si omogene în interior.

Este o analiza explorativa, de tip multidimensional, care are ca scop gruparea unor entitati informationale, cu natura fizica sau abstracta, în clase sau clustere alcatuite din entitati informationale cu grad ridicat de similaritate.

Este definita ca un instrument care are ca scop reducerea unor multimi de obiecte, sau chiar de variabile, la un numar mai restrâns de entitati informationale, care sunt clasele sau clusterele.

Tipul de inf utilizate in analiza cluster:
Problema cea mai importanta a oricarui tip de analiza cluster este aceea a modului în care poate fi masurata proximitatea, respectiv gradul de apropiere sau gradul de departare, dintre obiecte si dintre clustere.
In general, masurarea gradului de proximitate dintre obiecte se face cu ajutorul a doua grupe de indicatori, cunoscute sub numele de indicatori de similaritate si indicatori de disimilaritate. Indicatorii de similaritate si indicatorii de disimilaritate pot fi utilizati atât în analizele cluster efectuate pe obiecte, cât si în analizele cluster efectuate pe variabile.
Cu cât valoarea unui indicator de similaritate este mai mare, cu atât obiectele sau variabilele pentru care acest indicator se evalueaza pot fi considerate a fi mai asemanatoare, respectiv mai apropiate. De asemenea, o valoare foarte mica a indicatorului de similaritate evidentiaza faptul ca cele doua obiecte sau cele doua variabile sunt mai departate între ele.
Indicatorii de disimilaritate sunt marimi numerice care exprima cât de deosebite sau cât de departate sunt doua obiecte sau doua variabile. Indicatorii de disimilaritate se mai numesc si indicatori sau coeficienti de deosebire sau de distantare a obiectelor sau variabilelor. Cu cât valoarea unui indicator de disimilaritate este mai mare, cu atât cele doua obiecte sau cele doua variabile pentru care se calculeaza sunt mai diferite, adica mai distantate între ele.
Cea mai importanta si cea mai utilizata categorie de indicatori de disimilaritate este reprezentata de indicatorii de tip distanta. De multe ori însa, conceptul de distanta este utilizat si pentru a desemna indicatori de similaritate, cu toate ca acestia exprima gradul de apropiere dintre doua entitati informationale.
Informaaiile utilizate, în ultima instanta, în analiza cluster sunt reprezentate sub forma unor matrici simetrice de tip obiecte×obiecte, numite, dupa caz, matrici de proximitate, matrici de similaritate, matrici de asociere, matrici de incidenta, matrici de disimilaritate sau matrici de distante. Atât liniile, cât si coloanele matricilor de acest fel se refera la obiectele analizate, astfel încât numarul lor este egal cu numarul de obiecte supuse analizei. Elementele acestor matrici sunt marimi numerice care exprima proximitatea dintre perechile de obiecte care eticheteaza rândurile si coloanele matricilor.
În cazul particular al clasificarii variabilelor, informatiile utilizate efectiv în analiza sunt reprezentate sub forma unor matrici de tipul variabile×variabile. Elementele acestor matrici sunt marimi numerice care exprima gradul de proximitate dintre perechile de variabile aflate în liniile si coloanele acestor matrici.

38. Definiti analiza cluster si aratati cum se clasifica metodele de analiza cluster

Definitie: Analiza cluster poate fi definita ca reprezentând o multime de principii, metode si algoritmi de clasificare, având ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.
Din punct de vedere strict teoretic, analiza cluster poate fi privita ca reprezentând o modalitate specifica de construire a uneia sau a mai multor partitii pe multimea obiectelor analizate. Orice partitie de acest fel defineste o solutie cluster, adica un anumit mod de grupare pe clase a obiectelor multimii supuse studiului. Din punct de vedere strict matematic, analiza cluster poate

fi privita ca o modalitate de alegere a celei mai adecvate partitii sau submultimi din cadrul familiei de parti a multimii de obiecte analizate.
Clasificare: Din punct de vedere al naturii lor, al modului de operare si al tipului de solutii pe care le furnizeaza, metodele de analiza cluster pot fi împartite în doua mari categorii: metode de tip ierarhic si metode de tip iterativ sau de partitionare.
Algoritmii sau metodele de tip ierarhic au ca scop producerea mai multor solutii cluster, solutii numite ierarhii cluster. Caracteristica principala a acestor algoritmi consta în faptul ca numarul de clustere nu este cunoscut aprioric.
Exista doua categorii de algoritmi de clasificare ierarhica: algoritmi de agregare si algoritmi de dezagregare.
Algoritmii de clasificare ierarhica furnizeaza mai multe solutii, de tip multinivel, care se numesc ierarhii cluster si care difera între ele prin numarul de clustere pe care le includ si prin gradul de agregare al clusterelor.
Algorimii sau metodele de **tip iterativ** au ca scop producerea unei structuri cluster formata dintr-o singura solutie cluster. O astfel de structura cluster se numeste structura cluster uninivel si contine o singura cluster, care include un numar fixat de clustere. În cazul metodelor de clasificare prin partitionare, numarul de clustere este cunoscut aprioric.
În functie de natura criteriului utilizat în procesul propriu-zis de clasificare, metodele de analiza cluster pot fi împartite în doua categorii: ***metode euristice si metode algoritmice***
Metodele euristice includ procedurile de clasificare dezvoltatea pe baza unei anumite euristici. O euristica este o modalitate intuitiva de solutionare a unei anumite probleme particulare. Euristicile reprezinta seturi de reguli sau de recomandari cu caracter general, deduse pe baza unor rationamente teoretice sau pe baza unor observatii statistice. Prin natura lor, metodele de clasificare ierarhica sunt metode euristice.
Metodele algoritmice includ procedurile de clasificare de tip formal, bazate pe existenta unui anumit algoritm de solutionare a problemei. Un algoritm este o multime de finita si complet definita de operatii, paÕi sau proceduri, a caror executie determina obtinerea unui anumit rezultat sau a unei anumite solutii. Orice algoritm se compune din trei parti esentiale: initializarea, procedura sau schema iterativa si criteriul de oprire.

Analiza cluster de tip ierarhic
metoda de clasificare bazata pe gruparea obiectelor pe baza de agregare succesiva în clase din ce în ce mai largi de obiecte sau de dezagregare succesiva în clase din ce în ce mai mici.
Se imparte in:- Metode de clasificare ierarhica prin agregare (care se imparte in Metoda agregarii simple, Metoda agregarii complete, Metoda agregarii medii, Metoda centroidului, Metoda lui Ward,

- Metode de divizare: numite si metode de tip descendent, sunt analoage cu metodele aglomerative, cu deosebirea ca derularea acestora se desfasoara într-o maniera inversa. Ca si în cazul metodelor de agregare, solutiile obtinute cu ajutorul metodelor divizative sunt ierarhii de clustere, care pot fi reprezentate prin intermediul arborilor cluster sau dendrogramelor.

Metoda agregarii simple este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care distanta dintre cei mai apropiati vecini este cea mai mica, în comparative cu alte perechi de clustere.

Metoda agregarii complete este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care distanta dintre cei mai departati vecini este cea mai mica, în comparative cu alte perechi de clustere.

Metoda agregarii medii este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care distanta medie dintre toate perechile formate cu obiecte din cele doua clustere este cea mai mica, în comparatie cu alte perechi de clustere.

Metoda centroidului este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care distanta dintre centroizii celor doua clustere este cea mai mica, în comparative cu alte perechi de clustere.

Metoda lui Ward este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care suma patratelor abaterilor la nivelul clusterului rezultat din comasare este cea mai mica, în comparatie cu alte perechi de clustere.

Algoritmi de partitionare: includ o serie de metode de analiza cluster, cu mult mai performante decât metodele de clasificare ierarhica. Dintre cei mai importanti algoritmi de partitionare, mentinem: algoritmul celor K-medii si algoritmul celor K-medoizi.

39. Definiti conceptul de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme

Distanta reprezinta unul dintre cele mai importante si mai frecvent utilizate concepte din domeniul analizei datelor. În acelasi timp, distanta constituie una dintre cele mai relevante modalitati de sumarizare a informatiilor manipulate în analiza datelor, mai ales în situatiile în care sunt investigate interdependentele dintre fenomene si procese. Ca marime, distanta se calculeaza pentru a evalua apropierea sau departarea dintre obiectele sau caracteristicile care se supun studiului, pentru a masura gradul de similitudine sau nesimilitudine dintre acestea, din punct de vedere al caracteristicilor studiate. Definirea si interpretarea conceptului de distanta presupune, în mod implicit, existenta unui spatiu în raport cu care are loc nu numai definirea, ci si evaluarea numerica a distantei. Spatiul în care este posibil a fi definit ca o distanta se numeste spatiu metric si poate fi spatiul variabilelor sau spatiul observatiilor.

Corespunzator celor doua modalitati de reprezentare, în spatiul variabilelor si în spatiul observatiilor, distanta poate fi utilizata pentru a evalua apropierea sau departarea dintre puncte ale unui spatiu multidimensional, puncte ce pot reprezenta atât obiecte, cât si caracteristici.

În functie de modul în care distanta este evaluata, adica în functie de modul în care se evalueaza gradul de departare sau apropiere dintre doua obiecte, exista mai multe tipuri importante de distante: distanta euclidiana, distanta statistica, distanta standadizata, distanta Mahalanobis etc.

40. Formulati criteriul general al clasificarii si aratati cum se evalueaza variabilitatea

inter si intra cluster (cazul uni-dimens)

Criteriu general de clasificare: Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o variabilitate minima în interiorul claselor si o variabilitate maxima între clase

41. Formulati criteriul general al clasificarii si aratati cum se evalueaza variabilitatea inter si intra cluster (cazul n-dimens)

Pentru evaluarea variabilitatii inter si intra cluster se utilizeaza metoda lui Ward.
Fie clusterul AB, clusterul obtinut din combinarea clusterului A cu clusterul B, atunci suma distantelor **inter cluster** (a vectorilor) este:

$$\mathrm{SSE}_A = \sum_{i=1}^{n_A} (\mathbf{y}_i - \bar{\mathbf{y}}_A)'(\mathbf{y}_i - \bar{\mathbf{y}}_A),$$

$$\mathrm{SSE}_B = \sum_{i=1}^{n_B} (\mathbf{y}_i - \bar{\mathbf{y}}_B)'(\mathbf{y}_i - \bar{\mathbf{y}}_B),$$

$$\mathrm{SSE}_{AB} = \sum_{i=1}^{n_{AB}} (\mathbf{y}_i - \bar{\mathbf{y}}_{AB})'(\mathbf{y}_i - \bar{\mathbf{y}}_{AB}),$$

Unde $\mathbf{y}AB = (nA\mathbf{y}A + nB\mathbf{y}B)/(nA + nB)$ *iar* nA, nB, si $nAB = nA + nB$ sunt nr din punctele A si B, respectiv AB.

Evaluarea variabilitatii **intra cluster:**
Daca A este compus doar din yi, iar B doar din yj, atunci SSE*A* si SSE*B* sunt egale cu zero.
Atunci ecuatia se reduce la :

$$I_{ij} = \mathrm{SSE}_{AB} = \tfrac{1}{2}(\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \tfrac{1}{2}d^2(\mathbf{y}_i, \mathbf{y}_j).$$

42. Metode de evaluare a distantelor dintre clustere

Evaluarea distantelor dintre clustere
O problema dificila care apare în analiza cluster, este legata de necesitatea evaluarii distantelor dintre clase sau clustere.
Dificultatea acestei probleme este data de faptul ca distantele dintre clase sau clustere sunt, de fapt, distante între multimi de obiecte sau distante între multimi de variabile.
Problema evaluarii distantelor dintre clustere apare în special în cazul analizei cluster de tip ierarhic, în care construirea arborelui de clustere poate fi facuta pe baza comasarii succesive sau divizarii succesive a clusterelor. Comasarea clusterelor este numita amalgamare sau agregare, iar divizarea clusterelor este numita dezagregare.
Teoretic, procesul de agregare sau dezagregare succesiva a clusterelor se bazeaza pe definirea unei distante limita între clustere, distanta numita si prag de agregare, respectiv prag de dezagregare. În principiu, decizia de comasare a doua clustere sau de divizare a unui cluster este luata numai daca distanta dintre aceste clustere este mai mica, respectiv mai mare decât distanta limita fixata.
Daca în cazul evaluarii gradului de apropiere sau departare dintre doua obiecte lucrurile sunt relativ simple, fiind suficient

sa se calculeze una din distantele mentionate mai sus, în cazul în care este necesar a fi evaluat gradul de apropiere sau departare dintre doua clustere lucrurile devin ceva mai complicate si presupun existenta unei metode specifice de evaluare.
Distanta dintre doua clustere este, de fapt, o distanta dintre doua multimi de puncte, adica o distanta mai dificil de evaluat.
Ca distanta între doua multimi de puncte, distanta dintre doua clustere poate fi masurata cu ajutorul uneia dintre mai multe metode posibile.
Dintre metodele propuse pentru evaluarea distantelor dintre clustere mentionam: metoda celor mai apropiati vecini, metoda celor mai departati vecini, metoda distantei medii între perechi, metoda centroidului si metoda lui Ward etc.

Metoda celor mai apropiati vecini
Metoda celor mai apropiati vecini evalueaza distanta dintre doua clustere ca fiind distanta minima dintre toate perechile posibile de forme din cele doua clustere. Aceasta înseamna ca distanta dintre doua clustere este masurata prin distanta dintre cele mai apropiate obiecte apartinând celor doua clase.

Definitie: Metoda celor mai apropiati vecini evalueaza distanta dintre doua clustere ca distanta între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt cele mai apropiate între ele în sensul distantei utilizate.
În figura urmatoare este vizualizata distanta dintre doua clustere, evaluata dupa metoda celor mai apropiati vecini.
Metoda celor mai departati vecini
Metoda celor mai departati vecini este metoda dupa care distanta dintre doua clase este masurata prin distanta dintre cele mai departate obiecte apartinând celor doua clustere. Pe baza acestei metode, doua clustere sunt considerate a fi mai apropiate sau mai departate, în functie de proximitatea dintre cele mai departate obiecte din cele doua clustere.

Definitie: Metoda celor mai departati vecini evalueaza distanta dintre doua clustere ca distant între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt cel mai departate între ele în sensul distantei utilizate.
Calculul distantei dintre doua clustere cu ajutorul metodei celor mai departati vecini se face pe baza datelor din matricea
distantelor dintre obiectele din cele doua clustere, prin identificarea în aceasta matrice a elementului cu valoarea cea mai mare.

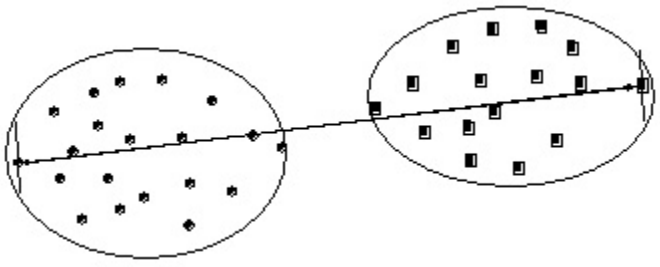


Figura 8.5: Distanaa dintre doua clustere în cazul metodei celor mai departati vecini

Pentru evaluarea distantelor dintre obiectele cele mai departate din cele doua clustere poate fi utilizata oricare dintre metodele cunoscute de calcul a distantelor dintre obiecte, în functie de natura variabilelor care definesc obiectele supuse clasificarii.

Metoda distantei medii dintre perechi

Metoda **distantei medii dintre perechile de obiecte** evalueaza distanta dintre doua clustere prin intermediul distantei medii dintre toate perechile posibile de obiecte care apartin celor doua clustere.

Definitie: Metoda distantei medii dintre perechi evalueaza distanta dintre doua clustere ca medie a distantelor dintre oricare doua obiecte care apartin celor doua clustere, unul primului cluster, iar celalalt din celui de-al doilea cluster.

Evaluarea distantei dintre doua clustere cu ajutorul metodei distantei medii între perechile de obiecte se face pe baza datelor din matricea distantelor dintre obiectele din cele doua clustere, calculând media acestor distante.

În figura urmatoare este sugerata o interpretare geometrica a modului de calcul a distantei dintre clustere cu ajutorul metodei distantei medii dintre perechi.

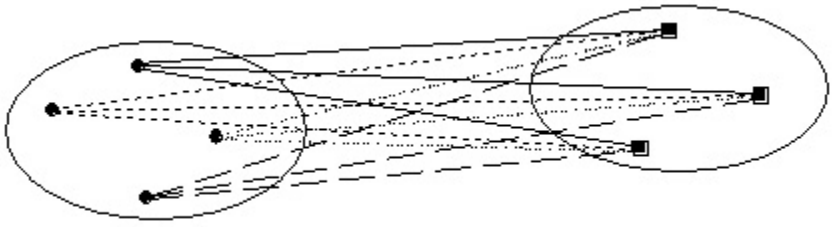


Figura 8.6: Ilustrarea grafica a metodei distantei medii dintre perechi

Ca si în cazul celorlalte doua metode, pentru evaluarea distantelor dintre obiectele celor doua clustere, poate fi utilizata oricare dintre metodele cunoscute de calcul al distantelor dintre obiecte.

Metoda centroidului

Metoda centroidului este metoda dupa care distanta dintre doua clustere este masurata ca distanta între centroizii celor doua clustere. În acest fel, doua clustere sunt considerate mai apropiate sau mai departate, în functie de gradul de apropiere sau de departare dintre centroizii lor.

Centroidul sau centrul de greutate al unui cluster reprezinta obiectul, real sau abstract, ale carui caracteristici au ca valori chiar mediile caracteristicilor obiectelor care compun clusterul respectiv.

Definitie: Metoda centroidului evalueaza distanta dintre doua clustere ca distanta între centroizii celor doua clustere.

Evaluarea distantei dintre doua clustere cu ajutorul metodei centroidului se face calculând mai întâi centroizii celor doua clustere, dupa care se evalueaza distanta dintre clustere ca distanta între acesti centroizi.

Figura urmatoare ilustreaza interpretarea geometrica a calculului distantelor dintre clustere cu ajutorul metodei centroidului.

În aceasta figura, centroizii celor doua clustere sunt marcati prin cele doua puncte de dimensiune mai mare.

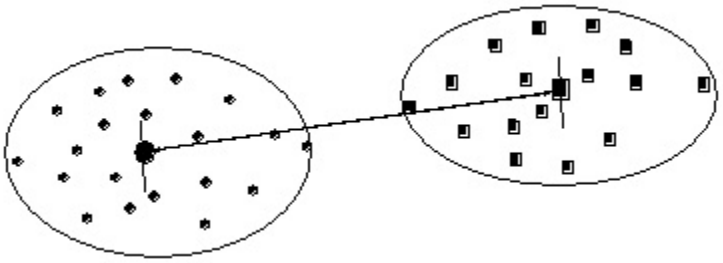


Figura 8.7: Distanta dintre clustere în cazul metodei centroidului
Deoarece centroidul este vectorul mediilor corespunzatoare tuturor obiectelor dintr-un cluster, în calculul distantei dintre doua clustere cu ajutorul metodei centroidului sunt luate în considerare, în mod implicit, toate obiectele din fiecare cluster.

Metoda lui Ward
Metoda lui Ward este o metoda de evaluare a distantei dintre doua clustere, care se bazeaza pe maximizarea gradului de omogenitate a clusterelor sau, ceea ce este acelasi lucru, pe minimizarea variabilitatii intracluster. De regula, gradul de omogenitate a unui cluster se considera a fi cu atât mai mare, cu cât suma totala a patratelor abaterilor intracluster este mai mica.
Elementul caracteristic al metodei lui Ward este reprezentat de faptul ca prin comasarea a doua clustere se urmareste obtinerea unei omogenitati maxime la nivelul tuturor clusterelor care apartin unei configuratii date a obiectelor pe clustere. În acest sens, se poate spune ca distanta Ward dintre doua clustere masoara variabilitatea intracluster cumulata, pe care o induce comasarea celor doua clustere la nivelul configuratiei cluster rezultate.

Definitie: Metoda lui Ward evalueaza distanta dintre doua clustere suma totala a patratelor abaterilor la nivelul
configuratiei cluster rezultate din comasarea celor doua clustere pentru care se evalueaza distanta. Spre deosebire de alte metode de calcul a distantelor între clustere, distanta Ward ofera o serie de avantaje. Aceste avantaje decurg din faptul ca ea este singura dintre metodele de evaluare a distantelor dintre clustere, care exprimt distantele din punct de vedere al minimizarii variabilitatii intracluster sau, ceea ce înseamna acelasi lucru, din punct de vedere al maximizarii variabilitatii intercluster.

43. Descrieti analiza cluster de tip ierarhic si mentionati care sunt cele doua categorii de clasificare ierarhica

Ierahia cluster ofera posibilitatea cercetatorului de a alege o anumita configurare a obiectelor pe clase, ceea ce înseamna, implicit, si alegerea unui anumit numar de clase.

Include metodele de clusterizare prin agregare si metodele de clusterizare prin divizare.

Pentru fiecare dintre cele doua tipuri de clusterizare exista mai multe proceduri specifice, între care mentionam: metoda agregarii simple, metoda agregarii complete, metoda agregarii medii, metoda lui Ward etc.

44. Descrieti metoda agregarii simple de analiza cluster

1. Distanta Euclidiana, care mai este cunoscuta si sub numele de norma de tip , este distanta cea mai frecvent utilizata în problemele de analiza cluster. Ea se calculeaza ca radacina patrata a sumei patratelor diferentelor coordonatelor celor doua obiecte sau variabile pentru care se evalueaza distanta.

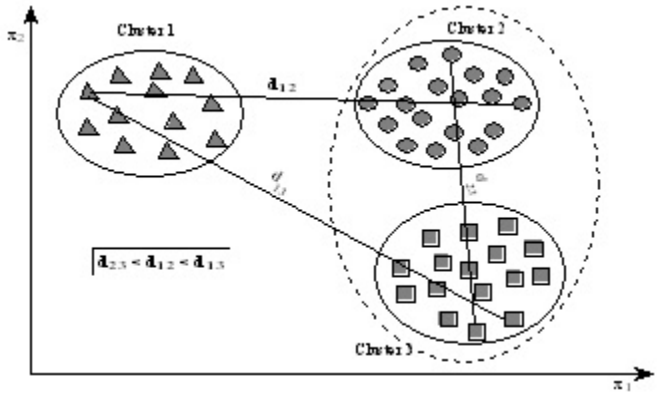
2. Distanta Manhattan, numita si distana rectangulara, distanta "City-Block" sau norma de tip , se calculeaza ca suma a valorilor absolute ale diferentelor coordonatelor celor doua obiecte sau celor doua variabile analizate.

3. Distanta Cebîsev, cunoscuta si sub numele de "maxim al dimensiunilor" sau norma de tip , este o distanta de tip valoare absoluta si se determina ca fiind valoarea maxima a valorilor absolute ale diferentelor dintre coordonatele obiectelor sau variabilelor.

4. Distanta Mahalanobis reprezinta singurul tip de distanta care ia în considerare, într-o maniera completa, gradul de dispersare al multimii de obiecte sau al multimii de variabile analizate, precum si gradul de corelare al respectivelor entitati informationale.

45. Descrieti metoda agregarii complete de analiza cluster

Metoda agregarii complete reprezinta o clasificare ierarhica de tip ascendent care comaseaza clusterele ce au cea mai mica distanta intre cei mai apropiati vecini



46. Descrieti metoda agregarii medii de analiza cluster

Metoda agregarii medii este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care distanta medie dintre toate perechile formate cu obiecte din cele doua clustere este cea mai mica, în comparatie cu alte perechi de clustere.

47. Descrieti metoda centroidului de analiza cluster

EVALUAREA DISTANTELOR DINTRE CLUSTERE

Metoda centroidului este metoda dupa care distanta dintre doua clustere este masurata ca distanta între centroizii celor doua clustere. În acest fel, doua clustere sunt considerate mai apropiate sau mai departate, în functie de gradul de apropiere sau de departare dintre centroizii lor. Centroidul sau centrul de greutate al unui cluster reprezinta obiectul, real sau abstract, ale carui caracteristici au ca valori chiar mediile caracteristicilor obiectelor care compun clusterul respective.

Definitie: Metoda centroidului evalueaza distanta dintre doua clustere ca distanta între centroizii celor doua clustere. Evaluarea distantei dintre dou| clustere cu ajutorul metodei centroidului se face calculând mai întâi centroizii celor doua clustere, dupa care se evalueaza distanta dintre clustere ca distanta între acesti centroizi. Figura urmatoare ilustreaza interpretarea geometrica a calculului distantelor dintre clustere cu ajutorul metodei centroidului. În aceasta figura, centroizii celor doua clustere sunt marcati prin cele doua puncte de dimensiune mai mare.

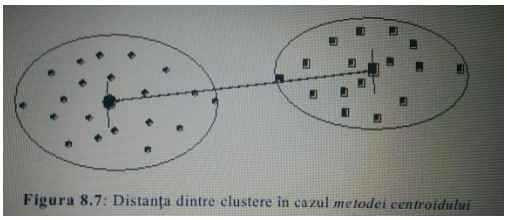


Figura 8.7: Distanţa dintre clustere în cazul *metodei centroidului*

Deoarece centroidul este vectorul mediilor corespunzatoare tuturor obiectelor dintr-un cluster, în calculul distanTei dintre doua clustere cu ajutorul metodei centroidului sunt luate în considerare, în mod implicit, toate obiectele din fiecare cluster.

METODE SI TEHNICI DE ANALIZA CLUSTER

Metoda centroidului este o metoda de clasificare ierahica ascendenta, în care distantele dintre clustere sunt evaluate cu ajutorul metodei centroidului. Ideea de baza a metodei centroidului este aceea de obtinere a unui nou cluster prin comasarea a doua clustere existente, în functie de distanta cea mai mica dintre centroizii clusterelor care sunt verificate în scopul comasarii.

Definitie: Metoda centroidului este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasific|rii acele doua clustere pentru care distanta dintre centroizii celor doua clustere este cea mai mica, în comparatie cu alte perechi de clustere. Doua clustere sunt comasate într-un nou cluster daca si numai daca distanta dintre centroizii lor este cea mai mica dintre toate distantele dintre centroizii oricaror doua clustere care apartin configuratiei cluster disponibile. În figura urmatoare este vizualizat modul de comasare a dou| clustere folosind metoda centroidului.

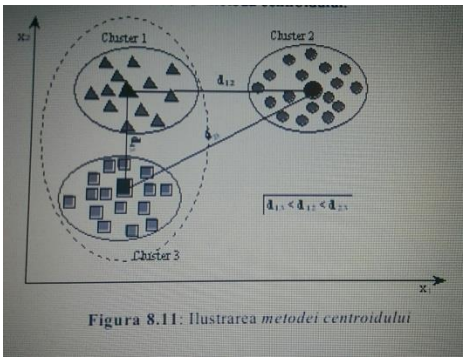


Figura 8.11: Ilustrarea *metodei centroidului*

48. Decrieti metoda lui Ward de analiza cluster

EVALUAREA DISTANTELOR DINTRE CLUSTERE

Metoda lui Ward este o metoda de evaluare a distantei dintre doua clustere, care se bazeaza pe maximizarea gradului de omogenitate a clusterelor sau, ceea ce este acelasi lucru, pe minimizarea variabilitatii intracluster. De regula, gradul de omogenitate a unui cluster se considera a fi cu atât mai mare, cu cât suma totala a patratelor abaterilor intracluster este mai mica. Elementul caracteristic al metodei lui Ward este reprezentat de faptul ca prin comasarea a doua clustere se urmareste obtinerea unei omogenitati maxime la nivelul tuturor clusterelor care apartin unei configuratii date a obiectelor pe clustere. În acest sens, se poate spune ca distanta Ward dintre doua clustere masoara variabilitatea intracluster cumulata, pe care o induce comasarea celor doua clustere la nivelul configuratiei cluster rezultate. În acest sens, distanta Ward poate fi definita sub forma urmatoare:

Definitie: Metoda lui Ward evalueaza distanta dintre doua clustere suma totala a patratelor abaterilor la nivelul configuratiei cluster rezultate din comasarea celor doua clustere pentru care se evalueaza distanta. Spre deosebire de alte metode de calcul a distantelor între clustere, distanta Ward ofera o serie de avantaje. Aceste avantaje decurg din faptul ca ea este singura dintre metodele de evaluare a distantelor dintre clustere, care exprima distantele din punct de vedere al minimizarii variabilitatii intracluster sau, ceea ce înseamna acelasi lucru, din punct de vedere al maximizarii variabilitatii intercluster.

METODE SI TEHNICI DE ANALIZA CLUSTER

Metoda lui Ward, cunoscuta si sub numele de metoda minimei variante intracluster, este una dintre cele mai cunoscute si mai eficiente metode de clasificare ierarhica prin agragare. În virtutea acestei metode atribuirea unui obiect la un cluster se face numai daca aceasta atribuire minimizeaza varianta din cadrul clusterului. Pe post de criteriu general de afectare a obiectelor la clustere este considerata minimizarea sumelor elementelor de pe diagonala matricii comune de covarianta a clusterelor, adica minimizarea trasei matricii de covarianta intraclase. Metoda lui

Ward este o metoaa de evaluare a distantei dintre dou| clustere care se bazeaza pe maximizarea gradului de omogenitate a clusterelor.
Definitie: Metoda lui Ward este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care suma patratelor abaterilor la nivelul clusterului rezultat din comasare este cea mai mica, în comparatie cu alte perechi de clustere. Metoda lui Ward nu este o metoda propriu-zisa de calcul a distantelor dintre clustere, ci o metoda de formare a clusterelor pe baza maximizarii gradului de omogenitate a clusterelor. Ca masura a gradului de omogenitate a clusterelor este utilizata suma patratelor abaterilor, numita suma patratelor abaterilor intracluster. Gradul de omogenitate a unui cluster se considera a fi cu atât mai mare cu cât suma abaterilor intracluster este mai mica. Distanta Ward se evalueaza pentru toate combinatiile posibile de comasare într-un singur cluster a oricaror doua clustere din configuratia initiala.

49. Descrieti algoritmul k-means

Algoritmul k-means asignează un punct acelui cluster al cărui centru, numit centroid, este cel mai apropiat de el. Centrul este media tuturor punctelor din cluster – coordonatele acestuia sunt determinate ca medie aritmetică, în funcție de metrica aleasă. De exemplu, daca setul de date are 3 dimensiuni iar clusterul are două puncte $X = (x_1,x_2,x_3)$ și $Y = (y_1,y_2,y_3)$, atunci centroidul Z devine $Z = (z_1,z_2,z_3)$, unde $z_1 = \frac{x_1 + y_1}{2}$, $z_2 = \frac{x_2 + y_2}{2}$ și $z_3 = \frac{x_3 + y_3}{2}$.

Pașii de execuție ai algoritmului sunt:

- Alegerea numărului de clustere k
- Generarea la întâmplare a k clustere și determinarea centrelor acestora sau generarea la întâmplare a unor puncte considerate centrele acestora
- Asignarea fiecărui punct către clusterul cu centrul cel mai apropiat, utilizând metrica aleasă
- Recalcularea centrelor clusterelor
- Repetarea celor doi pași precedenți până când se îndeplinește criteriul de convergență ales

Principalele avantaje ale algoritmului sunt simplitatea și viteza sa de execuție care îi permite să lucreze cu seturi mari de date. Dezavantajul său este faptul că nu oferă același rezultat de la execuție la execuție, din moment ce clusterele rezultante depind de inițializările care se fac la întâmplare la începutul rulării algoritmului. Alt dezavantaj este faptul că trebuie să se cunoască de la început numărul de clustere, ceea ce nu e posibil întotdeauna.

50. Ce este dendrograma (arborele de clasificare ierarhica) si cum se construieste aceasta

Analiza cluster de tip ierarhic sau *arborescent (dendograma)*este o metoda de clasificare bazata pe gruparea obiectelor pe baza de *agregare succesiva* în clase din ce în ce mai largi de obiecte sau de *dezagregare succesiva* în clase din ce în ce mai mici. Ipoteza fundamentala a analizei cluster de tip ierarhic este aceea la nivelul multimilor supuse studiului exista mai multe niveluri de structurare naturala a obiectelor pe grupe sau clase, evidentiindu-se o imbricare sau o includere, de tip arborescent, a structurilor continute la nivel latent în cadrul acestor multimi.

În cea mai mare parte a lor, algoritmii de clasificare ierahica sunt algoritmi de tip *euristic*. Exista însa si o categorie aparte algoritmi de clasificare ierarhici, reprezentati de algoritmii de tip *model formal*, care genereaza structurile cluster pe baza maximizarii verosimilitatii.
Rezultatul utilizarii analizei cluster de tip ierarhic îl reprezinta o multime de structuri particulare de clustere, numita *arbore al clasificarii* sau *arbore ierarhic*. Structurile cluster care alcatuiesc arborerele de clasificare includ un numar de clustere diferit. O solutie cluster ce corespunde unui nivel mai ridicat de agregare contine un numar de clustere mai mic cu 1 decât o solutie cluster corespunzatoare proximului nivel ierarhic inferior. Aceasta înseamna ca structurile cluster de tip ierarhic sunt caracterizate prin nivele diferite de agregare, cuprinse între un nivel minim si un nivel maxim.
Structura cluster cu cel mai înalt nivel de agregare este formata dintr-un singur cluster, care include toate obiectele supuse clasificarii. Structura cluster cu cel mai redus nivel de agregare este formata dintr-un numar de clustere egal cu numarul de obiecte analizare, fiecare cluster incluzând un singur obiect. Numarul de clustere din dou| structuri cluster succesive difera printr-o unitate, structura cluster cu nivel mai înalt de agregare continând cu un cluster mai putin decât structura cluster precedent.
Cu cât nivelul de agregare al structurilor cluster este mai ridicat, cu atât similaritatile dintre obiectele unui cluster sunt mai reduse, adic| clusterele sunt mai eterogene. Acest lucru se explica prin faptul ca un cluster de la un nivel de agregare mai înalt contine un numar mai mare de obiecte decât un cluster de la un nivel de agregare mai redus.

51. Cum se alege numarul de clustere în cazul clasificarilor de tip ierarhic

In cazul clasificarii de tip ierarhic, se va alege din multimea de solutii claster, o singura solutie. Desi alegerea se face in principal in functie de scopurile urmarite, pentru a se obtine o clasificare semnificativa, este necesar ca alegerea partitiei sa se faca pe o evaluare cat mai exacta a calitatii tuturor partilor incluse in ierarhia cluster.

52. Formulati problema generala a recunoasterii supervizate a formelor si mentionati cateva domenii de utilizare

Problema este ca in multe domenii de activitate este necesara gruparea, clasificarea si diferentierea anumitor entitati sub forma unor clase, a caror limitare trebuie sa fie foarte clara. Exista insa situatii in care informatiile de care se dispune nu sunt suficiente pentru a face aceste clasificari in mod corect. Acest lucru este des intalnit in cazul obiectelor de tip multidimensional. In acest caz diferentierea nu mai poate fi facuta numai pe cale ituitiva, fiind necesar sa se apeleze la o serie de metode statistico-matematice.
Domenii de utilizare: analiza financiara, marketing, medicina, biologie, meteorologie.

53. Definiti scopurile recunoasterii supervizate a formelor si descrieti tipul informatiilor utilizate în recunoasterea supervizata

În mod frecvent, în analiza datelor apare necesitatea studierii unor populatii care sunt *eterogene* din punct de vedere al caracteristicilor analizate, fapt care complica procesul de cunoastere a acestor populatii si impune efectuarea unui demers stiintific specific. Expresia cea mai semnificativa a populatilor de tip eterogen este întâlnita în special în domeniul statisticii, econometriei si analizei datelor, fiind reprezentata chiar de cantitatile foarte mari de informatie care trebuie prelucrata, sintetizata si interpretata.

În cazul cercetarii unor populatii de acest tip, pentru ca rezultatele investigarii sa capete consistenta si relevanta, este necesara o împartire, o divizare a acestor populatii în subpopulatii cu un anumit grad de omogenitate, urmând ca analizele si procesul de modelare implicate în studierea respectivei populatii sa se faca în mod diferentiat, pentru fiecare subpopulatie în parte.

Formularea unor concluzii corecte si robuste cu privire la manifestarea populatiilor caracterizate de un grad mai mare sau mai mic de eterogenitate nu este posibila decât daca analiza ia în considerare structurarea acestor populatii pe categorii.

În alte situatii, cum sunt cele în care sunt analizate diverse entitati economico-sociale, considerate a proveni din populatii cu caracteristici foarte diferite, exista interesul de a identifica, de a recunoaste, originea acestor entitati, si de a obtine o încadrare corecta a acestora în anumite clase reprezentative pentru populatia de origine. Situatiile de acest fel depasesc sfera economico-financiara, ele întâlnindu-se în mod frecvent într-o mare varietate de alte domenii importante ale stiintei, cum ar fi: informatica, biologia, antropologia, medicina, sociologia, geologia, meteorologia etc.

54. Ce sunt clasificatorii de tip liniar. Descrieti logica discriminarii liniare si spatiul discriminat

Prima modalitate de abordare a problemelor de clasificare cu ajutorul tehnicilor de analiza discriminanta dateaza din anul 1933 si a fost propusa de Fisher. Ulterior abordarile de acest tip s-au dezvoltat în mod constant, iar aplicatiile bazate pe analiza discriminanta s-au extins la din ce în ce mai multe domenii de activitate si s-au diversificat din ce în ce mai mult.
Cele mai multe si cele mai utile aplicatii ale analizei discriminant bazata pe criteriul lui Fisher sunt întâlnite în domeniul financiar-bancar, domeniu în care tehnicile de tip se numesc *tehnici de credit-scoring* si constituie cele mai importante instrumente pentru fundamentarea deciziilor privind acordarea de credite.

Metoda de analiza discriminanta propusa de Fisher este o metoda parametrica, caracterizata prin simplitate si robustete si care ofera posibilitati de interpretare foarte utile pentru analiza. Simplitatea acestei metode decurge din faptul ca utilizarea sa nu necesita decât evaluarea unor estimatii pentru parametrii populatiei si claselor acesteia, parametrii reprezentati de medii, variante sau covariante. Aceasta reprezinta un avantaj foarte important al analizei discriminante de tip Fisher, în comparatie, de exemplu, cu tehnicile de analiza discriminanta bazate pe criteriul Bayes-ian, tehnici a caror utilizare presupune cunoasterea probabilitatilor apriorice.

Fundamentul teoretic al analizei discriminante de tip Fisher este reprezentat de analiza variantei. Criteriul lui Fisher defineste o modalitate de deducere a functiilor discriminant pe baza analizei comparative dintre *variabilitatea intragrupala* si *variabilitatea intergrupala*, la nivelul claselor sau grupelor populatiei analizate. Functiile discriminant deduse pe baza criteriului lui Fisher se mai numesc si *functii scor* si sunt functii *liniare*.

Dupa cum am mai mentionat, criteriul fundamental care sta la baza împartirii multimii de obiecte **Ω** în submultimile ω1,ω2,...ωk este un criteriu mixt, care urmareste *minimizarea variabilitatii intragrupale* si *maximizarea variabilitatii intergrupale*. Utilizarea acestui criteriu combinat asigura cea mai buna diferentiere a claselor sau grupelor populatiei **Ω** .

Ideea care sta la baza criteriului lui Fisher este aceea a determinarii unor *directii* sau *axe*, astfel încât, de-a lungul acestora, clasele multimii **Ω** sa se *diferentieze cât mai mult între ele* si, în acelasi timp, fiecare clasa sa aiba un *grad de omogenitate cât mai mare*. Cu alte cuvinte, criteriul lui Fisher are ca scop determinarea unor *directii* de-a lungul carora variabilitatea intergrupala sa fie cât mai mare, iar variabilitatea intragrupala sa fie cât mai mica. Proiectiile obiectelor pe axele definite de aceste directii reprezinta noi coordonate ale obiectelor si se numesc *scoruri discriminant*.

Dintr-un anumit punct de vedere, analiza discriminanta poate fi considerata| ca fiind asemanatoare cu analiza componentelor principale, care are ca scop general identificarea unor axe în raport cu care variabilitatea obiectelor sa fie maxima. Deosebirea principala dintre analiza discriminanta si analiza componentelor principale este legata de faptul ca în cadrul analizei componentelor principale spatiul cauzal este considerat în integralitatea sa, fara a se face nici o diferentiere între elementele acestuia din punct de vedere al unui anumit criteriu.

În cazul analizei componentelor principale variabilitatea este privita ca o caracteristica generala a populatiei analizate, fara a se tine seama de existenta unei eventuale structurari a acestei populatii pe grupe sau clase. În consecinta, variabilitatea care face obiectul analizei componentelor principale este considerata ca un tot unitar, fara a exista posibilitatea descompunerii acesteia în raport cu o anumita structura a spatiului cauzal analizat.

Spre deosebire de aceasta, în cazul analizei discriminante se considera ca populatia analizata este structurata pe grupe sau clase, iar variabilitatea acestei populatii poate fi descompusa sub forma a doua componente importante: *variabilitatea intergrupala* si *variabilitatea intragrupala*.

În plus, fata de diferenta mentionata, în analiza discriminanta noile directii care trebuie identificate nu trebuie sa fie în mod obligatoriu ortogonale, spre deosebire de analiza componentelor principale în care directiile de variabilitate maxima trebuie sa verifice proprietatea de ortogonalitate.

Cea mai importanta problema a criteriului lui Fisher de discriminare între clasele unei populatii este legata de descompunerea variabilittaii acestei populatii. Vom detalia modul în care poate fi descompusa variabilitatea populatiei în raport cu cele doua sensuri ale acesteia: *variabilitatea simpla* - exprimata prin intermediul sumei totale a patratelor abaterilor si *variabilitatea mixta* sau *compusa* - masurata prin intermediul matricii produselor mixte ale abaterilor. Este evident ca variabilitatea mixta poate fi definita numai pentru cazul obiectelor multidimensionale.

Asa cum am precizat mai înainte, determinarea functiilor discriminant este echivalenta cu gasirea unor directii, sau vectori, în raport cu care variabilitatea intragrupala sa fie minima, iar variabilitatea intergrupala sa fie maxima. Aceste directii vor defini axele spatiului discriminat si pot fi identificate sub forma unor combinatii liniare de variabilele descriptor selectate în analiza.

55. Definiti functiile discriminant liniare, variabilele discriminant si scorurile discriminant

Functia discriminant liniara duce la obtinerea unui clasificator. De asemenea, pot fi folosite si alte criterii, cum ar fi criteriul minimizarii costului clasificarii, criteriul lui Bayes sau criteriul probabilitatilor aposteriorice si altele.
Variabilele discriminant:

Componentele vectorului $\tilde{\beta}$ reprezintă coeficienţii funcţiei discriminant liniare $D(x^c)$, ceea ce înseamnă că funcţia discriminant are forma:

$$D(x^c) = \tilde{\beta}_1 \cdot x_1^c + \tilde{\beta}_2 \cdot x_2^c + \ldots + \tilde{\beta}_n \cdot x_n^c.$$

Înlocuind variabilele centrate x_i^c cu $x_i - \mu_i$, vom obţine exprimarea funcţiei discriminat în funcţie de variabilele discriminant originale, respectiv:

$$D(x) = \tilde{\beta}_1 \cdot (x_1 - \mu_1) + \tilde{\beta}_2 \cdot (x_2 - \mu_2) + \ldots + \tilde{\beta}_n \cdot (x_n - \mu_n).$$

Izolând termenii care conţin mediile variabilelor descriptor, funcţia discriminant poate fi scrisă sub forma:

$$D(x) = -(\tilde{\beta}_1 \cdot \mu_1 + \tilde{\beta}_2 \cdot \mu_2 + \ldots + \tilde{\beta}_n \cdot \mu_n) + \tilde{\beta}_1 \cdot x_1 + \tilde{\beta}_2 \cdot x_2 + \ldots + \tilde{\beta}_n \cdot x_n.$$

În concluzie, putem spune că funcţiile discriminant ale lui Fisher sunt funcţii liniare de forma următoare:

$$D(x) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n.$$

unde $\beta_0 = -(\beta_1 \cdot \mu_1 + \beta_2 \cdot \mu_2 + \ldots + \beta_n \cdot \mu_n)$ reprezintă termenul liber, iar coeficienţii $\beta_1, \beta_2, \ldots, \beta_n$ sunt componente ale unui vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$.

În consecinţă, variabila discriminant corespunzătoare funcţiei discriminant $D(x)$ este definită astfel:

$$d = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n.$$

iar valoarea acesteia pentru o anumită formă x, adică scorurul discriminant, reprezintă evaluarea funcţiei discriminat D în punctul respectiv. Media şi varianţa variabilei discriminant d (necentrată) sunt definite de următoarele relaţii:

$$E(d) = \beta_0 + \beta_1 \cdot \mu_1 + \beta_2 \cdot \mu_2 + \ldots + \beta_n \cdot \mu_n = \beta_0 + \beta^t \cdot \mu$$

$$Var(d) = \beta^t \cdot E[(x - \mu) \cdot (x - \mu)^t] \cdot \beta = \beta^t \cdot \Sigma \cdot \beta = \beta^t \cdot \Sigma_w \cdot \beta + \beta^t \cdot \Sigma_b \cdot \beta$$

8.7.4 Definirea funcţiilor discriminant ale lui Fisher

Am prezentat mai înainte modul în care poate fi dedusă o funcţie discriminant de tip Fisher. Criteriul pe baza căruia a fost dedusă o funcţie discriminant de acest tip este un criteriu mixt, care vizează în mod simultan două aspecte: minimizarea variabilităţii intragrupale şi maximizarea variabilităţii intergrupale.

O funcţie discriminat de tip Fisher se determină ca o combinaţie liniară de variabilele discriminant, combinaţie ai cărei coeficienţi sunt componente ale unui vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Din această modalitate de definire rezultă, în mod implicit, că pot fi identificate mai multe funcţii discriminant.

Numărul maxim posibil de funcţii discriminant care pot fi identificate pe baza criteriului lui Fisher este egal cu numărul de valori proprii *distincte* şi *strict pozitive* ale matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Deoarece această matrice este de dimeniune $n \times n$, în situaţia în care ea este strict pozitiv definită şi are rangul maxim, rezultă că numărul total de funcţii discriminant care pot fi determinate este egal cu n.

Vom prezenta în continuare modul în care pot fi determinate toate funcţiile discriminant posibile. Pentru aceasta vom nota cele n valori proprii ale matricii $\Sigma_w^{-1} \cdot \Sigma_b$ cu $\lambda_1, \lambda_2, \ldots, \lambda_n$ şi vom presupune că ele sunt ordonate din punct de vedere al valorilor pe care le au astfel:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0.$$

Vom nota cu $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(n)}$ cei n vectori proprii ai matricii $\Sigma_w^{-1} \cdot \Sigma_b$, asociaţi, în ordine, cu valorile proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Prima funcţie discriminat se defineşte cu ajutorul vectorului propriu $\beta^{(1)}$, care corespunde celei mai mari valori proprii, şi are forma următoare:

$$D_1(x) = \beta_0^{(1)} + \beta_1^{(1)} \cdot x_1 + \beta_2^{(1)} \cdot x_2 + \ldots + \beta_n^{(1)} \cdot x_n.$$

Deoarece această funcţie corespunde celei mai mari valori posibile a raportului dintre varianţa intergrupală şi varianţa intragrupală, ea asigură *cea mai bună separabilitate* a claselor, din punct de vedere al criteriului mixt menţionat mai sus. Aceasta înseamnă că proiecţiile obiectelor pe noua axă determinată de vectorul de coeficienţi $\beta^{(1)}$ pot fi separate pe clase care se diferenţiază în cel mai mare grad posibil şi care au cel mai mare grad posibil de omogenitate.

În mod similar, cea de-a doua funcţie discriminat se defineşte cu ajutorul vectorului propriu care corespunde celei de-a doua valori proprii, respectiv:

$$D_2(x) = \beta_0^{(2)} + \beta_1^{(2)} \cdot x_1 + \beta_2^{(2)} \cdot x_2 + \ldots + \beta_n^{(2)} \cdot x_n.$$

56. Descrieti clasificatorul Bayesian si aratati cum poate fi utilizat acesta in predictia apartenentei formelor

Clasificatorul Bayesian

Acest algoritm de data mining are la baza notiuni fundamentale din teoria probabilitatilor. Astfel,unul dintre conceptele utilizate in cadrul acestuia este reprezentat de probabilitatea bayesiana a unui eveniment, care se defineste ca fiind gradul de incredere al unei persoane asupra aparitiei acelui eveniment. Relatiile probabilistice dintre variabilele unei multimi sunt reprezentate sub forma unui model grafic numit retea bayesiana ce poate manipula cu usurinta multimi incomplete de date.

Algoritmul este utilizat in special atunci cand dimensiunea setului de date de intrare este foarte mare. In figura urmatoare este ilustrata modalitatea de clasificare a obiectelor in verde (GREEN) sau rosu (RED). Prin aplicarea algoritmului Bayesian asupra unui set de date, orice obiect nou va fi incadrat intr-una din aceste doua categorii, dupa cum este reprezentat in figura11:

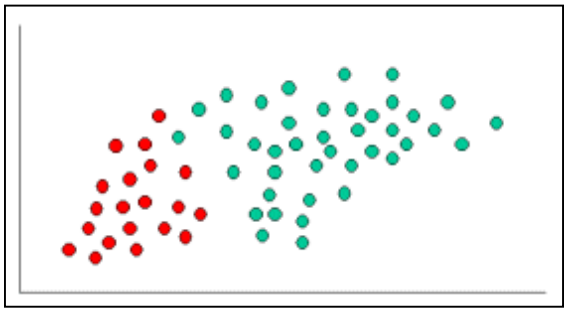


Figura 11. Set de date

Probabilitatea ca un obiect nou sa apartina uneia din cele doua categorii se determina astfel:

$$p(green) = \frac{nr.obiecte\ de\ tip\ green}{nr\ obiecte\ totale}$$

$$p(red) = \frac{nr.obiecte\ de\ tip\ red}{nr\ obiecte\ totale}$$

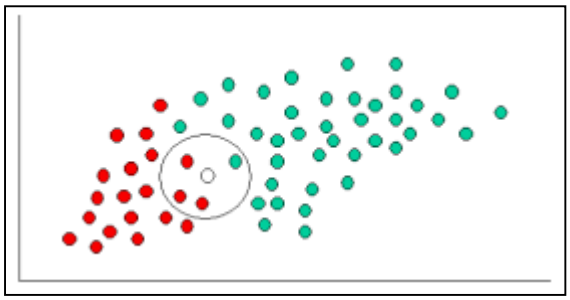


Figura 12. Instanta noua

In cazul unui nou element,se vor calcula urmatoarele probabilitati:

$$p'(green) = \frac{nr.obiecte \quad tip\ green\ din\ proxima\ vecinatate\ a\ noul\ element}{nr\ obiecte\ totale\ de\ tip\ green}$$

$$p'(red) = \frac{nr.obiecte \quad tip\ red\ din\ proxima\ vecinatate\ a\ noul\ element}{nr\ obiecte\ totale\ de\ tip\ red}$$

Ulterior, pentru stabilirea apartenentei elementului nou se va calcula clasificatorul Bayesian in ambele cazuri si se va analiza urmatoarea relatie:

$$p(A) * p'(A) < p(B) * p'(B)$$

- daca relatia este adevarata, atunci obiectul va fi incadrat in clasa A
- daca relatia este falsa, atunci obiectul va fi incadrat in clasa B

unde A si B reprezinta cele doua clase luate in considerare in vederea aplicarii algoritmului.

In conditiile in care se lucreaza cu mai multe caracteristici de va utiliza urmatoarea relatie de calcul:

$$R_k(n,m) = \frac{count\ \left(x_k^i = 1\ AND\ y^i = m\right)}{count\ \left(y^i = m\right)}$$

unde : $R_k(n,m)$ reprezinta ponderea instantelor din clasa m pentru care valoarea caracteristicii k are valoarea n.

57. Descrieti forma clasificatorului Bayesian in cazul normalitatii si homoscedasticitatii claselor

Spre deosebire de analiza discriminanta de tip Fisher, analiza Bayesiana presupune cunoasterea probabilitatilor apriorice.
Teorema lui Bayes reprezintă un prim mijloc de a determina probabilitatea unui eveniment Ai (componentă a unei repartiţii) în situaţia în care se ştie că apariţia acestuia este influenţată de îndeplinirea unui alt eveniment independent B. Prin mulţime de evenimente mutual exclusive şi exhaustive se înţeleg acele evenimente care satisfac următoarele condiţii: oricare ar fi două evenimente din mulţimea respectivă, acestea nu pot avea loc simultan (exclusivitate); cu evenimentele din mulţime se pot descrie toate stările în care se află sistemul la care aceste evenimente fac referire (exhaustivitate). În aceasta relaţie, expresia nu trebuie considerată în sensul probabilităţii de apariţie a evenimentului B atunci când se cunoaşte că evenimentul Ai a avut loc, deoarece estimarea evenimentului B este anterioară estimării lui Ai. Interpretarea corectă a acestei expresii din cadrul teoremei lui Bayes este de probabilitate ca evenimentul B să fi avut deja loc ştiind că apariţia sa a fost urmată de apariţia evenimentului Ai.

58. Descrieti clasificatorul liniar Fisher si aratati cum poate fi utilizat acesta in predictia apartenentei formelor

Separarea claselor in spatiul formelor se poate realiza prin mai multe tipuri de abordari, printre care si criterial functiilor discriminant liniare ale lui Fisher.

Cele mai multe şi cele mai utile aplicaţii ale analizei discriminant bazată pe criteriul lui Fisher sunt întâlnite în domeniul financiar-bancar, domeniu în care tehnicile de tip se numesc *tehnici de credit-scoring* şi constituie cele mai importante instrumente pentru fundamentarea deciziilor privind acordarea de credite.

Metoda de analiză discriminantă propusă de Fisher este o metodă parametrică, caracterizată prin simplitate şi robusteţe, şi care oferă posibilităţi de interpretare foarte utile pentru analiză. Simplitatea acestei metode decurge din faptul că utilizarea sa nu necesită decât evaluarea unor estimaţii pentru parametrii populaţiei şi claselor acesteia, parametri reprezentaţi de medii, varianţe sau covarianţe. Aceasta reprezintă un avantaj foarte important al analizei discriminante de tip Fisher, în comparaţie, de exemplu, cu tehnicile de analiză discriminantă bazate pe criteriul Bayes-ian, tehnici a căror utilizare presupune cunoaşterea probabilităţilor apriorice.

Fundamentul teoretic al analizei discriminante de tip Fisher este reprezentat de analiza varianţei. Criteriul lui Fisher defineşte o modalitate de deducere a funcţiilor discriminant pe baza analizei comparative dintre *variabilitatea intragrupală* şi *variabilitatea intergrupală*, la nivelul claselor sau grupelor populaţiei analizate. Funcţiile discriminant deduse pe baza criteriului lui Fisher se mai numesc şi *funcţii scor* şi sunt funcţii *liniare*.

După cum am mai menţionat, criteriul fundamental care stă la baza împărţirii mulţimii de obiecte Ω în submulţimile $\omega_1, \omega_2, ..., \omega_K$ este un criteriu mixt, care urmăreşte *minimizarea variabilităţii intragrupale* şi *maximizarea variabilităţii intergrupale*. Utilizarea acestui criteriu combinat asigură cea mai bună diferenţiere a claselor sau grupelor populaţiei Ω.

Ideea care stă la baza criteriului lui Fisher este aceea a determinării unor *direcţii* sau *axe*, astfel încât, de-a lungul acestora, clasele mulţimii Ω să se *diferenţieze cât mai mult între ele* şi, în acelaşi timp, fiecare clasă să aibă un *grad de omogenitate cât mai mare*. Cu alte cuvinte, criteriul lui Fisher are ca scop determinarea unor *direcţii* de-a lungul cărora variabilitatea intergrupală să fie cât mai mare, iar variabilitatea intragrupală să fie cât mai mică. Proiecţiile obiectelor pe axele definite de aceste direcţii reprezintă sunt noi coordonate ale obiectelor şi se numesc *scoruri discriminant*.

Dintr-un anumit punct de vedere, analiza discriminantă poate fi considerată ca fiind asemănătoare cu analiza componentelor principale, care are ca scop general identificarea unor axe în raport cu care variabilitatea obiectelor să fie maximă. Deosebirea principală dintre analiza discriminantă şi analiza componentelor principale este legată de faptul că în cadrul analizei componentelor principale spaţiul cauzal este considerat în integralitatea sa, fără a se face nici o diferenţiere între elementele acestuia din punct de vedere al unui anumit criteriu.

În cazul analizei componentelor principale variabilitatea este privită ca o caracteristică generală a populaţiei analizate, fără a se ţine seama de existenţa unei eventuale structurări a acestei populaţii pe grupe sau clase. În consecinţă, variabilitatea care face obiectul analizei componentelor principale este considerată ca un tot unitar, fără a exista posibilitatea descompunerii acesteia în raport cu o anumită structură a spaţiului cauzal analizat.

Spre deosebire de aceasta, în cazul analizei discriminante se consideră că populaţia analizată este structurată pe grupe sau clase, iar variabilitatea acestei populaţii poate fi descompusă sub forma a două componente importante: *variabilitatea intergrupală* şi *variabilitatea intragrupală*.

În plus, faţă de diferenţa menţionată, în analiza discriminantă noile direcţii care trebuie identificate nu trebuie să fie în mod obligatoriu ortogonale, spre deosebire de analiza componentelor principale în care direcţiile de variabilitate maximă trebuie să verifice proprietatea de ortogonalitate.

Cea mai importantă problemă a criteriului lui Fisher de discriminare între clasele unei populaţii Ω este legată de descompunerea variabilităţii acestei populaţii. Vom detalia modul în care poate fi descompusă variabilitatea populaţiei în raport

59. Descrieti clasificatorul Mahalanobis si aratati cum poate fi utilizat acesta in predictia apartenentei formelor

4.2.4 Distanţa Mahalanobis

Distanţa standardizată ia în considerare numai variabilitatea individuală ce caracterizează observaţiile variabilelor, ceea ce echivalează, în mod implicit, cu faptul că, în calculul acestei distanţe variabilele sunt presupuse a fi necorelate.

O generalizare a distanţei standardizate, care, spre deosebire de distanţa standardizată, ia în considerare şi variabilitatea interacţiunii dintre variabile, o reprezintă ***distanţa Mahalanobis***.

Distanţa Mahalanobis ia în considerare atât variabilitatea individuală conţinută în observaţiile efectuate asupra variabilelor, cât şi variabilitatea comună conţinută în respectivele observaţii.

Pentru a fi senzitivă în raport cu variabilitatea individuală, în construcţia distanţei Mahalanobis sunt implicate varianţele variabilelor, iar pentru a fi senzitivă în raport cu variabilitatea comună, în construcţia distanţei Mahalanobis sunt implicate covarianţele şi coeficienţii de corelaţie.

Definiţie: În cazul bidimensional, în care se consideră obiecte având câte două caracteristici, x_k şi x_l, ***distanţa Mahalanobis*** dintre două obiecte o^i şi o^j este dată de relaţia:

$$d_{Mah}(o^i,o^j) = \frac{1}{1-r^2}\cdot\left(\frac{(x_k^i - x_k^j)^2}{s_k^2} - 2r\cdot\frac{(x_k^i - x_k^j)(x_l^i - x_l^j)}{s_k s_l} + \frac{(x_l^i - x_l^j)^2}{s_l^2}\right),$$

unde r reprezintă coeficientul de corelaţie dintre cele două variabile ce reprezintă caracteristicile obiectelor, s_k^2 şi s_l^2 reprezintă varianţele, iar s_k şi s_l reprezintă abaterile standard ale celor două variabile.

Este important să observăm că distanţa standardizată şi distanţa euclidiană sunt cazuri particulare ale distanţei Mahalanobis. Într-adevăr, dacă cele două variabile ce caracterizează obiectele sunt necorelate, adică $r=0$, distanţa Mahalanobis coincide cu distanţa standardizată. Pe de altă parte, dacă varianţele variabilelor sunt egale cu unitatea şi variabilele sunt necorelate, distanţa Mahalanobis coincide cu distanţa euclidiană.

Definiţie: În cazul obiectelor *multidimensionale*, adică al obiectelor caracterizate prin intermediul a n variabile, ***distanţa Mahalanobis*** este definită de mărimea:

$$d_{Mah}(o^i,o^j) = (x^i - x^j)^t S^{-1}(x^i - x^j)$$

unde x^i şi x^j sunt vectori n-dimensionali ale căror componente sunt reprezentate de valorile caracteristicilor obiectelor o^i şi o^j, iar S este *matricea de covarianţă*.

Dacă cele n variabile ce caracterizează obiectele sunt necorelate, matricea de covarianţă S este o matrice diagonală, elemenele diagonale ale acesteia reprezentând varianţele variabilelor. În cazul în care variabilele sunt standardizate şi necorelate, matricea de covarianţă S este matricea unitate, ceea ce înseamnă că distanţa Mahalanobis se reduce la distanţa euclidiană.

• *Distanţa Mahalanobis*

Distanţa *Mahalanobis* este una dintre cele mai cunoscute, mai importante şi mai frecvent utilizate distanţe. Ea este o formă generalizată a conceptului de distanţă şi se calculează sub formele următoare:

$$d(o_p, o_j) = (x^{(i)} - x^{(j)})^t \cdot \Sigma_{n\times n}^{-1} \cdot (x^{(i)} - x^{(j)}); \qquad d(x_p, x_q) = (y^{(p)} - y^{(q)})^t \cdot \Sigma_{T\times T}^{-1} \cdot (y^{(p)} - y^{(q)}),$$

unde $x^{(i)}$ şi $x^{(j)}$ sunt vectori coloană reprezentând liniile i şi j din matricea de observaţii X, $y^{(p)}$ şi $y^{(q)}$ sunt vectori coloană reprezentând liniile p şi q din matricea de observaţii Y, iar Σ^{-1} este notaţia pentru inversa matricii de covarianţă, matrice calculată în spaţiul variabilelor - în primul caz, respectiv în spaţiul observaţiilor - în al doilea caz. Se poate observa că, în cazul în care matricea de covarianţă Σ este egală cu matricea unitate, distanţa Mahalanobis se reduce la distanţa Euclidiană pătrată.

Distanţa Mahalanobis reprezintă singurul tip de distanţă care ia în considerare, într-o manieră completă, *gradul de dispersare* al mulţimii de obiecte sau al mulţimii de variabile analizate, precum şi *gradul de corelare* al respectivelor entităţi informaţionale. Utilizarea distanţei Mahalanobis este recomandată, mai ales în situaţiile în care variabilele care descriu obiectele sunt corelate între ele. Distanţa Mahalanobis este utilizată şi în cazul tehnicilor de clasificare controlată, pe baza acestei distanţe fiind dezvoltat chiar un criteriu operaţional de discriminare.

60. Descrieti modul de stabilire a abilitatii predictive a unui clasificator si matricea corectitudinii clasificarii

Definiție: **Clasa, grupa sau clusterul** reprezintă o entitate informațională distinctă și cu semnificație concretă, formată din totalitatea obiectelor ale căror caracteristici sunt identice sau diferă foarte puțin și care sunt semnificativ diferite de caracteristicile obiectelor din alte clase sau grupe.
Definiție: **Clasificatorul sau criteriul de clasificare** reprezintă regula sau mulțimea de reguli pe baza cărora obiectele care aparțin mulțimii analizate sunt afectate sau atribuite unor clase sau grupe bine definite. În funcție de natura regulilor utilizate în procesul de clasificare, există mai multe categorii de clasificatori:

- clasificatori ierarhici
- clasificatori de cost minim
- clasificatori de distanță minimală
- clasificatori de tip Bayes-ian
- clasificatori euristici etc.

Sub cea mai generală formă a sa, **problema de clasificare** poate fi formulată în termenii teoriei deciziei, iar metodele de clasificare pot fi definite sub forma unor instrumente decizionale specifice.
Explicarea apartenenței obiectelor mulțimii la cele K clase presupune, de fapt, deducerea sau identificarea unui criteriu de clasificare sau a unei reguli de clasificare, care să descrie modul de structurare a obiectelor populației pe clase. **Criteriul de clasificare** mai este cunoscut și sub numele de clasificator.
Problema generală a clasificării: Fiind dată o mulțime de obiecte, se cere să se determine criteriul sau regula care să descrie apartenența obiectelor la clasele sub forma cărora se structurează respectiva mulțime de obiecte. În funcție de cunoașterea sau necunoașterea apriorică a apartenenței la cele K clase a obiectelor care aparțin eșantionului extras din populația , metodele de clasificare se împart în două mari categorii: **de clasificare controlată** și **de clasificare necontrolată**. Odată ce criteriul de clasificare a fost stabilit, el poate fi folosit, în continuare, pentru efectuarea de predicții privind apartenența la o anumită clasă a unor noi obiecte, din afara eșantionului existent, obiecte a căror apartenență nu este cunoscută aprioric.
După ce criteriul de clasificare a fost identificat, și cu condiția ca apartenența obiectelor aparținând eșantionului disponibil să fie cunoscută, el poate fi utilizat și pentru verificarea corectitudinii cu care acesta poate face clasificarea, adică pentru testarea calității clasificatorului.
Calitatea criteriului de clasificare poate fi testată chiar pe obiectele din eșantionul pe care acest criteriu a fost identificat. În acest scop, fiecare obiect din eșantion, a cărui apartenență la o anumită clasă este cunoscută în mod efectiv, este reclasificat cu ajutorul respectivului criteriu, iar rezultatul noii clasificări este comparat cu clasificarea reală. Testarea clasificatorului poate să conducă la o clasificare corectă a unor obiecte din eșantionul analizat și la o clasificare incorectă a altor obiecte din acest eșantion. Aceasta înseamnă că utilizarea clasificatorului respectiv poate să conducă la situația în care obiectele care aparțin în mod real unei anumite clase să fie clasificate fie în clasa corectă, fie incorect, în oricare din celelalte clase. **Modul în care un clasificator asigură clasificarea obiectelor cu apartenență cunoscută poate fi descris prin**

intermediul unei matrici, numită matricea corectitudinii clasificării sau, mai simplu, matricea clasificării, care conține informațiile necesare pentru a aprecia corectitudinea clasificării obiectelor. Dacă vom considera un eșantion format din T obiecte, care aparțin claselor w1, w2, wK , atunci matricea de clasificare are forma din tabelul următor.

Matricea clasificării

Tabelul 8.2

Clase reale	Clase de predicție				Obiecte de clasificat
	ω_1	ω_2	...	ω_K	
ω_1	T_{11}	T_{12}	...	T_{1K}	$T_{1\cdot}$
ω_2	T_{21}	T_{22}	...	T_{2K}	$T_{2\cdot}$
...	...	...	...	...	...
ω_K	T_{K1}	T_{K2}	...	T_{KK}	$T_{K\cdot}$
Obiecte clasificate	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot K}$	T

Un element **Tij** al matricii de clasificare arată numărul de obiecte aparținând în mod real clasei **wi** și care, prin utilizarea tehnicilor de recunoaștere a formelor, sunt clasificate în clasa **wi**. Definind în acest fel elementele matricii de clasificare, rezultă că numărul de obiecte clasificate corect este reprezentat de suma elementelor de pe diagonala principală a matricii clasificării, respectiv:

$$\begin{pmatrix}\text{Număr de obiecte}\\ \text{clasificate corect}\end{pmatrix} = T_{11} + T_{22} + \ldots + T_{KK} = \sum_{i=1}^{K} T_{ii}.$$

Similar, numărul de obiecte clasificate incorect este reprezentat de suma elementelor aflate în afara diagonalei principale a matricii clasificării, respectiv:

$$\begin{pmatrix}\text{Număr de obiecte}\\ \text{clasificate incorect}\end{pmatrix} = T - (T_{11} + T_{22} + \ldots + T_{KK}) = \sum_{i,j;i\neq j} T_{ij}.$$

Suma valorilor dintr-o linie a matricii de clasificare reprezintă numărul de obiecte din clasa de proveniență ce corespunde liniei respective, indiferent de clasele în care au fost clasificate acestea. Astfel, $T_{k\cdot}$ reprezintă numărul de obiecte din clasa de proveniență ω_k, indiferent de clasa în care acestea au fost clasificate. În mod similar, suma valorilor dintr-o coloană a matricii de clasificare reprezintă numărul de obiecte clasificate în clasa corespunzătoare coloanei, indiferent de clasa de proveniență a obiectelor. Rezultă că $T_{\cdot k}$ reprezintă numărul de obiecte clasificate în clasa ω_k, indiferent de clasa de proveniență a acestora.

Pe baza informațiilor din matricea de clasificare pot fi definiți o serie de indicatori care caracterizează corectitudinea clasificării. Printre aceștia menționăm:

- gradul de clasificare *corectă*:

$$p_c = \frac{\text{Număr de obiecte clasificate corect}}{\text{Număr total de obiecte clasificate}} = \frac{T_{11} + T_{22} + \ldots + T_{KK}}{T};$$

- gradul de clasificare *incorectă*:

$$p_{inc} = \frac{\text{Număr de obiecte clasificate incorect}}{\text{Număr total de obiecte clasificate}} = \frac{T - (T_{11} + T_{22} + \ldots + T_{KK})}{T}.$$

Împreună cu alți indicatori specifici, cei doi indicatori definiți anterior sunt folosiți pentru a aprecia calitatea unui clasificator, adică măsura în care acesta reușește să detecteze în mod corect apartenența obiectelor la clasele populației analizate. O clasificare este cu atât mai corectă, cu cât valoarea indicatorului p_c este mai mare.

Totalitatea activităților desfășurate în contextul unui proces de recunoaștere a formelor, împreună cu mulțimea de metode și tehnici utilizate în scopul stabilirii apartenenței formelor la anumite clase sau grupe, determină conceptul cunoscut sub numele de *sistem de recunoaștere a formelor*.