Probleme de managementul resurselor şi euristici

Activitățile de management al resurselor sunt desfăşurate într-un context larg, pot implica un personal divers şi pot urmări o mare varietate de scopuri (v. Fig. 1.).

Personal	**Context**	**Scop**	**...**
Cercetător	Întreprindere	Identificare potențiale utilizări şi beneficii	...
Manager	Lucrare	Maximizare şanse de succes	
Inginer	Proiect	Minimizare costuri	
...	...	...	

Fig. 1. O clasificare a contextului în care au loc activitățile de management al resurselor

Oricare ar fi aceşti parametrii care definesc cadrul în care sunt desfăşurate, ele presupun desfăşurarea unui algoritm decizional, bazate pe intrări (sau date de intrare) şi ieşiri (decizii). De cele mai multe ori, operațiile raționale care stabilesc deciziile pe baza datelor de intrare sunt euristice, adică sunt derivate din experiența specifică domeniului de operare şi folosesc reguli "de bun simț" (în engleză "common sense").

Uzual, în viața noastă de zi cu zi la fel ca şi în cercetarea ştiințifică noi operăm cu ***probleme***. În informatică şi ramurile derivate ale acesteia (cum e cazul bio-informaticii şi chemo-informaticii) o problemă are o semnificație precisă, foarte apropiată cu cea de ***algoritm***. Un algoritm este în esență o rețetă specificând ce să facem în anumite condiții pentru a obține un anumit obiectiv. Un algoritm necesită două resurse pentru a ***rezolva*** o problemă, şi anume timp (cu sensul de timp de execuție, mărime corelată cu numărul de instrucțiuni elementare) şi spațiu (pentru stocarea datelor de intrare şi a variabilelor).

Nu toate problemele sunt de aceeaşi ***complexitate*** şi acelaşi lucru este valabil şi pentru algoritmii de rezolvare. Astfel, unele probleme au complexitate exponențială, ceea ce înseamnă că cel mai bun algoritm rezolvă problema într-un timp de execuție ce creşte exponențial în funcție de dimensiunea (volumul, mărimea) datelor de intrare. Acest tip de probleme sunt numite ***dificile***, deoarece chiar şi cel mai bun algoritm (care există sau ar putea exista) va fi probabil nepractic cu date de intrare din practică. De exemplu, o problemă dificilă este următoarea (în care timpul de explorare al spațiului de căutare a posibilităților este exponențial):

Fiind date un număr de maşini de recoltat, un număr de câmpuri de recoltat, un număr de şoferi (şi dacă dorim un număr de spații de depozitare) să se găsească cea mai bună cale să se organizeze recoltarea la o fermă, pentru fiecare maşină de recoltat implicând un câmp şi un şofer

Fig.2. Problemă de management al resurselor

Dacă o problemă este dificilă, atunci căutarea ***optimului*** frecvent iese în afara timpului disponibil pentru aplicațiile reale. Chiar dacă există această problemă, există totuşi o serie de probleme întâlnite în practică când obținerea optimului nu este necesară (obligatorie). De cele mai multe ori o ***soluție bună*** este suficientă. Într-adevăr, presupunând că problema dificilă este organizarea recoltatului la o fermă, un algoritm permițând ca costul de recoltare să fie redus de la 40000 lei/săptămână la 10000 lei/săptămână este de un real folos pentru fermă, chiar dacă un algoritm optimal (care găseşte minimul global) ar mai putea încă să îmbunătățească organizarea reducând costul la 8000 lei/săptămână. Mai mult, desigur că algoritmul care permite reducerea costului la 10000 lei/săptămână este preferat celui care reduce costul la 8000 lei/săptămână dacă timpul de execuție al acestuia din urmă este excesiv de mare, de exemplu mai mare decât săptămâna care se organizează.

Un bun exemplu în acest sens este

Deoarece cele mai multe probleme dificile au fost împrejurul nostru de foarte mulți ani, pentru o varietate de probleme dificile unul sau mai mulți ***euristici*** au fost deja concepuți. Aceştia sunt seturi de reguli gândite pentru a rezolva o problemă anume, uzual bazați pe bunul

simț (în ceea ce privește soluția așteptată) prin evitarea erorilor grosolane, dar care nu sunt gândiți pentru a produce totdeauna soluția cu exactitate și respectiv să fie capabili să producă o soluție pentru orice valori de intrare.

Chiar dacă cei mai mulți euristici sunt foarte mult ad-hoc și dependenți de problema dată, odată cu dezvoltarea informaticii cercetătorii au reușit să formuleze trei euristici care sunt foarte generali, și anume aplicabili la o mare varietate de probleme dificile. Din cauza acestei generalități pe care o posedă, aceștia au căpătat numele de ***meta-euristici***. Toți trei sunt stocastici în natura lor (*a fi stocastic: Implicând sau conținând una sau mai multe variabile aleatoare, implicând șansa sau probabilitatea*), doi dintre aceștia (SA și GA) fiind bazați pe procese naturale care au loc în jurul nostru din totdeauna. Împreună cu **"*călirea simulată*"** (în engleză SA - "simulated annealing") și **"*căutarea tabu*"** (în engleză TS - "Tabu Search") sunt și **"*algoritmii genetici*"** (în engleză GA - Genetic Algorithm).

Este evident deci în acest punct că rezultatul activităților de management, deciziile, sunt puternic influențate de datele pe care le posedăm, au o specificitate ridicată cu privire la acestea (adică orice schimbare în datele de intrare poate influența decizia) și, în același timp, sunt obținute ca urmare a aplicării unui euristic așa încât optimalitatea acestora (sau cât de aproape ne aflăm de optim) este influențată de acesta.

Putem să definim ce ***evaluează calitatea*** unui euristic. Sunt trei criterii care trebuie considerate:

÷ viteza: cât de repede se obține soluția;
÷ precizia: cât de departe de află acea soluție de optimul global;
÷ scopul: cât de mare este subsetul datelor de intrare în raport cu setul tuturor valorilor posibile pentru care euristicul operează în raport cu anterioarele două criterii;

Chiar dacă decizia de management este o decizie umană, tot la fel de bine se aplică aici și perceptele după care decizia este luată în mod automatizat - mai devreme sau mai târziu, oricare manager capătă o serie de automatisme, care evidențiază că de fapt și-a alcătuit proprii euristici.

O problemă importantă legată de ***complexitatea algoritmică*** este reprezentată de teorema **"*inexistenței mesei pe gratis*"** (în engleză NFLT - "No Free Lunch Theorem"), teoremă care utilizând aceste trei criterii de mai sus arată că toți algoritmii sunt strict echivalenți, ceea ce înseamnă că pentru doi algoritmi A și B, pentru fiecare set de date pentru care A performează mai bine decât B există un set de date pentru care B performează mai bine decât A.

Interpretarea simplă care se dă acestei teoreme în termeni comuni și anume că oricât ai încerca să-ți faci algoritmii tăi mai isteți, este un efort în van deoarece ei vor performa la fel ca orice alt algoritm, nu este una corectă. Ceea ce teorema într-adevăr spune este că dacă se mediază performanțele tuturor algoritmilor pe toate datele posibile, atunci ei vor performa la fel. Revenind la termeni comuni, șmecheria este desigur să nu încerci să hrănești toți algoritmii pe care îi realizezi pe toate datele cu putință, ci să încerci să îți dedici algoritmul la un ***domeniu de aplicabilitate***, și aici să iei în considerare și să valorifici prin implementare în algoritm orice structură specială este posibil să existe în datele cu care intenționezi să hrănești algoritmul.

De aici rezultă că scopul algoritmului care performează bine trebuie să fie restrâns la setul de date care prezintă structurile speciale identificate. Următoarele categorii de probleme pot fi subiect de rezolvare folosind euristici:

Probleme de decizie. O problemă de decizie este definită pentr-o întrebare cu răspuns de tipul da/nu pe un set (infinit) de date de intrare; din acest motiv problemele de decizie sunt echivalente cu obținerea setului de date de intrare pentru care răspunsul problemei este da. Problemele de decizie sunt legată de problemele de optimizare atâta timp cât problema este obținerea celui mai bun răspuns la problemă.

Probleme de clasificare. O problemă de clasificare pentru obiecte dintr-un domeniu dat este în separarea acestor obiecte în clase mai mici, și producerea de criterii de determinare dacă un obiect anume dintr-un domeniu este într-o anume clasă sau nu. Una dintre cele mai faimoase probleme de clasificare este problema formulată de Carl LINNAEUS (23 Mai 23 1707 - 10 Januarie 10 1778) a clasificării viețuitoarelor după clase, ordine, genuri și specii.

Probleme de optimizare. O problemă de optimizare este o problemă de găsire a celei mai bune soluții dintre toate soluțiile posibile. În mod formal, o problemă de optimizare este un cvadruplu (I,f,m,g) unde:

÷ I - set de instanțe;
÷ f(·) - setul soluțiilor posibile definite pe I;
÷ m(·,·) - măsura definită pe produsul soluțiilor posibile şi instanțelor
÷ g - min. sau max. - funcția obiectiv
÷ scopul este găsirea optimului lui x: $m(x,f(x))=g\{m(y,f(y), y \in I\}$

Pentru fiecare problemă de optimizare există o problemă de decizie care este asociată şi a cărei întrebare este dacă există o soluție posibilă pentru o anumită măsură m_0.

Analiza calitativă şi cantitativă şi procedeul analitic

În trecut, rezultatele analizelor în medicină erau obținute în mod ***calitativ***, de aceea, majoritatea diagnosticelor erau bazate pe simptoame şi/sau examinările cu raze X, deşi era cunoscut faptul că multe boli fiziologice erau însoțite de schimbări chimice în lichidele metabolice.

Uneori erau utilizate teste pentru a detecta componenții normali sau anormali în diferite probe recoltate pentru analiză. Aceste teste în procedee prin intermediul cărora a devenit posibilă determinarea ***cantitativă*** a componenților incluşi.

Pe măsură ce precizia a crescut şi au fost stabilite proporțiile normale, a devenit clar că rezultatele de laborator au putut fi folosite în scopul precizării diagnosticelor. În prezent, pentru examinarea medicală generală a unui bolnav sau pentru a diagnostica un ansamblu specific de simptoame este nevoie de o serie de analize cantitative ale unor probe recoltate din corpul omenesc. În viitor, astfel de probe se estimează că vor deveni din ce în ce mai numeroase, iar rezultatele analizelor vor putea fi la îndemâna medicului, jucând un rol esențial la stabilirea diagnosticului. În mod curent, peste două miliarde de probe sunt executate anual în laboratoarele clinicilor medicale şi acest număr creşte mereu.Majoritatea acestor teste includ determinarea glucozei, ureei, proteinelor, sodiului, calciului, HCO_3^-/H_2CO_3, acidului uric şi pH.

Prima etapă în realizarea unui procedeu analitic o constituie ***stabilirea obiectivului*** care se urmăreşte. Numai identificând clar scopul propus, se poate imagina o ***cale logică*** care să conducă la rezolvarea corectă a problemei.

Se pot pune mai multe întrebări. De exemplu:

÷ Cu ce fel de date se operează: calitative sau cantitative?
÷ Ce informație se caută?
÷ Care este precizia cerută?
÷ Este un sistem simplu (mic) sau complex (mare)?
÷ Decizia urmează să influențeze major desfăşurarea curentă a activităților sau are efect local?
÷ Ce obstacole de implementare există?
÷ Care şi câte sunt resursele de personal implicate?
÷ Există infrastructură şi personal corespunzător pentru implementare?

O importantă sarcină care-i revine managerului este de a alege acea soluție care să conducă la ***cea mai bună rezolvare a scopului urmărit***.

Există cazuri în care libertatea de alegere este limitată. De exemplu analizele privind apa sau produsele farmaceutice trebuie să fie efectuate prin procedee aprobate de standardele legale, astfel încât soluțiile de implementare a activităților desfăşurate trebuie să țină seama de aceste standarde.

Ştiința, aşa cum o cunoaştem noi astăzi, ne oferă răspunsuri la o serie de probleme practice. În fapt, principiile şi legile chimice, fizice şi chiar matematice au luat naştere din observarea fenomenelor. În acest sens, conceptul de funcție matematică este strâns legat de

conceptul de măsurare. Definiția funcției matematice este reprezentarea informațională a modalității noastre de observare.

O serie de concepte sunt caracteristice raționamentului analitic şi pavează calea de la observație la decizie. Astfel, ***observația*** este o activitate ce consistă în recepționarea cunoaşterii prin intermediul simțurilor sau al instrumentelor. Observația presupune existența unui observator şi a unei observabile iar recepționarea cunoaşterii realizează abstractizarea rezultatului observației (de exemplu sub formă de numere sau imagini). ***Măsurarea*** este o activitate ce presupune executarea a două operații: observarea şi înregistrarea rezultatelor observației şi depinde de: natura obiectului (material) observat, natura fenomenului (imaterial) observat, de modalitatea de măsurare şi înregistrare a rezultatelor observației. Măsurarea presupune identificarea prealabilă a elementului (`e`) sau ***elementelor*** supuse observației şi are ca rezultat obținerea unei ***proprietăți*** (`|(e)`) a elementului (`e`) observat. O serie de măsurători presupune existența unei colecții de elemente distincte - ***mulțime*** - în care ordinea poate să nu fie relevantă. Mulțimea vidă (`∅`) este o mulțime care nu conține nici un element. Când proprietatea (rezultatul unei observații) înregistrată folosind exact una din exact două valori posibile denumite nefavorabile (şi notate `F` sau `0`) şi respectiv favorabile (şi notate `T` sau `1`) spunem că operăm cu ***valori de adevăr***. Mulțimea valorilor de adevăr ({0, 1} sau {F,T}) este o mulțime este o mulțime în care elementele sunt convențional ordonate (0<1, F<T). Negația logică (`!`) este operația prin intermediul căreia se ajunge de la o valoare de adevăr la cealaltă, în timp ce ***identitatea*** logică (`≡`) lasă valoarea de adevăr neschimbată şi exprimă faptul că rezultatul unei operații de măsurare asupra a două elemente este acelaşi. Prin intermediul valorii de adevăr aplicată elementelor unei mulțimi se ajunge la conceptul de submulțime. ***Apartenența*** este proprietatea unui element de a face parte (`∈`) sau nu (`∉`) dintr-o submulțime. Asocierea notată $S_1 \times S_2$ şi definită de $S_1 \times S_2 = \{(e_1,e_2) \mid e_1 \in S_1, e_2 \in S_2\}$ este ***produsul cartezian*** al mulțimilor S_1 şi S_2. Submulțimile lui $S_1 \times S_2$ se numesc ***relații binare*** iar dacă $S_1 = S_2$ acestea se mai numesc endorelații (în Fig.3 sunt redate proprietățile (endo)relațiilor binare).

Relații binare			
RE	Reflexive	(a,a) ∈ RE	=, ⊆, |, ≤
CR	Coreflexive	(a,b) ∈ CR atunci a≡b	=
QR	Cvasi-reflexive	(a,b) ∈ QR atunci (a,a), (b,b) ∈ QR	lim
IR	Ireflexive	(a,a) ∉ IR	≠, ⊥, <
SY	Simetrice	(a,b) ∈ SY atunci (b,a) ∈ SY	=, CD, CM
NS	Anti-simetrice	(a,b), (b,a) ∈ NS atunci a≡b	≤
AS	Asimetrice	(a,b) ∈ AS atunci (b,a) ∉ AS	IH, <
TS	Tranzitive	(a,b), (b,c) ∈ TS atunci (a,c) ∈ TS	=, ≤, <, ⊆, |, ⇒, IH
TL	Totale	(a,b) ∈ TL sau (b,a) ∈ TL	≤
TC	Trihotome	exact una din (a,b) ∈ TL, (b,a) ∈ TL, a≡b	<
ED	Euclidiene	(a,b), (a,c) ∈ ED atunci (b,c) ∈ ED	=
SE	Seriale	∃b : (a,b) ∈ SE	≤
UQ	Unicitate	(a,b), (a,c) ∈ UQ atunci b≡c	f(·)
EQ	Echivalențe	atunci RE, SY, TS	=, ~, ≡, CM, CD, ||
PO	Ordine parțială	atunci RE, NS, TS	|
TO	Ordine totală	atunci PO, TL	Alfabet, ≤
WO	Bine ordonate	atunci TO, SE	
⊥	Co-prime	cel mai mare divizor este 1	
VT	Adevăr vid	`dacă A atunci B` când A = Fals	
=	Egal	atunci RE, CR, SY, NS, TS, ED, EQ	
≤	Mai mic sau egal	atunci RE, NS, TS, TL, SE, PO, TO	
<	Mai mic	atunci IR, NS, AS, TS, TC, SE	
⊆	Submulțime	RE, NS, TS, SE, PO	
≠	Diferit	IR, SI	
DI	Distanță Euclidiană	RE, SI, TS, ED, SE, EQ	
IH	Moştenire	AS, TS	
CM	Congruență modulo n	EQ	
CD	Congruență div n	EQ	
lim	Limita unei serii	RE, QR	
f(·)	Funcție matematică	SE, UQ	
inj	Funcție injectivă	a ≠ b atunci f(a) ≠ f(b)	
srj	Funcție surjectivă	∃x : b=f(a)	
bij	Funcție bijectivă	INJ, SRJ	
Id	Nume	Definiție	Reprezentanți

Fig. 3. Caracteristici ale relațiilor binare

În ambele cazuri, al ***funcțiilor matematice***, şi al ***măsurătorilor experimentale*** avem asigurate două caracteristici ale relației între elementele observate şi proprietățile acestora (v. Fig. 3). astfel, pentru toate elementele observate posedăm o înregistrare a proprietății - având astfel asigurată ***serializarea*** (SE) - şi aceasta este unică (într-un moment de spațiu şi timp definit) având deci asigurată şi ***unicitatea*** (UQ). Nici o altă caracteristică cunoscută a relațiilor nu este adevărată în general nici pentru funcțiile matematice şi nici pentru funcția de măsurare, astfel încât putem spune că ceea ce realizează funcția de măsurare exprimă informațional o funcție matematică (v. Fig. 4).

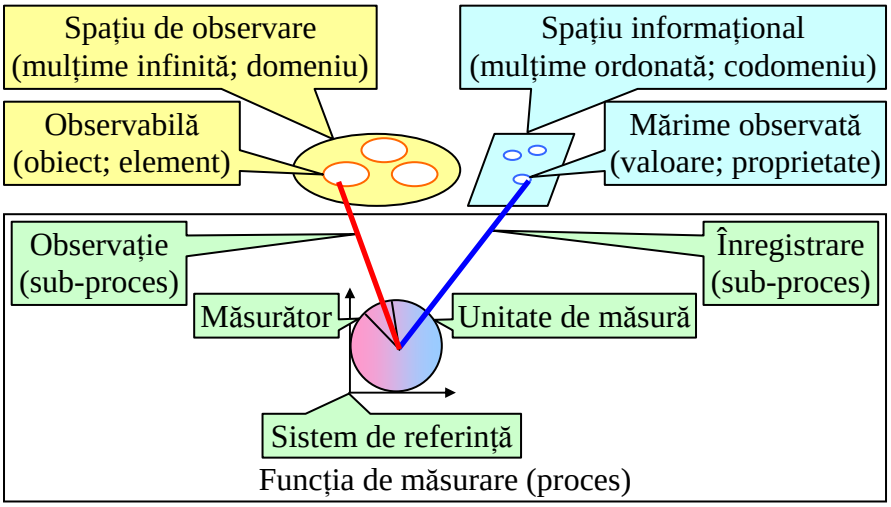


Fig. 4. Culegerea datelor experimentale este o funcție matematică

Există o serie de variabile implicit asociate funcției de măsură, cele mai importante fiind cele legate de spațiu (coordonatele observației) şi timp (momentul observației).

Pentru o mulțime finită S se poate defini o funcție (numită ***funcție de numărare***) iterativ astfel: $S_0 = S$; $S_1 = S \setminus \{s_1\}$; ... $S_i = S \setminus \{s_i\}$... Funcția $f(i) = s_i$ este o funcție de numărare pe mulțimea S, şi ne arată că orice mulțime finită e numărabilă. Alegerea elementelor s_1, ..., s_i ... din mulțimea S este instrumentul specific măsurării (presupune o observație, o înregistrare şi construcția unei submulțimi care să reunească elementele rămase).

Conceptul de funcție matematică este strâns legat de conceptul de măsurare, iar funcția de numărare este instrumentul specific cu ajutorul căruia se realizează o ordonare în spațiul informațional. Mai mult, dacă o mulțime S are n elemente, există exact n! posibilități de a enumera elementele sale prin intermediul funcției de numărare. Aşa cum se va vedea în continuare (vezi ***Nivelul de măsură***) din acest punct de vedere al legăturii cu măsurarea, de interes sunt funcțiile de numărare care aduc spațiul de observare (presupus format din elemente asupra cărora se poate aplica funcția de numărare) în spațiul informațional sub formă de numere binare (0 sau 1 prin intermediul funcției de măsurare), ordinale (naturale sau întregi) şi respectiv reale (în precizie infinită).

În preliminar, fie două mulțimi (presupus) finite A (spațiul de observare) şi B (spațiul informațional). Există exact $|B|^{|A|}$ posibilități de a construi funcții matematice $f:A \rightarrow B$ (posibilități de măsurare) care aduc elementele lui A în elemente din B.

În acest context, fie numărul total de elemente din spațiul de observare $\aleph_0$ - definim spațiul de observare drept un infinit numărabil. În raport cu acesta, numărul de posibilități de numărare este $\aleph_0!$, numărul de posibilități de măsurare care aduc elementele observate în mulțimea valorilor de adevăr ($\{0,1\}$) este $\aleph_1 = 2^{\aleph_0}$, numărul de posibilități de măsurare care aduc elementele observate în mulțimea numerelor întregi (sau naturale) - de exemplu definind o relație de ordine în legătură cu elementele observate este $\aleph_1 = 2^{\aleph_0}$ şi este egală cu numărul de posibilități de măsurare care aduc elementele observate în mulțimea numerelor reale (vezi Fig. 5). Cardinalitatea celor 3 operații descrise mai sus aduce o serie de consecințe redate în Fig. 6.

Funcție	Cardinalitate	Remarca
Observare	$\aleph_0$	Identifică elementele din spațiul de observare
Măsurare	$\aleph_1 = 2^{\aleph_0}$	Dă expresie proprietății elementelor din spațiul de observare
Numărare	$\aleph_0!$	Ordonează elementele din spațiul de observare

Fig. 5. Cardinalitatea observației, măsurării şi numărării

Operație	Convergență	Remarci
Observare vs. măsurare	$\lim_{n\to\infty} \frac{n}{2^n} = 0$	Operația de măsurare folosind valori de adevăr este informațional superioară operației de observare. Matematic nu există posibilitatea ca prin observare să se acopere întreg spațiul de posibilități de măsurare (cele două mulțimi posedă cardinalitate diferită).
Măsurare vs. numărare	$\lim_{n\to\infty} \frac{2^n}{n!} = 0$	Operația de numărare este informațional superioară operației de măsurare. Matematic nu există posibilitatea ca prin măsurare să se acopere întreg spațiul de posibilități de numărare (cele două mulțimi posedă cardinalitate diferită).
Măsurarea continuului	$\lim_{n\to\infty} \frac{n!}{2^{2^n}} = 0$	Dacă operația de măsurare ar da expresie unor funcții $f:\Re\to\Re$ atunci dacă observatorul imaginează realitatea prin funcții continue din nou măsurarea nu acoperă întreg spațiul de posibilități de enumerare (cele două mulțimi posedă cardinalitate diferită, cardinalitatea mulțimii funcțiilor $f:\Re\to\Re$ continue fiind egală cu cea a mulțimii numerelor reale $\Re$) iar dacă observatorul imaginează realitatea prin funcții oarecare măsurarea excede spațiul de posibilități de enumerare (cele două mulțimi posedă cardinalitate diferită, cardinalitatea mulțimii funcțiilor $f:\Re\to\Re$ fiind superioară posibilităților de enumerare).

Fig.6. Compararea observației cu măsurarea şi numărarea

Se desprinde o remarcă finală cu privire la scala de măsură: *dacă prin intermediul funcției de numărare avem reprezentarea informațională a spațiului de observare, atunci pentru o reprezentare nedegenerată a acestuia (proprietatea înregistrată să definească consistent în mod unic o posibilitate de enumerare a spațiului de observare) atunci măsurarea este (din păcate) insuficientă în acest scop.*

Astfel, prin măsurători imaginăm realitatea mai simplă decât enumerarea sa, dacă realitatea este formată din elemente distincte (obiecte) şi există posibilitatea să imaginăm realitatea mai complexă decât enumerarea unor elemente constitutive ale sale (cum ar fi secundele) atunci când realitatea este continuă (cum este timpul sau spațiul; dacă este timpul sau spațiul continuu sau discret a rămas însă încă ca problemă nerezolvată în fizică). Însă niciodată imaginea nu va fi atât de fidelă pe cât ne-am dori realității. Mai mult, având la dispoziție spațiul de observare discret (așa cum putem de fapt să realizăm observațiile), compararea măsurării cu numărarea duce la concluziile sintetizate în Fig. 7.

Funcție	Proprietate	Argumente
Numărare	Ordine	Funcția de numărare induce o relație de ordine în codomeniu (spațiul informațional)
Măsurare	Dezordine	Cardinalitatea măsurării infinit mai mică decât cardinalitatea numărării ($\lim_{n\to\infty} 2^n/n! = 0$)

Fig. 7. Numărare vs. măsurare şi ordine vs. dezordine

Argumentele din Fig. 7 ne arată că de exemplu în domeniul topologiei moleculare oricât

ne-am străduí să construim un descriptor (reprezentat printr-un număr) care să caracterizeze în mod unic o structură chimică acesta este mai devreme sau mai târziu contrazis de realitate (degenerarea descriptorilor de structură chimică nu poate fi evitată).

Nivele de măsură şi scale de măsură

Dacă degenerarea nu poate fi evitată prin intermediul funcției de măsurare, poate fi însă atenuată prin intermediul scării de măsură. Este de notat că nu toate scările de măsură introduc relație de ordine. Un exemplu natural este aici grupa sanguină, sau aminoacizii constituenți ai codului genetic între care nu există o relație de ordine naturală.

Să considerăm mulțimea cu 2 elemente în care ordinea elementelor nu este relevantă: C = {a,b}. Mulțimea submulțimilor acestei mulțimi este SC = {{},{a},{b},{a,b}}. O relație de ordine în mulțimea SC este definită prin numărul de elemente al (cardinalitatea) submulțimii. Relația de ordine "cardinalitate" nu este o relație de ordine strictă, existând două submulțimi cu acelaşi număr de elemente: 0 = |{}| < |{a}| = 1 = |{b}| < |{a,b}| = 2.

Ce fel de scală de măsură defineşte cardinalitatea? - pentru a afla răspunsul trebuie să ne întoarcem la observație şi anume să întrebăm: "Ce caracteristică se doreşte a fi evaluată?". Dacă răspunsul la această a doua întrebare este numărul de elemente al submulțimii observate, atunci într-adevăr mărimea măsurată este cantitativă - fiind echipată cu o relație de ordine - având submulțimea cu 0 elemente care este evident mai mică decât submulțimile cu 1 element şi care sunt evident mai mici decât submulțimea cu 2 elemente. Dacă se doreşte diferențierea submulțimilor mulțimii C, atunci cardinalitatea nu este suficientă. Putem să observăm însă numai mulțimile cu exact 1 element, pentru care măsura cardinalitate nu diferențiază: {a} şi {b}. În acest caz ne aflăm într-o situație tipică de măsură calitativă: "Submulțimea conține elementul 'a'?" - cu răspuns complementar cu răspunsul la întrebarea: "Submulțimea conține elementul 'b'?". S-a arătat astfel că procedura de definire a unei scale de măsură trebuie cel puțin verificată din punct de vedere al consistenței, sau, dacă scala este deja definită (cum a fost cazul cardinalității), se impune cel puțin verificarea consistenței acesteia în raport cu mărimea observată şi scopul urmărit. Mai mult, rezultă că chiar în absența unei relații de ordine între valorile măsurate ({a} şi {b}) pot exista însă alte tipuri de relații (ex. complementul logic: {a} = {a,b}\{b}), ceea ce face ca rezultatele unor măsurători să nu fie totdeauna independente.

În următoarea figură (Fig. 8) sunt clasificate după complexitate (definită de relațiile care se stabilesc între valorile înregistrate) scalele de măsură.

Scală	Tip	Operații	Structură	Statistici	Exemple
Binară	Logic	"=", "!"	Algebră booleană [7]	Modă, Fisher Exact [8]	Viu/Mort Fețele unei monezi
(multi) Nominală	Discret	"="	Mulțime standard	Modă, Hi pătrat	ABO (sistem grupe sanguine) Clasificarea organismelor vii
Ordinală	Discret	"=", "<"	Algebra comutativă	Mediana, Ordonare	Numărul de atomi în molecule
Interval	Continuu	"≤", "-"	Spațiu afin (unidimensional)	Media, StDev, Corelația, Regresia, ANOVA	Scala de temperatură Scala de Distanță Scala de Timp Scala de Energie
Raport	Continuu	"≤", "-", "*"	Spațiu vectorial (unidimensional)	GeoMean, HarMean, CV, Logaritm	Dulceața relativă la sucroză pH

Fig. 8. Scale de măsură

O ***scală de măsură este nominală*** dacă între valorile acesteia nu se poate defini o relație de ordine. De aici rezultă că în mod uzual scala de măsură nominală este caracteristică mărimilor calitative.

Scala (de măsură) binomială formată din doar două valori (între care nu există relație de ordine) cum ar fi: {Da, Nu}, {Viu, Mort}, {Vivo, Vitro}, {Prezent, Absent}, {Alcan saturat, Alt tip de compus}, {Număr întreg, Număr neîntreg}. Scala de măsură nominală care nu este binomială se mai numeşte şi scală de măsură multinomială.

Scala multinominală are un număr finit de elemente (valori) şi indiferent de numărul acestora, între ele există o legătură de complementaritate. Astfel, pentru o scală de măsură nominală formată din grupele sangvine {0, A, B, AB} o valoare care este diferită de oricare 3 din cele 4 valori este cu siguranță a 4-a dintre acestea.

O serie finită de valori poate să constituie o ***scală ordinală*** dacă elementele acesteia se află într-o relație de ordine. Astfel, de exemplu valorile {Prezent, Absent} enumerate între exemplele de scală binomială pot deveni scală ordinală dacă între valorile "Prezent" şi "Absent" se defineşte o relație de ordine ("Absent" < "Prezent"). Alte astfel de exemple sunt "Fals" < "Adevărat", 0 < 1, "Negativ" < "Nenegativ", "Nepozitiv" < "Pozitiv". Dintre exemplele de scale de măsură cu 3 valori unul este imediat: "Negativ" < "Zero" < "Pozitiv". Ceea ce deosebeşte suplimentar o scală ordinală de o scală nominală este faptul că nu este necesar ca scala ordinală să fie formată dintr-un număr finit (sau cunoscut) de elemente. Este necesar însă ca intre ele să existe o relație de ordine definită cel puțin printr-o funcție "Succesor" al unei valori şi complementul acesteia "Predecesor".

În ***scala interval*** distanța între atribute are o semnificație. De exemplu la măsurarea temperaturii, distanța între 30° şi 40° este aceeaşi cu distanța între 70° şi 80°. Intervalul între valori este interpretabil (are o semnificație fizică). Acesta este motivul pentru care are sens să calculăm media unei variabile de tip interval, ceea ce nu se aplică la scalele ordinale. Aşa cum 80° nu reprezintă de două ori mai cald decât 40°, pe scalele interval nu are sens raportul a două valori.

În final, pe ***scala raport*** totdeauna valoarea 0 are semnificație. În mod evident construcția unei scale raport presupune că cea mai mică valoare (care s-ar putea observa) este 0. Aceasta înseamnă că întotdeauna se poate evalua raportul a două măsuri pe o scală raport, aceasta fiind de asemenea o scală raport.

Este important de notat că calitatea unei scale de măsură nu dă şi acuratețea de măsură, sau densitatea valorilor posibile ale unei variabile în jurul valorii măsurate. Astfel, chiar dacă frecvent folosim ipoteza că o variabilă este continuă (între oricare două valori măsurate teoretic există cel puțin încă o valoare) în practică se întâmplă deseori ca valoarea intermediară a cărei existență este presupusă (sau demonstrată teoretic sau practic) să nu poată fi observată (măsurată) datorită preciziei de care dispunem în măsură. Este de notat deci că tipul scalei de măsură nu dă şi caracterul variabilei măsurate. Se pot la fel de bine măsura variabile discrete pe scale de măsură raport cum se pot măsura şi variabilele continue.

Măsurarea proprietăților biologice determină modalitatea de prelucrare şi interpretare a datelor obținute. Operația de măsurare se poate efectua doar cu ajutorul ***unei scări de măsură***. Din acest ultim unghi de vedere a problematicii măsurătorii rezultă că măsurătoarea este direct asociată cu tipul scării de măsură. Aşa cum rezultă din ce expuse mai sus, cât de exactă este o măsurătoare este la fel de important ca valoarea măsurătorii înseşi. Din acest motiv atunci când se exprimă valoarea unei măsurători aceasta este însoțită de precizie, în diferite forme de exprimare ale acesteia. Măsura referă o mărime supusă observației.

Astfel revenind la mărimile calitative şi cele cantitative discutate anterior, din acest punct de vedere ilustrat mai sus, al scalelor de măsură, o ***mărime este calitativă*** dacă pentru aceasta nu poate fi (sau cel puțin nu există) definită o scară de valori cel puțin ordonată. Dacă scara de valori a unei mărimi admite o relație de ordine (strictă) între elementele acesteia atunci mărimea este cantitativă.

Astfel, din punct de vedere al tipului scalei de măsură, o variabilă care numără moleculele dintr-un set de date este "la fel de" variabilă raport ca o variabilă care măsoară temperatura la care aceste molecule se află în mediul ambiant sau trec de la starea de agregare solidă la cea lichidă. Fig. 9 ilustrează dezordinea indusă de scalele de măsură folosind entropia

ca măsură de organizare a informației.

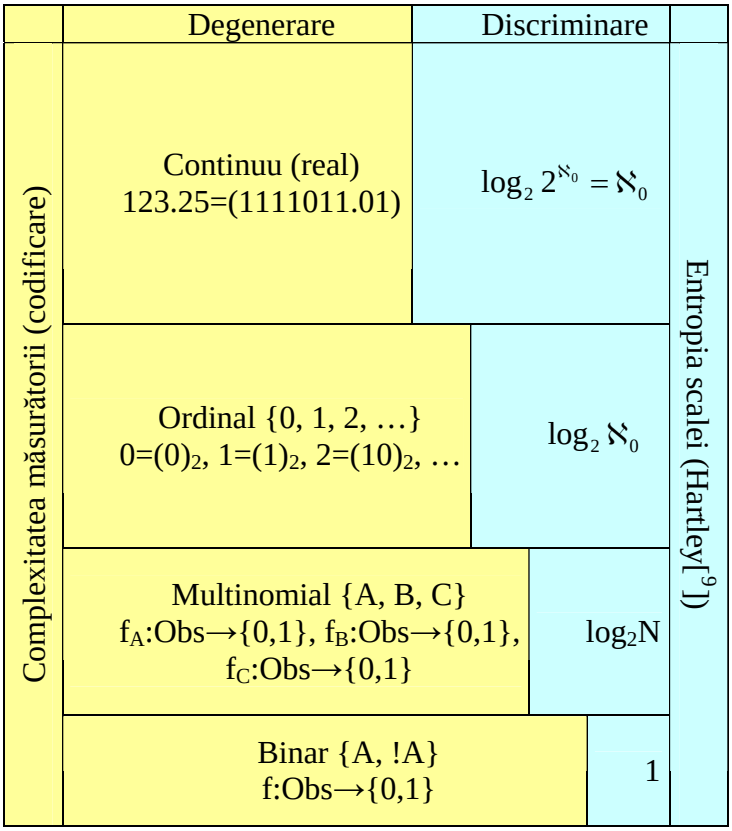


Fig. 9. Degenerare vs. discriminare

Revenind asupra ***spațiului de observare*** (vezi Fig. 4), în Fig. 10 este ilustrată structura arborescentă a relațiilor de incluziune care se stabilesc între observabilele fizice, în adâncime situându-se Universul (ca întreg spațiul de observare) iar la suprafață situându-se compușii chimici - ca formă de reprezentare a materiei cu compoziție (a atomilor) şi relații (legături între atomi) bine definite.

Structură	Proprietate
Univers	Întreg spațiul de observare
Energie radiantă	Viteză comparabilă cu viteza luminii
Radiații β, γ	Se diferențiază prin proprietăți
Materie	Întreg spațiul de observabile nerelativiste
Corp	Viteză mult mai mică decât viteza luminii
Ansamblu materiale	Compoziție (chimică) variabilă şi discontinuă
Material	Compoziție (chimică) variabilă dar continuă
Amestec substanțe	Compoziție definită
Substanțe eterogene	Compoziție (chimică) variabilă
Soluție	Stare de agregare solidă sau lichidă
Aliaj	Amestec de metale în stare solidă sau lichidă
Substanțe omogene	Compoziție (chimică) constantă
Compus chimic	Structură chimică definită şi unică

Fig. 10. Structura spațiului de observare

Sistemele posedă o structură intrinsecă care se reflectă prin intermediul funcției de măsură. În acest sens este ilustrativ exemplul structurii universului (Fig. 10).

Algoritmi genetici şi decizia asistată

Aşa cum s-a arătat în Fig. 2 problemele care necesită luarea unor decizii în viaţa de zi cu zi de cele mai multe ori sunt probleme dificile, în sensul celor menţionate anterior. Cu cât setul de date de intrare este mai voluminos, cu atât decizia se construieşte mai dificil şi de foarte multe ori complexitatea este una exponenţială de volumul datelor.

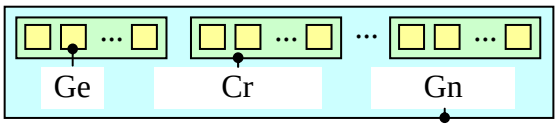
Aşa cum am anticipat, pentru a veni în sprijinul deciziei asistate, s-au dezvoltat o serie de euristici foarte generali, capabili să ofere răspuns la probleme specifice unei varietăţi de domenii de activitate. Una dintre aceste categorii de euristici este cea a algoritmilor genetici.

Algoritmii genetici posedă o caracteristică foarte importantă, şi anume sunt de inspiraţie naturală. Sunt algoritmi de căutare euristici adaptivi bazaţi pe ideile teoriei evoluţiei şi anume aduce conceptele de selecţie naturală şi genetică în arena simulării matematice cu ajutorul calculatorului. Mimica proceselor observate în evoluţia naturală a materiei organice în general serveşte drept instrument algoritmilor genetici în scopul de a rezolva probleme de decizie, clasificare, optimizare şi simulare. Elementele cheie la care se face apel în algoritmii genetici sunt:

÷ Modelul genetic (dualismul genotip - fenotip) aşa cum a fost el formulat şi argumentat încă de la primii paşi ai geneticii (Morgan & alţii, 1915; Fisher, 1918);
÷ Încrucişarea (dualismul caractere - gene) aşa cum a fost ea observată încă de la precursorii geneticii moderne (Lamarck, 1809; Mendel, 1866; Weismann, 1893);
÷ Mutaţia, aşa cum a fost ea observată încă de la precursorii geneticii moderne şi până în zilele noastre: întâmplătoare (Veies, 1902); deliberată prin expunerea la anumite condiţii (Patterson, 1928; Muller, 1928; Auerbach & alţii, 1947); sub presiunea factorilor de mediu: (Cains & alţii, 1988);
÷ Selecţia naturală sau supravieţuirea celui mai tare (Darwin, 1859).

Algoritmii genetici se materializează sub forma de programe evolutive şi sunt simulări pe calculator în care:

÷ Se operează asupra unei populaţii de reprezentări abstracte (Fig. 11) numite (după elementele genetice pe baza cărora au fost imaginate) cromozomi sau genotipuri ale unui ***genom***, la rândul său fiecare reprezentare abstractă a unui ***cromozom*** fiind compusă din ***gene***.



Legendă: Ge - Genă; Cr - Cromozom; Gn - Genom

Fig. 11. Spaţiul de căutare al unui algoritm genetic

Fiecare ***generaţie*** este compusă dintr-o populaţie de şiruri de caractere (sau alte forme de reprezentare abstractă) analog cu cromozomii ADN-ului. Fiecare element al populaţiei reprezintă un punct în spaţiul de căutare şi în acelaşi timp o soluţie posibilă. Ceea ce Fig. 11 reprezintă formal, şi anume spaţiul de căutare al unui algoritm genetic, poate avea multe variante de implementare, trei dintre ele fiind următoarele:

÷ Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile formulate în sistemul S (în engleză: S-system formalism, Savageau, 1976) care este un tip de formalism derivat din modelul de proces al reacţiilor stoechiometrice cu pre-echilibru ($\Sigma_i R_i \leftrightarrow \Sigma_j I_j \rightarrow \Sigma_k P_k$, unde R_i reactanţi, I_j intermediari, P_k produşi ai unei reacţii în care constantele de proces - constantă de viteză şi ordine parţiale de reacţie - sunt necunoscute şi se doresc a fi determinate), atunci următoarea este o posibilă implementare:
  - O genă: o constantă (un ordin parţial sau o constantă de viteză de reacţie) subiect al găsirii (optimizării);
  - Un cromozom: o posibilă cale de desfăşurare a reacţiei, având specificate toate ordinele parţiale şi constantele de viteză specificate;

- Genomul: toate căile de desfăşurarea a reacţiei prezente într-o iteraţie a algoritmului genetic;

÷ Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile de aliniament de secvenţe genetice (Notredame & Higgins, 1996) de ADN, ARN sau proteine în scopul identificării regiunilor de similaritate care pot fi sursă de relaţii structurale, funcţionale sau evolutive între secvenţe, atunci următoarea este o implementare posibilă:
- O genă: două (sau mai multe) poziţii corespunzătoare la două (sau mai multe) sub-secvenţe aliniate (sau mai exact pseudo-aliniate) şi lungimea aliniamentului acestora;
- Un cromozom: o posibilitate de aliniament pentru cele două (sau mai multe) secvenţe;
- Genomul: toate posibilităţile de aliniament de secvenţe stocate într-o iteraţie a algoritmului genetic;

÷ Dacă algoritmul genetic are ca scop o problemă de setare în managementul efectuat în scopul maximizării randamentului de producţie în câmp (Liu & alţii, 2001), o problemă dificilă de setare a parametrilor controlabili (sau alteori predictibili) pentru obţinerea unei productivităţi maxime, atunci următoarea este o implementare posibilă:
- O genă: una dintre următoarele: pH-ul solului, fertilizatori în termeni de cantitate de N, P şi K, cantitatea de materie organică din sol, gradul de creştere termică zilnică (o mărime medie între temperatura minimă şi maximă a zilei), potenţialul genetic (ce poate fi exprimat în termeni de randament care s-ar obţine dacă vremea, solul şi fertilitatea sunt toate optime), cantităţile de precipitaţii pe perioada de maximă vegetaţie pe lunile Mai, Iunie, Iulie şi August, densitatea de plantare şi factorul de rotaţie;
- Un cromozom: o stare de fapt care poate apare în practică în câmp;
- Genomul: toate stările de fapt stocate într-o iteraţie a algoritmului genetic;

Un scor sau ***şansă de supravieţuire*** a fiecărei soluţii este calculată (Fig. 12) pentru fiecare genotip cu ajutorul unei funcţii, numită şi ***funcţie obiectiv***.

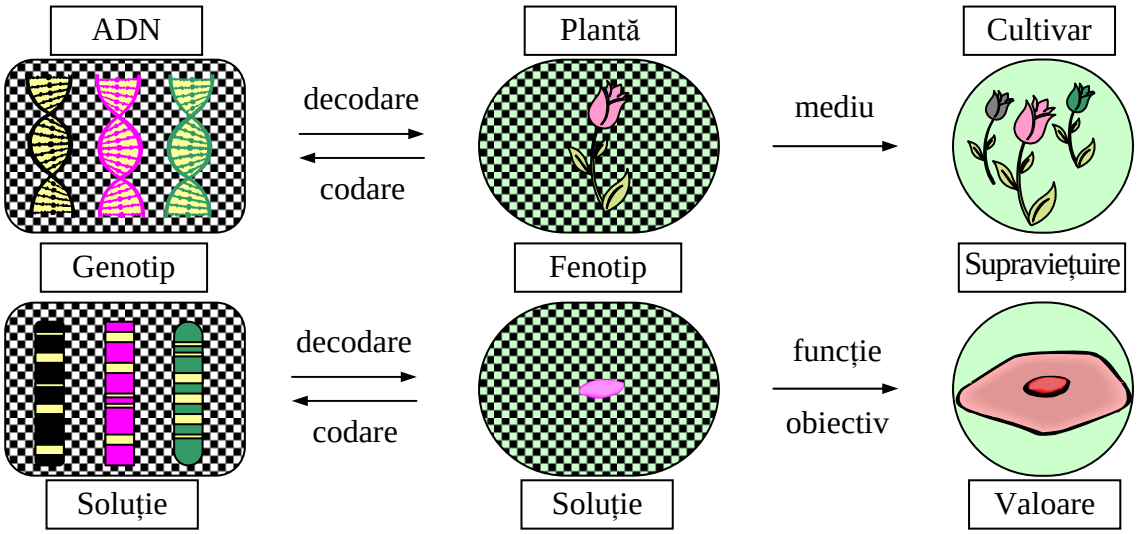


Fig. 12. Selecţia: genotip, fenotip şi supravieţuire

Valoarea funcţiei obiectiv este asociată cu abilitatea individului să supravieţuiască şi defineşte astfel ***fenotipul*** asociat genotipului. Dacă fiecare genotip reprezintă un punct în spaţiul de căutare şi în acelaşi timp o soluţie posibilă, prin intermediul selecţiei genotipul este concretizat în fenotip (operaţie care iterează reprezentarea soluţiilor posibile în spaţiul soluţiilor şi evaluează valoarea acestora). Principiul selecţiei naturale se exprimă astfel:

÷ Indivizii (fenotipurile) din populaţie concurează pentru supravieţuire (selecţie).

÷ Genele indivizilor selectaţi se propagă de la o generaţie la alta (datorită selecţiei);

÷ Fiecare generaţie devine mai potrivită mediului în care se află (prin penalizarea indivizilor

care eşuează a supravieţui).

Scorul este asociat fiecărui fenotip (soluţie) reprezentând abilitatea acestuia să concureze pentru resurse în mediu, pentru ***supravieţuire***. Scopul algoritmului genetic este ca să aplice încrucişarea şi mutaţia selectivă a fenotipurilor (prin intermediul decodării lor în genotipurile din care provin) pentru a produce descendenţi mai buni decât părinţii lor.

Evoluţia prin intermediul algoritmului genetic se realizează prin menţinerea unui eşantion din populaţie de un număr dat (sau uneori variabil) de genotipuri candidate la selecţie, care se poate face aplicând acelaşi operator. Astfel, selecţia şi supravieţuirea sunt două concepte asociate. Selecţie se face pentru operaţiile de încrucişare şi mutaţie asupra genotipurilor din eşantionul de material genetic, şi selecţie se face şi pentru supravieţuirea fenotipurilor în cultivarul de dimensiune limitată.

Pe parcursul evoluţiei, o parte din indivizii populaţiei sunt înlocuiţi de alţii. În acest mod se speră că de-a lungul generaţiilor soluţii mai bune vor ***răsări*** în timp ce cele mai slabe soluţii sunt înlăturate. Odată cu trecerea de la o generaţie la alta eşantionul va conţine din ce în ce mai bune soluţii decât generaţia anterioară.

În Fig. 13 este redată legătura care se stabileşte între scorul (exprimat prin funcţia Fitness(·) în tabel) şi regula de selecţie în funcţie de strategia (aşa cum este ea cunoscută în literatura de specialitate) folosită.

Strategie	Expresia funcţiei de scor	Selecţie	Comentarii
Proporţional	f_i=Fitness(Cromozom_i)	$p_i=f_i/\Sigma_i f_i$	Şansa de selecţie este proporţională cu scorul (utilizând probabilitatea p_i în selecţie)
Deterministic		i | f_i = max. sau min.	Selecţia indivizilor este făcută pe baza celui mai tare (sau celui mai slab) individ (elitism)
Turnir		(f_i,f_j) max. sau min.	Perechi de indivizi concurează între ei pentru selecţie (din nou cel mai tare sau cel mai slab)
Normalizare	$g_i=(f_i-N_0)(f_{max.}-f_{min.})/(N_1-N_0)$	$p_i=g_i/\Sigma_i g_i$	O scală fixă [N_0,N_1] normalizează scorul fenotipurilor între generaţii diferite
Ranguri	h_i=Rank(f_i)($f_{max.}-f_{min.}$)/Size	$p_i=h_i/\Sigma_i h_i$	Şansa este proporţională cu rangul scorului unde: Rank(·): rangul; Size: volum genom

Fig. 13. Selecţie şi scor de selecţie în algoritmii genetici

Ceea ce Fig. 12 reprezintă formal, şi anume selecţia şi supravieţuirea fenotipurilor poate avea multe variante de implementare, trei dintre ele fiind următoarele:

÷ Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile formulate în sistemul S (Savageau, 1976), atunci următoarea este o posibilă implementare:
- Şirul, corespunzător unui genotip: o listă de valori constante subiect al optimizării şi asociat cu un experiment virtual;
- Soluţia, corespunzătoare genotipului (şi cromozomului din Figura 2-1): seria de timp a elementelor experimentului virtual (pentru o reacţie chimică prin soluţie se înţeleg seriile de timp ale concentraţiilor reactanţilor, intermediarilor şi produşilor de reacţie pe parcursul desfăşurării reacţiei);
- Valoarea, corespunzătoare scorului: suma pătrată a diferenţelor dintre valorile observate (ca serie sau serii de timp) şi valorile estimate (de fenotip) ale uneia (sau mai multor) observabile (cum ar fi concentraţie sau concentraţii de intermediari);

÷ Dacă algoritmul genetic are ca subiect rezolvarea unei probleme dificile de aliniament de secvenţe de aminoacizi (Notredame & Higgins, 1996), atunci următoarea este o posibilă implementare:
- Şirul, corespunzător unui genotip: o listă de perechi (sau de mai multe) poziţii de sub-secvenţe aliniate urmată de lungimea fiecărei sub-secvenţe;
- Soluţia, corespunzătoare fenotipului (şi genotipului din Figura 2-1): o serie de valori conţinând poziţii de rupere şi lungimi de translatare necesare pentru a alinia secvenţele;
- Valoarea, corespunzătoare scorului: o funcţie de scor dând (uzual sub forma unei

sume) costul total pentru toate ruperile şi deplasările necesare pentru a alinia secvenţele, utilizând un cost predefinit pentru o rupere şi pentru deplasarea unei unităţi în secvenţă;

÷ Dacă algoritmul genetic setarea parametrilor necesari pentru obţinerea unei bune producţii în câmp (Liu & alţii, 2001), atunci următoarea este o posibilă implementare:

- Şirul, corespunzător unui genotip: o listă de valori ce corespund unui experiment virtual şi constituie obiect al optimizării; Valorile din şir pot fi: pH-ul solului, fertilizatori în termeni de cantitate de N, P şi K, cantitatea de materie organică din sol, gradul de creştere termică zilnică (o mărime medie între temperatura minimă şi maximă a zilei), potenţialul genetic (ce poate fi exprimat în termeni de randament care s-ar obţine dacă vremea, solul şi fertilitatea sunt toate optime), cantităţile de precipitaţii pe perioada de maximă vegetaţie pe lunile Mai, Iunie, Iulie şi August, densitatea de plantare şi factorul de rotaţie;
- Soluţia, corespunzătoare fenotipului (şi genotipului din Fig. 11): un şir de valori caracterizând soluţia, cuprinzând valori obţinute prin aplicarea de funcţii care să exprime: calitatea solului, calitatea vremii, managementul de cultivare, potenţialul genetic şi efectul unor evenimente întâmplătoare;
- Valoarea, corespunzătoare scorului: suma pătratelor diferenţelor între randamente observate (în serii de experimente anterioare) şi estimate (de fenotip) ale randamentelor;

Operatorul de încrucişare realizează împerecherea între fenotipuri; fenotipurile (uzual două) sunt ***selectionate*** din populaţie folosind operatorul de selecţie; o porţiune de recombinat de-a lungul şirului de gene ale genotipurilor asociate fenotipurilor este aleasă (întâmplător sau deterministic) şi valorile celor două porţiuni de şiruri sunt schimbate între ele (Fig. 14), rezultând astfel din această împerechere doi descendenţi care sunt direct selectaţi pentru a face parte din noua generaţie de populaţie; încrucişarea este făcută în speranţa că dacă se recombină porţiuni de genotipuri de succes, atunci acest proces este probabil să producă descendenţi chiar mai buni decât părinţii din care provin;

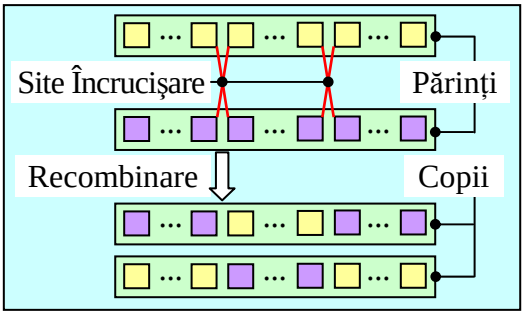


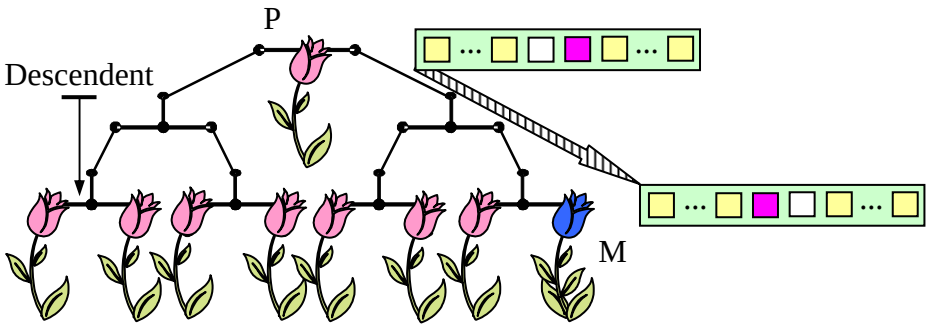
Fig. 14. O încrucişare dublă implicând ruperea şi reunirea cromozomilor părinţilor

Mutaţia este operatorul care introduce modificări noi (inexistente în populaţia unei generaţii); ceea ce este caracteristic în general mutaţiei şi implicit şi operatorului acesteia corespondent în algoritmii genetici este că ea se petrece cu o probabilitate scăzută, fiind deci aplicată cu o probabilitate scăzută (cu probabilitatea de 1/8 în Fig. 15); operatorul de mutaţie poate implementa o mutaţie:

÷ Întâmplătoare: când o porţiune a unui individ selectat va suferi schimbarea valorilor stocate în genele sale cu alte valori existente în materialul genetic al populaţiei şi are rolul menţinerii diversităţii în populaţie pentru a preveni populaţia să prezinte o convergenţă prematură;

÷ Deliberată: când expunerea la anumite condiţii se transpune în folosirea unei reguli predeterminate de modificare a valorilor genelor;

÷ Sub presiunea factorilor de mediu, când valorile genelor se schimbă în raport cu scorul fenotipului supus modificării genetice;



Legendă: P: Părinte M: Mutant

Fig. 15. Mutația

O serie de caracteristici posedă algoritmii genetici, aşa cum sunt enumerate în continuare:

÷ Utilizând doar selecția singură un algoritm nu va reuşi decât să copieze (cloneze) cel mai bun individ (fenotip) al său în întreaga populație;

÷ Utilizând mutația singură un algoritm va reuşi doar să inducă parcurgerea întâmplătoare a spațiului de căutare;

÷ Utilizând încrucişarea şi selecția un algoritm va reuşi să conveargă către o soluție bună dar nu sub-optimală (în apropierea celei optime);

÷ Mutația şi selecția (fără încrucişare) într-un algoritm creează algoritmi paraleli, toleranți la perturbații în căutarea de puncte de maxim local (în terminologia în engleză: hill-climbing);

Utilizarea tuturor operatorilor (mutație, încrucişare si selecție) asigură unui algoritm toate caracteristicile de definire ale unui algoritm genetic (Fig. 16);

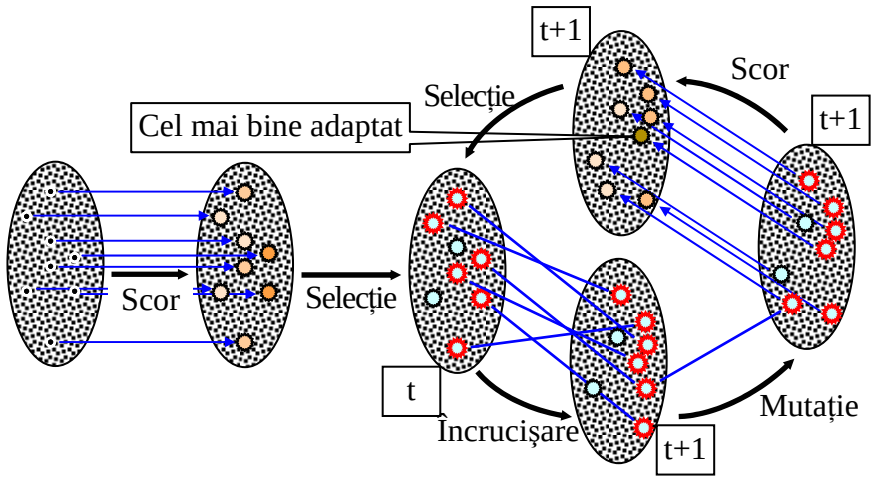


Fig. 16. Schema ilustrativă a modului de lucru al unui algoritm genetic clasic

Într-un algoritm genetic clasic (de genul celui ilustrat în Fig. 16), pentru a rezolva o problemă, se generează întâmplător sau se inițiază cu valori predefinite o populație de un volum dat de genotipuri (Fig. 11); cerințele preliminare algoritmului genetic este existența funcției obiectiv cu ajutorul căreia se evaluează scorul unui fenotip în populație; algoritmul genetic iterează astfel:

÷ Repetă
- Pasul_1: Utilizând operatorul de selecție (Fig. 12) selectează doi cromozomi;
- Pasul_2: Utilizând o funcție discretă de probabilitate pentru alegerea porțiunii de încrucişat încrucişează cei doi părinți şi creează descendenții acestora (Fig. 14);
- Pasul_3: Cu o mică probabilitate şi utilizând o funcție discretă de probabilitate pentru alegerea porțiunii de mutat efectuează mutația unui genotip (Fig. 15), eventual un descendent al încrucişării din pasul anterior;
- Pasul_4: Inițializează o nouă populație cu noile fenotipuri (de la paşii 2 şi 3

anteriori);

- Pasul_5: Completează utilizând operatorul de selecție aplicat populației de părinți noua populație cu fenotipuri (până cel puțin la refacerea numărului inițial de membrii);
- Pasul_6: Refă valorile funcției de scor ale noii populații în conformitate cu noua compoziție a acesteia;

÷ Până când cel mai bun fenotip al populației satisface o condiție impusă (condiție care reprezintă condiția de sfârşit a algoritmului).

Exemple de probleme dificile şi soluții de decizie asistată

În genetică

Răspunsul la probleme dificile de ***evoluție*** se caută adesea folosind algoritmi genetici. Astfel, genomul cloroplastului din *Manihot esculenta* şi evoluția atpF în familia *Malpighiales* sunt subiectul cercetărilor în (Daniell & alții, 2008), coniferele genului *Taxus* şi evoluția genelor paclitaxe biosintetice TS şi DBAT sunt subiectul lucrării (Hao & alții, 2009), evoluția parfumului trandafirilor chinezești sunt subiect al (Scalliet & alții, 2008), iar al plantelor cățărătoare *Hemiptera* şi *Psylloidea* în asociere cu *Anacardiaceae* sunt subiect al unui studiu sistematic în (Burckhardt & Basset, 2000).

Studiul arborilor filogenetici utilizând corespondența cu setul potrivirilor perfecte în grafuri complete (Jäntschi & Diudea, 2009) a constituit subiectul lucrării (Diaconis & Holmes, 1998). Autorii au arătat că corespondența menționată produce o metrică de distanță între arborii filogenetici, şi devine astfel o cale pentru enumerarea tuturor arborilor într-un număr minim de paşi. Identificarea arborelui filogenetic este o problemă dificilă, şi în cadrul acesteia autorii au arătat că efectuând produsul a două potriviri care este cunoscut sub denumirea de algebra Brauer (Brauer, 1937), se permite o implementare simplă a unui algoritm genetic.

Problemele legate de eşantioane mari de taxoni în estimarea filogenetică sunt discutate în (Lemmon & Milinkovitch, 2002), unde un algoritm genetic meta-populațional (metaGA) implicând mai multe populații de arbori care sunt forțate să coopereze în căutarea arborelui optim a fost găsit potrivit. Un rezultat important se desprinde din (Lemmon & Milinkovitch, 2002), şi anume că frecvențele cu care arborii şi clicile prelevate utilizând algoritmul metaGA pot corespunde la estimatorii nedeplasați ai probabilităților ulterioare (Huelsenbeck & alții, 2001).

O altă analiză de arbore filogenetic în liniile majore ale *Brachycera* a fost realizat în (Wiegmann & alții, 2003) şi indică că *Brachycera* este originată în Triasicul târziu sau în Mezozoicul timpuriu şi toate liniile majore inferioare ale zburătoarelor *Brachycera* au avut origini contemporane în Jurasicul mijlociu înainte de originile plantelor de flori (angiospermelor). Autorii au obținut o rezoluție mărită a filogeniei pentru *Brachycera*, şi estimările revizuite ale epocii zborului îmbunătățește contextul temporal al interferențelor evolutive şi comparațiilor genomice între organismele zburătoare model. Secvențele de nucleotide au fost aliniate manual cu un editor de aliniere interactiv numit Genetic Data Environment 2.2 (Smith & alții, 1994). Datele filogenetice au inclus 2220 de caractere din 28S rDNA (cuprinzând 608 variabile şi 294 parsimonic (*parsimonie: adoptarea celor mai simple presupuneri în formularea teoriei sau interpretarea datelor, în special în acord cu regula lamei de ras a lui Ockham (principiu atribuit logicianului William of OCKHAM, care subliniază că trebuie eliminate toate acele presupuneri care nu fac nici o diferență în predicțiile observate ale ipotezelor explicatoare sau teoriei); în latină: lex parsimoniae - **entia non sunt multiplicanda praeter necessitatem***) informative corespunzător la toate datele; 493 variabile şi 296 informative în *Brachycera* şi 101 caractere morfologice (Yeates, 2002). Analiza filogenetică a setului de date combinat a fost efectuată cu opțiunea parsimonie din programul PAUP (Fink, 1986).

Un studiu extins cu privire la evoluția timpurie şi diversificarea furnicilor a fost raportat

în (Brady & alţii, 2006). O importantă parte a acestui studiu este reprezentată de elaborarea metodelor de studiu care au fost folosite, şi care se regăsesc descrise în informația suplimentară lucrării menționate. Astfel, autorii au folosit o serie de programe, toate acestea operând cu algoritmi genetici:

÷ Pentru alinierea secvențelor: Clustal X (Larkin & alţii, 2007);
÷ Pentru datarea divergenței (estimarea lungimii ramificației) şi inferență filogenetică (analiză parsimonică; inferența arborilor optimi de probabilitate maximă; comparația unui set de amplasări ale grupurilor de ieşire în arborele de grupări interne folosind testul Shimodaira-Hasegawa): PAUP* v4.0b10 (Fink, 1986);
÷ Pentru obținerea modelelor de substituție nucleotidică: ModelTest v3.06 (Posada & Crandall, 1998);
÷ Pentru analiza neparametrică de încărcare a probabilității maxime: GARLI v0.94 (Schultz & alţii, 2006), derivat din GAML (Lewis, 1998);
÷ Pentru analiza Bayes: MrBayes v3.1.2 (Ronquist & Huelsenbeck, 2003);
÷ Pentru datarea divergenței (estimarea lungimii ramificației) utilizând abordarea probabilității penalizate: r8s v1.7 (Sanderson, 2002; Sanderson, 2003).

Într-un studiu ulterior (Schultz & Brady, 2008), cercetările asupra furnicilor au avut ca rezultat comunicarea identificării de relicve încă în viață de specii de furnici *attine* care ocupă poziții filogenetice care sunt de tranziție între sistemele agricole. Metodologia folosită include ca mai sus analiza filogenetică (parsimonie, probabilitate maximă şi datarea divergenței), un model nucleotidic de tip Bayes şi un model MCMC al codonului, şi în plus o nouă abordare, topografia filogenetică a sistemelor agricole:

÷ Taxelor terminale le-au fost asociate stări într-un caracter cu şase stări reprezentând patru sisteme agricole de *attine* şi agricultura tăietorilor de frunze (nu, inferior, mediu, superior, tăietor de frunze, *coral-fungus*);
÷ Cinci specii (*Myrmicocrypta n. sp. Brazil*, *Mycetagroicus triangularis*, *Cyphomyrmex n. sp.*, *Cyphomyrmex morschi*, *Trachymyrmex irmgardae*, şi *Pseudoatta n. sp.*) ale căror stări au fost asociate la 'necunoscut' şi *Trachymyrmex papulatus* a primit starea 'agricultură inferioară', asocieri de stări bazate pe o colecție de grădină din Argentina (o a doua colecție din aceeaşi localitate a cultivat o grădină tipică de *attine* înalte);
÷ Evoluția caracterelor a fost optimizată într-un arbore de consens Bates codon-model (cu lungimile ramurilor) sub ambele parsimonie folosind MacClade [10] şi probabilitate maximă folosind modulul StochChar al programului Mesquite [11];
÷ În parsimonie, optimizările stărilor ancestrale au fost neambigue. În ipoteza modelului Markov cu k stări şi 1 parametru (Lewis, 2001), probabilitatea ca fiecare sistem agricol să se ridice din cel mai recent strămoş al clicii de furnici corespunzătoare a fost, ca proporție din probabilitatea totală distribuită între cele şase stări ale caracterului, de 0.9831 pentru inferior, 0.9995 pentru mediu, 0.9905 pentru superior, 0.9924 pentru tăietorii de frunze şi 0.9998 pentru *coral-fungus*.

Altă analiză filogenetică au fost condusă utilizând algoritmi genetici pentru producție setului de reguli necesare pentru a modela distribuțiile populaționale geografice ale maimuțelor păianjen şi bocitoare prin caracterizarea nişelor sale ecologice (Ortiz-Martinez & alţii, 2008). Datorită proceselor întâmplătoare implicate în model, fiecare model obținut cu un singur set de date este diferit; pentru a captura variabilitatea autorii au elaborat 100 de modele pentru fiecare specie şi apoi au selectat 10 modele care dau cea mai mică eroare de suprapunere şi omisiune, urmând procedura descrisă în (Anderson & alţii, 2003). Autorii au putut să obțină că maimuțele păianjen ocupă un areal mai mare şi o diferență de altitudine mai mare decât maimuțele plângătoare. Validarea modelului a fost făcute pentru maimuțele păianjen, fiind suficiente date disponibile pentru această specie; validarea modelului a indicat că distribuția prezisă a speciei este statistic mai mare decât cea aşteptată de întâmplare.

În biotehnologie

(Lee & alţii, 1999) au realizat estimarea parametrilor folosind o aplicaţie hibridă având înglobată metoda simplex deal-coborâtoare (*metoda simplex deal-coborâtoare: aici se face referire la metoda elaborată de Nelder & Mead cunoscută şi sub numele de metoda Nelder-Mead şi este o metodă de optimizare numerică pentru optimizarea problemelor fără constrângeri multidimensionale, metodă care aparţine unei clase mai generale de algoritmi de căutare*) - vezi şi (Nelder & Mead, 1964) ca operator adiţional într-un algoritm genetic. În timpul evoluţiei, la fiecare pas al iteraţiei algoritmul hibrid operează astfel încât metoda simplex este folosită pentru selecţia unui procent din porţiunea superioară a populaţiei (în acord cu funcţia de scor) pentru a produce noi soluţii candidate pentru generaţia următoare. Restul populaţiei este generat folosind schema de reproducere a unui algoritm genetic clasic (cuprinzând selecţie, încrucişare şi mutaţie). Algoritmul a fost aplicat pentru optimizarea a trei cinetici de reacţie, şi autorii au remarcat îmbunătăţiri semnificative comparat cu cazul clasic. Reacţiile investigate au fost după cum urmează:

÷ Carboxilarea fosfo-enol-piruvatului (PEP) la oxalo-acetat (OAA) catalizat de P-enol-piruvat (PPC), când dioxidul de carbon este transformat la fosfat (Pi): $CO_2 + PEP \rightarrow OAA + Pi$

÷ Transformarea adenozin-tri-fosfatului (ATP) la adenozin-di-fosfat (ADP) în prezenţa OOA transformat în PEP catalizat de carboxi-kinaza fosfo-enol-piruvatului (PCK): $OAA + ATP \rightarrow PEP + ADP + CO_2$

÷ Transformarea PEP la piruvat (Pyr) în prezenţa ADP (transformat la ATP) catalizată de kinaza piruvatului (PyKi): $PEP + ADP \rightarrow Pyr + ATP$

Pizarro şi alţii au raportat o transformare a unui algoritm genetic clasic adaptată la caracteristicile unui model capabil să explice rata de creştere în fermentarea industrială la fermentarea acidului acetic (Pizarro & alţii, 2001). În abordarea făcută de autori, fiecare cromozom reprezintă o posibilă combinare a valorilor pentru fiecare din cei cinci parametrii de optimizat, reprezentaţi în cod binar. S-a definit aici un domeniu permis pentru valorile fiecărui parametru prin implementarea acestuia în codificarea binară a valorilor. Populaţia iniţială a fost constituită din valori selectate la întâmplare. Programul realizat decodează aceste valori ale parametrilor pentru fiecare cromozom şi apoi le foloseşte pentru a simula un proces de fermentare cu fiecare secvenţă de parametrii. Algoritmul de simulare rezolvă un sistem de ecuaţii diferenţiale având date constantele de viteză şi concentraţii viabile ale biomasei, pe baza relaţiilor între formarea produsului, consumul de substrat şi creşterea celulară utilizând algoritmul Runge-Kutta (*Metodele Runge-Kutta: sunt o importantă familie de metode implicite şi explicite de analiză numerică care au ca scop aproximarea soluţiilor ecuaţiilor diferenţiale ordinare. Aceste tehnici au fost propuse de C. Runge în 1895 şi completate de M. W. Kutta în 1902*). Concentraţiile iniţiale sunt considerate ca fiind acelea ale secvenţei reprezentative a procesului, în timp ce raţia iniţială biomasă viabilă/total este codificată în parametrii cromozomului. Algoritmul genetic implementat are două condiţii de stop importante: când nu se mai înregistrează valori reale pozitive pentru una din concentraţii, şi când timpul de proces în simulare a atins timpul total de proces al secvenţei reprezentative. O nouă generaţie cu acelaşi număr de cromozomi este formată prin aplicarea operatorilor de reproducere (aici înţeles cu sensul de copiere), încrucişare şi mutaţie. Cromozomii cu cea mai bună abilitate de supravieţuire obţin cel mai mare scor şi cea mai mare probabilitate de succes în adaptare (identic est: mai aproape de 1), şi au mai mare şansă să fie selectaţi şi copiaţi în noua generaţie. Încrucişarea uniformă este folosită şi cei mai buni cinci cromozomi din fiecare generaţie trec în generaţia următoare neschimbaţi. Aceşti cromozomi sunt numiţi elitişti. Cromozomii fraţi şi cei care ies din domeniu sunt blocaţi folosind o buclă de repetiţie cu filtru. Când aceşti cromozomi sunt descoperiţi după încrucişare, alţi cromozomi obţinuţi de asemenea din încrucişare îi substituie, iar dacă sunt descoperiţi după mutaţie, ei sunt înlocuiţi de cromozomii originali în aceleaşi poziţii dar sunt mutaţi din nou cu aceeaşi şansă de mutaţie. În acest proces, mutabilitatea nu este mărită, şi numărul de cromozomi rămâne constant. Procesul se opreşte după cinci generaţii fără schimbări mai mari decât un procent fixat al răspunsului mediu al cromozomilor elitişti.

Algoritmul este rulat de cinci ori la fiecare rulare a programului. O execuție finală are loc în care populația este compusă din cei mai buni cromozomi găsiți în fiecare din execuțiile anterioare. Concentrațiile acetice în fermentatoarele fabricii Vinagrerıas Riojanas SA (Logrono, Spain), obținute prin NIR (infraroşu apropiat), au fost studiate aplicând această metodologie. Datele au fost culese pe o perioadă de 4 luni fără a interveni în parametrii proceselor industriale, cum e cazul condițiilor de oxigenare şi temperaturii. Temperatura medie a fost de 29.5°C şi condițiile de oxigenare au fost suficiente pentru a asigura necesarul de oxigen, astfel încât oxigenul a devenit un substrat ne-limitator. Astăzi fermentatoarele industriale lucrează discontinuu cu schimbări (în parametrii de mediu). Bazinele bioreactoarelor studiate au fost tot timpul hrănite cu vin alb de aceeaşi origine. Timpul de proces a fost de aproximativ 30-31 ore şi pe această durată de timp 218 secvențe complete au fost obținute. O concentrație medie a secvențelor de fermentare a fost calculată din datele experimentale şi a servit în modelarea procesului. Variabilitatea concentrației în cadrul secvențelor este datorată erorilor analitice şi factorilor care nu pot fi controlați în procesul industrial, cum ar fi concentrația în etanol a vinului între procesele de fermentare. Astfel, modelul obținut folosind valoarea medie nu modelează această varianță.

Estimarea parametrilor cinetici ai poli-esterificării între acidul gras dimeric şi etilen-glicol a constituit subiectul investigației folosind un algoritm genetic clasic uşor modificat (Guangzhu & alții, 2006). Lucrarea arată că modelul dezvoltat de autori este util pentru poli-esterificarea acidului dimeric cu etilen-glicol catalizat de acidul para-toluen-sulfonic. Astfel, autorii au folosit 28.1g (0.05moli) de acid gras dimeric, 3.11g (0.05moli) de etilen-glicol şi 0.5% acid para-toluen-sulfonic (ca şi catalizator) amestec care a fost pus într-un balon cu fund rotund (cu trei capete), care a fost echipat cu un agitator şi un tub pentru azot. Azotul a fost introdus în balon pentru a înlătura oxigenul şi a preveni oxidarea materialelor. Balonul a fost plasat într-o baie de ulei cu o temperatură de 170°C. După 30 min. de reacție, azotul a fost oprit şi o pompă de vid a fost folosită pentru a scoate apa din reactant. Reacția a continuat 8-10 ore în vid. Cantitatea de acid în reactant a fost măsurată la anumite momente de timp pe durata desfăşurării reacției. Estimarea parametrilor a fost realizată în trei paşi. În primul rând, ordinul de reacție a fost confirmat utilizând presupunerea unei activități chimice egale. În al doilea rând, experimentele au fost gândite să permită estimarea parametrilor de constantă de viteză a reacției între carboxil şi monomer şi respectiv între hidroxil şi polimeri. Excesul de monomer a fost adăugat reactantului după ce acesta a reacționat pentru câteva ore cu o proporție de materiale de 1:1, şi reacțiile au putut fi ignorate exceptând cea pentru monomerul adăugat şi polimeri. În final, valorile obținute au fost introduse pentru a obține ecuațiile de viteză şi a obține valorile vitezelor de reacție între carboxil pe monomer şi hidroxil pe monomer, şi între carboxil pe polimeri şi hidroxil pe polimeri.

Modificând un algoritm genetic prin utilizarea de operații genetice ARN asupra unui model ADN şi utilizarea de operatori de mutație şi încrucişare îmbunătățită, (Tao & Wang, 2007) au reuşit să realizeze estimarea parametrilor pentru două cazuri: cracarea termică a unui ulei greu folosind un model cu trei mase şi unitate cu fluid catalitic de cracare cu fracționator (care converteşte uleiuri cu masă moleculară mare în produse hidrocarbonate mai uşoare). În ambele cazuri s-a arătat că metodologia dezvoltată este efectivă în estimarea parametrilor procesului chimic.

Un algoritm genetic pentru căutare cu ajutorul calculatorului a fost raportată recent în (Wollman & alții, 2008). Algoritmul la care se face referire utilizează măsurătorile experimentale pentru a descoperi maşinăria de mecanică moleculară care se află în spatele procesului. Au fost efectuate măsurători în serii de timp mari efectuate in vivo şi pe celule ou perturbate experimental cu scopul de a identifica modelele mecaniciste ce stau la baza coordonării generatoarelor de forțe mitotice în celulele ou de *Drosofila*. Algoritmul a fost capabil să caute şi să elimine mii de modele posibile şi să identifice şase strategii distincte pentru integrarea motorului microtubulelor care corelează cu datele avute la dispoziție. Multe caracteristici ale acestor şase strategii prezise au fost conservate, incluzând un mecanism

persistent condus de kinesin-5 combinat cu inhibiția anafazei B-specifică a caracteristicilor kinesinice şi profile de activare-deactivare pentru motoarele mitotice cheie. Abordarea de inginerie inversă a utilizat în mod indirect date cantitative pentru a realiza o căutare exhaustivă cu calculatorul şi a identifica astfel construcția mecanică a celulei ou care poate să explice datele observate. Strategia a permis examinarea unui număr mare de parametrii posibili şi mecanisme alternative utilizând modele grosiere dintre care ulterior au fost rafinate modelele promițătoare incluzând componente adiționale şi rezultând astfel modele mult mai detaliate. Aşa cum autorii subliniază, schema de lucru sugerată poate fi uşor adaptată şi la celule ou mitotice ale altor organisme şi in vitro (unde poate fi gândită diferit), şi, în fapt, la multe alte sisteme biomecanice pentru care există suficiente date cantitative.

În sisteme agricole şi horticole

Aplicațiile algoritmilor genetici în probleme specifice sistemelor agronomice au constituit subiectul unor analize critice ale literaturii de specialitate (Hashimoto, 1997; Mayer & alții, 1999). O serie de aplicații importante au fost raportate în literatura de specialitate de atunci încoace, şi noi perspective de cercetare au fost anunțate (Anisimova & Liberles, 2007).

Sisteme de decizie bazate pe algoritmi genetici pot elabora modele capabile să stabilească priorități (Smith, 2001), să configureze sisteme de producție, şi să elaboreze managementul resurselor (Kuo & Liu, 2003; Wardlaw & Bhaktikul, 2004).

Orientat pe aspectele fundamentale, în (Annevelink, 1992) se comunică realizarea unui sistem menit să asiste decizia şi managementul în sisteme horticole şi implementat sub formă de program utilizabil pe un calculator personal (PC). O remarcă este necesară aici: în general programele bazate pe algoritmi genetici sunt mari consumatoare de resurse de memorie şi timp; adaptarea acestor programe pentru a fi folosite pe calculatoarele PC obişnuite este astfel notabilă). Sistemul, denumit IMAG IPP posedă un nivel de planificare tactică, şi un mediu interactiv pentru planificarea spațiului în nivelul de planificare operațională.

Crearea unui sistem de decizie care să fie utilizat în cadrul unei metodologii de control optimal a constituit subiectul lucrării (Seginer & alții, 2007). Sistemul de decizie a fost elaborat pentru operarea unui sistem de control al umidității într-un solar cu ventilație, în care umiditatea a fost folosită drept caracter dominant de control.

Formularea considerentelor teoretice care trebuie să stea la baza elaborării a unui model dinamic pentru controlul producției (Buwalda & alții, 2006) şi utilizarea acestuia în scopul optimizării randamentului şi consumului energetic (Henten & alții, 2006) pentru cultivarea ardeilor dulci (*Capsicum annuum*) sunt subiecte ale preocupărilor actuale.

Optimizarea irigării (Montazar & alții, 2008) şi identificarea regulilor optime de cultivare (Bozorg-Haddad & alții, 2009) pentru exploatarea zonelor aride prin cultivarea de grâu, orz, porumb, sfeclă de zahăr, floarea soarelui, castraveți, ceapă, cartofi, roşii, fasole, linte, lucernă şi peri sunt dintre cele mai recente comunicări în ceea ce priveşte utilizarea algoritmilor genetici.

Optimizarea sistemelor de producție folosind un model de vegetație cu variabile independente pentru sistemele de producție a salatei verzi în două medii de dezvoltare: în solar şi pe parcele, au constituit subiectul cercetărilor în (Seginer & Ioslovich, 1999), când o serie de concluzii de importanță practică au fost obținute:

÷ Plantele de toate vârstele (situate în diferite stadii de dezvoltare) pot creşte împreună într-un singur compartiment climatizat;
÷ Spațierea trebuie planificată pentru a menține o densitate de plantare constantă;
÷ Densitatea optimă de plantare este o funcție crescătoare de cantitatea de lumină şi o funcție descrescătoare de temperatura disponibile;
÷ Dacă prețul de producție este mare în raport cu prețul de întreținere a suprafeței cultivate (în textul lucrării făcându-se aici referire la chirie) şi costul energetic, atunci intensitatea optimă de cultivare se înregistrează pentru o operare în solar în defavoarea operării pe parcele; opusul este adevărat când chiria este mare;
÷ Diferența de preț care se cere a fi plătită pentru suplimentarea iluminării este mică atunci

când lumina naturală este mai intensă şi de durată.

Modelele de creştere pot fi folosite ca instrumente ale ***simulării*** pentru estimare cantitativă. Astfel, recent (Rodkaew & alţii, 2004) s-a raportat un algoritm genetic care înglobează teoria matematică a lui Lindenmayer (Lindenmayer, 1968) pentru creşterea de soia pentru boabe.

Extinzând rezultate anterioare şi bazat pe măsurători experimentale întinse pe durata a doi ani de zile consecutivi, s-a elaborat (Salomez & Hofman, 2007) un model de creştere a salatei desfăcute (în engleză: Butterhead lettuce) care exprimă greutatea în funcție de schimbările de temperatură în sol şi radiațiile cu lungime de undă mică.

Simulatoare bazate pe algoritmi genetici au fost aplicate cu succes în predicția producției de alune supusă la contaminarea cu alfa-toxine (Henderson & alții, 2000), monitorizarea creşterii utilizând date obținute de la sateliți (Boken & alţii, 2008), evaluarea efectului metalelor grele şi PCBs (bifenili policloruraţi) asupra pico-planctonului (fracțiunea din plancton compusă din celule cu diametrul între 0.2 şi 2 μm care pot fi deopotrivă fotosintetice şi heterotrofice) marin (Caroppo & alţii, 2006), a deşeurilor militare asupra organismelor marine în (Jäntschi & Bolboacă, 2008-Marine), a toxicității de fenolii para-substituiți asupra *Tetrahymena pyriformis* (Jäntschi & alţii, 2008-Tetrahymena), precum şi analiza asocierilor complexe între proprietățile solului şi abundența de ovăz sălbatic (Diaz & alții, 2005). Evenimentele rare cum este cazul temperaturilor extreme pot fi înglobate în modelele bazate pe algoritmi genetici care simulează creşterea plantelor, aşa cum se arată în lucrarea (Kysely & Dubrovsky, 2005).

Studii sistematice ale relațiilor care se stabilesc între fenotipuri şi proprietățile acestora au fost recent realizate la vinurile de masă pentru câteva componente ale acestora (Larsen & alţii, 2006), la epistasisul plantelor cu autopolenizare (Cui & Wu, 2005), la activitatea hemoglutinativă a extracțiilor de *Curcuma aromatica* în raport cu identitatea secvenței putative (Tiptara & alţii, 2008), precum şi pentru genotipizarea *Ficus carica L* (Masi & alţii, 2005).

O abordare recentă (Letort & alții, 2008) reține atenția realizând predicția trăsăturilor fenotipice sub diferite condiții de mediu în vederea elaborării strategiilor de înmulțire şi îmbunătățirii trăsăturilor dorite.

Maşini capabile de învățare bazate pe algoritmi genetici pot servi în ***clasificare***. Astfel, se raportează obținerea de astfel de sisteme capabile de discriminarea automată a semințelor (Chtioui & alţii, 1996; Chtioui & alţii, 1997; Chtioui & alţii, 1998), ciupercilor (Hruschka & alții, 2003), şi a imaginilor de plante stocate în baze de date (Zhu & alții, 2008), precum şi pentru diferențierea secvențelor la genomii unor specii şi varietăți de iarbă (Saski & alții, 2007).

Nu în cele din urmă, algoritmii genetici îşi găsesc utilizarea în probleme de decizie, clasificare, optimizare şi simulare pentru resursele naturale aşa cum rezultă din cercetările care au fost realizate care sunt menționate în continuare.

Astfel, decizia este subiectul abordat pentru construcția politicilor strategice energetice în (Dagdeviren & Eraslan, 2008), clasificarea la forme de relief în (Moore & alții, 2003), pentru datare geologică bazată pe rația uraniu/plumb în (Lundmark & alții, 2007), în timp ce optimizarea sistemelor de asigurare a resursei energetice în horticultură este subiectul cercetărilor raportate în (Husmann & Tantau, 2001), optimizarea tratamentului termic la fructe în (Morimoto & alții, 1997), şi managementul resurselor de apă în (Chen, 1997). Simularea serveşte pentru predicția potențialului solar (Bălan & alții, 2008; Sirdas & Sahin, 2008), precum şi al potențialului resurselor de apă (Anandhi & alții, 2008; Chen & alții, 2008).

Identificarea seturilor de resurse naturale care maximizează reprezentarea diversității regionale şi menținerea pe termen lung a biodiversității (Cabeza & Moilanen, 2001), precum şi rolul schimbărilor climatice în modelarea studiilor de impact (Fowler & alții, 2007) sunt alte preocupări de actualitate care au fost abordate cu ajutorul algoritmilor genetici.

Variante, adaptări şi alternative ale formalismului algoritmilor genetici

Există multe variante şi adaptări ale algoritmilor genetici menite să îmbunătățească performanțele acestora pentru un anume tip de probleme. Menționarea tehnicilor derivate şi/sau bazate pe tehnica algoritmilor genetici este suficientă pentru problematica abordată:

÷ Optimizarea bazată pe strategia coloniilor de furnici (în engleză: Ant colony optimization) - (Bouktir & Slimani, 2005);
÷ Algoritmi bacteriologici (în engleză: Bacteriologic algorithms) - (Benoit & alţii, 2005);
÷ Metoda entropiei încrucişării (în engleză: cross-entropy method) - (Boer & alţii, 2005);
÷ Algoritmi culturali (în engleză: Cultural algorithms) - (Kobti & alţii, 2004);
÷ Strategii evolutive (în engleză: Evolution strategies) - (Schwefel, 1995);
÷ Programare evolutivă (în engleză: Evolutionary programming) - (Fogel & alţii, 1966);
÷ Optimizare extremistă (în engleză: Extremal optimization) - (Bak & Sneppen, 1993);
÷ Adaptare Gausiană (în engleză: Gaussian adaptation) - (Kjellström, 1991);
÷ Programare genetică (în engleză: Genetic programming) - (Banzhaf & alţii, 1997);
÷ Algoritmi memetici (în engleză: Memetic algorithm) - (Smith, 2007);
÷ Alte variate, colectate în (Davis, 1991).

Alte abordări conjugă algoritmii genetici cu alte concepte. Următoarele se pot menţiona:

÷ Utilizarea maşinilor cu suport vectorial (în engleză: Support Vector Machines) - (Brown & alţii, 2000);
÷ Analiza de localizare a asemănărilor structurale prin histograme secvenţiale (acronim în engleză: SPLASH) - (Califano, 2000);
÷ Setul neregulat (în engleză: Rough set) - (Hvidsten & alţii, 2001).

Baze de date şi sisteme de gestiune a bazelor de date

Calculatoarele au fost folosite încă din 1950 pentru *stocarea şi procesarea datelor*. Un deziderat major al *sistemelor informatice* este de a realiza produse software care să localizeze eficient datele pe suportul fizic şi să-l încarce rapid în memoria internă pentru procesare. La baza unui sistem informatic se află un *set de fişiere* memorate permanent pe unul sau mai multe suporturi fizice.

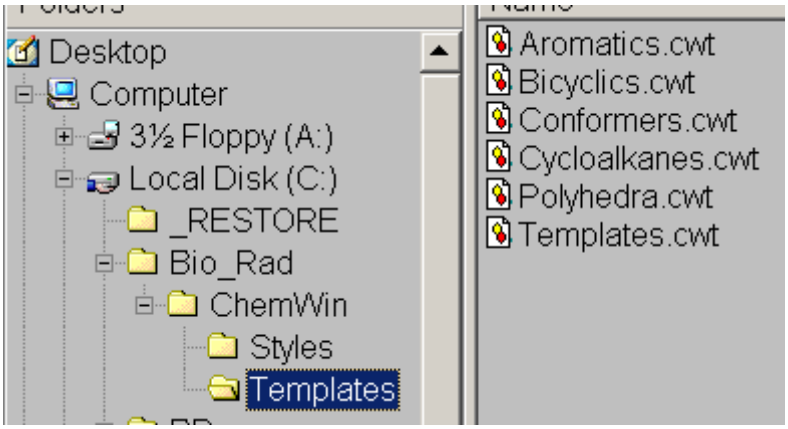


Fig. 17. Un sistem de gestiune a datelor chimice bazat pe şabloane

Gama largă de aplicaţii ale informaticii necesită acces rapid la mari cantităţi de date. Iată câteva exemple:

- sistemele computerizate de marcare din supermarketuri trebuie să traverseze întreaga linie de produse din magazin;
- sistemele de rezervare a locurilor la liniile aeriene sunt folosite în mai multe locuri simultan pentru a plasa pasageri la numeroase zboruri la date diferite;
- calculatoarele din biblioteci stochează milioane de intrări şi accesează citaţii din sute de publicaţii;
- sistemele de procesare a tranzacţiilor în bănci şi casele de brokeraj păstrează conturi care generează fluxul mondial de capital;

- motoarele de căutare World Wide Web scanează sute de pagini Web pentru a produce răspunsuri cantitative la interogări aproape instantaneu;
- sute de mici întreprinzători şi organizaţii utilizează bazele de date pentru a stoca orice de la inventare şi personal la secvenţe ADN şi informaţii despre obiecte provenite din săpături arheologice.

Un produs software care presupune managementul fişierelor suportă descompunerea logică a unui fişier în *înregistrări*. Fiecare înregistrare descrie o entitate şi constă dintr-un număr de *câmpuri*, unde fiecare câmp dă valori unei anumite proprietăţi (sau atribut) al entităţii.

Last_name	First_name	Acct_nbr	Address_1	City
Davis	Jennifer	1023495.0000	100 Cranberry St.	Wellesley
Jones	Arthur	2094056.0000	10 Hunnewell St	Los Altos
Parker	Debra	1209395.0000	74 South St	Atherton
Sawyer	Dave	3094095.0000	101 Oakland St	Los Altos
White	Cindy	1024034.0000	1 Wentworth Dr	Los Altos

Fig. 18. Descompunerea informaţiei în înregistrări

Un fişier simplu cu înregistrări este adecvat pentru date comerciale cu complexitate redusă, cum ar fi inventarul dintr-un magazin sau o colecţie de conturi curente pentru clienţi.

Un *index* al unui fişier constă dintr-o *listă de identificatori* (care disting înregistrările) *împreună cu adresele înregistrărilor*. De exemplu numele poate fi folosit pentru a identifica înregistrările unor persoane. Deoarece indexurile pot fi mari ele sunt uzual structurate într-o formă ierarhică şi sunt navigate cu ajutorul pointerilor. Formele ierarhice arborescente sunt frecvent folosite datorită vitezei mari de traversare.

Problemele reale ale procesării datelor solicită frecvent legarea datelor din mai multe fişiere. Astfel, în mod natural s-au conceput structuri de date şi programe de manipulare a datelor care să suporte legarea înregistrărilor din fişiere diferite.

3 modele de baze de date au fost create pentru a suporta legarea înregistrărilor de tipuri diferite:

÷ ***modelul ierarhic***: tipurile înregistrărilor sunt legate într-o structură arborescentă (de exemplu înregistrările unor angajaţi s-ar putea grupa după o înregistrare care să descrie departamentele în care aceştia lucrează); IMS (Information Management System produs de IBM) este un exemplu de astfel de sistem;

÷ ***modelul reţea***: se pot crea legături arbitrare între diferitele tipuri de înregistrări (de exemplu înregistrările unor angajaţi s-ar putea lega pe de o parte de o înregistrare care să descrie departamentele în care aceştia lucrează şi pe de altă parte supervizorii acestora care sunt de asemenea angajaţi);

÷ ***modelul relaţional***: în care toate datele sunt reprezentate într-o formă tabelată simplă.

În modelul relaţional descrierea unei entităţi particulare este dată de setul valorilor atributelor, stocate sub forma unei linii în tabel şi numită relaţie. Această legare a n valori de atribute furnizează cea mai potrivită descriere a entităţilor din lumea reală.

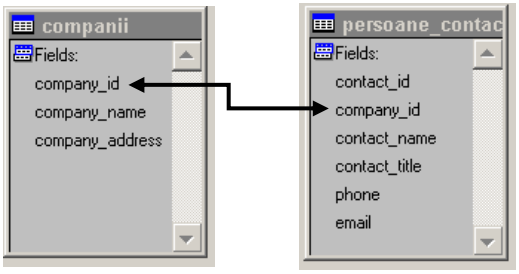


Fig. 19. Un sistem relaţional de evidenţe

Modelul relaţional suportă ***interogări*** (cereri de informaţii) care implică mai multe tabele

prin asigurarea unor legături între tabele (operația *join*) care combină înregistrări cu valori identice ale unor atribute ale acestora.

Statele de plată, de exemplu, pot fi stocate într-un tabel iar datele personalului beneficiar în altul. Informațiile complete pentru un angajat pot fi obținute prin reunirea acestor tabele (*join*) pe baza numărului personal de identificare.

Pentru a suporta o varietate de astfel de structuri de baze de date, o largă varietate a software denumită ***sistem de gestiune a bazelor de date*** este necesară pentru a stoca şi reda datele şi pentru a pune la dispoziția utilizatorului posibilitatea de a interoga și actualiza baza de date.

Gestiunea datelor presupune o structurare a acestora realizată prin definirea bazelor de date. Pentru ca exploatarea bazelor de date să fie eficientă, e necesar ca acestea să aibă un grad înalt de abstractizare. Din punct de vedere practic, este normal să se definească mai multe nivele de abstractizare. Putem lua în considerare:

÷ ***Nivelul fizic*** *(sau intern).* La acest nivel se găsesc toate detaliile legate de reprezentarea datelor pe un suport de memorie;

÷ ***Nivelul logic*** *(sau conceptual).* Se ia in considerare aspectul semantic al datelor; contează conținutul efectiv al lor, precum şi relațiile (legăturile) dintre acestea; se descriu toate bazele de date folosind structuri relativ simple în funcție de necesitățile impuse de anumite aplicații;

÷ ***Nivelul extern***. Acest nivel de abstractizare este cel în care se poate descrie conținutul unor baze de date; are în vedere simplificarea interacțiunii utilizator - bază de date.

Pentru descrierea bazelor de date facem apel la noțiunea de *structură* de date care reprezintă un ansamblu de instrumente conceptuale care permit descrierea datelor, a legăturilor dintre ele, semantica lor sau *constrângerile* la care ele sunt supuse.

Bazele de date evoluează în timp. Mulțimea informațiilor conținute în baza de date la un moment dat defineşte ***instanțierea*** bazei de date.

În 1970, Ted Codd (IBM, părintele SQL), nemulțumit de performanțele COBOL şi IMS formulează principiul de lucru al bazelor de date relaționale. Codd afirmă că SGBD trebuie să recunoască comenzi simple şi trebuie să fie aproape de utilizatori prin punerea împreună a comenzilor potrivite pentru găsirea a ceea ce se doreşte. Ceea ce Codd numeşte model relațional se bazează pe două puncte cheie:

÷ să furnizeze un mod de descriere a datelor cu numai cu structura lor naturală, ceea ce înseamnă că trebuie realizat acest lucru fără impunerea nici unei structuri adiționale pentru scopuri de reprezentare în calculator;

÷ de asemenea, să furnizeze baza pentru un limbaj de date de nivel înalt care va conduce la o maximă independență între programe, pe de o parte, şi reprezentarea în calculator, pe de altă parte.

O bază de date relațională extinde conceptul de tabele; este compusă dintr-o mulțime de tabele între care se definesc relații în sens matematic.

Să presupunem că avem T_1, T_2, ..., T_m m ***tabele*** într-o ***bază de date***. Fiecare dintre aceste tabele are o ***structură*** ($T_i = \{C_{i0}, C_{i1}, ...\}$) ce conține ***câmpuri*** (C_{ij}). Pentru a defini ***relații*** ($R \subseteq T_1 \times ... T_m$) între aceste tabele, este necesar ca cel puțin un câmp din fiecare tabelă să suporte o ***relație de ordine strictă*** (nota bene: nu e necesară existența logică a acestei construcții; ea se poate construi şi din structura fizică a informației din tabele, cum ar fi numărul înregistrării). Fie aceste câmpuri C_{i0}. Asta înseamnă că ***valorile*** (v_{i0k}, k=1,...) din înregistrările corespunzătoare acestor câmpuri C_{i0} sunt ***ordonate strict*** ($v_{i01} < v_{i02} < ...$). Nota bene: nu e necesar ca relația de ordonare strictă să fie strict crescătoare, cum nu e necesar ca valorile v_{i01}, v_{i02}, ... să fie stocate în înregistrări consecutive; este necesară doar existența relației de ordine strictă, care să permită referirea individuală a fiecărei valori, şi prin aceasta, identificarea în mod unic a fiecărei înregistrări k: (v_{i0k}, v_{i1k}, ...). Relația R între tabele este în fapt o submulțime a $C_{10} \times C_{20} \times ... C_{m0}$. Reprezentarea a relației R este:

R	C_{10}	...	C_{m0}
r_1	c_{101}	...	c_{m01}
...	...	...	...
r_n	c_{10n}	...	c_{m0n}

În mod uzual, pentru mulțimea $T_1 \times ... T_m$ se folosește noțiunea de univers (U). Elementele universului U se numesc atribute. Câmpurile C_{i0} se notează (pentru simplitate) A_i. Mulțimea valorilor atributelor A_i (v_{i0k}, $k \geq 1$) se notează cu D_i. Elementele relației r_1,..., r_n se numesc tuple şi se notează cu t_1, ..., t_n. Folosind aceste notații, relația R devine:

R	A_1/D_1	...	A_m/D_m
t_1	a_{11}	...	a_{1m}
...	...	...	...
t_n	c_{n1}	...	a_{nm}

Coloanele acestui tablou se identifică prin atributele A_i şi domeniile corespunzătoare D_i, scriind A_i/D_i ($1 \leq i \leq m$). Mulțimea ordonată a atributelor $A = A_1, ..., A_m$ care definesc relația R se numeşte ***schemă relațională***. Facem distincție între schema relațională A şi ***instanțierea*** acesteia (t_1, ... t_n). Convenim să notăm relația R de schemă *A*, sub forma: r(*A*) sau r(A_1, A_2, …, A_m). Dacă luam în considerare tuplul t_i care definește linia i din tabloul R de mai sus, adică $t_i \Leftrightarrow a_{i1}...a_{im}$, convenim ca să folosim aceeaşi notație t_i pentru $t_i = (a_{i1}, ..., a_{im}) \in D_1 \times ... D_m$. Convenim, de asemenea să notăm $t_i[A_j] = a_{ij} \in D_j$, $1 \leq i \leq n$, $1 \leq j \leq m$. De asemenea, dacă avem $K = (A_{j1}, A_{j2}, ..., A_{jk})$, $k \leq m$, atunci $t_i[K] = (a_{ij_1}, a_{ij_2}, ..., a_{ij_k})$.

Recapitulând, principalele concepte utilizate la descrierea logică (conceptuală), respectiv formală, apoi uzuală şi fizică a elementelor de organizare a datelor sunt:

formal	*uzual*	*fizică*
relație	tablou	fişier
tuplu	linie	înregistrare
atribut	coloană	câmp
domeniu	tip de dată	tip de dată

Cu alte cuvinte modelul relațional este caracterizat de:

÷ independența datelor față de hardware şi modul de memorare;

÷ navigarea automată sau un limbaj de nivel înalt neprocedural pentru accesarea datelor;

În loc ca să se proceseze câte o înregistrare, programatorul utilizează limbajul pentru a specifica operații unice care trebuie realizate asupra întregului set de date.

Limbajele de generația a 4-a (4^{th} GLs) sunt mai aproape de limbajul uman ca limbajele de nivel înalt (de generația a 3-a, 3^{th} GLs). Primele dintre acestea sunt FOCUS (Information Builders) SQL (IBM), QBE (Querry by example, IBM), dBASE (succesorul lui SQL).

Necesitatea pentru mai multă flexibilitate şi performanță din partea modelelor de date cum ar fi de a suporta aplicațiile științifice sau inginereşti a făcut ca să se extindă conceptul de model relațional aşa încât intrările în tabele să nu mai fie simple valori ci să poată fi programe, texte, date nestructurate mari în formă binară sau orice alt format solicitat de utilizator. Un alt progres s-a făcut prin încorporarea conceptului de *obiect* devenit esențial în limbajele de programare. În bazele de *date orientate obiect* toate datele sunt obiecte. Obiecte se pot lega între ele printr-o *relație de apartenență* pentru a forma o familie mai largă şi mai diversă de obiecte (în anii '90 au fost lansate primele sisteme de management orientat obiect OODMS). Datele care descriu un transport pot fi stocate, de exemplu, ca familie mai largă care poate conține automobile, vapoare, vagoane, avioane. Clasele de obiecte pot forma *ierarhii* în care obiecte individuale pot moşteni proprietăți de la obiectele situate deasupra în ierarhie. Bazele de date multimedia, în care vocea, muzica şi informația video se stochează împreună cu informațiile de tip text, devin tot mai frecvente şi îşi imprimă trendul în dezvoltarea sistemelor de gestiune a bazelor de date orientate obiect.

O secvență tipică pentru un limbaj 4^{th} GL este:

FIND ALL RECORDS WHERE NAME IS "TUCKER"

SQL (Structured Query Language) este un limbaj standard industrial pentru crearea,

actualizarea şi interogarea sistemelor de management ale bazelor de date relaționale.

Prima versiune standardizată a SQL a apărut în 1986 și conține construcțiile de bază ale limbajului pentru definirea şi manipularea tabelelor de date. O revizie în 1989 a adăugat limbajului extensii pentru integritatea referențială și generalizează constrângerile de integritate. O altă extensie în 1992 furnizează facilități în manipularea schemelor şi administrarea datelor şi de asemenea substanțiale îmbunătățiri în ceea ce privește definirea și manipularea datelor. Dezvoltarea sistemului este în desfășurare pentru a face din acesta un limbaj computațional complet pentru definirea şi managementul obiectelor complexe persistente. Aceasta include generalizarea şi specializarea ierarhiilor, moştenire multiplă, tipuri de dată utilizator, generatoare şi construcții declarative, suport pentru sistemele bazate pe cunoştințe, expresii interogative recursive și instrumente adiționale de administrare a datelor. Include de asemenea tipuri abstracte de date, identificatori de obiecte, metode, moştenire, polimorfism, încapsulare şi toate celelalte facilități care sunt asociate uzual cu managementul datelor de tip obiect.

În prezent, industria bazelor de date reprezintă poate cel mai important segment al industriei de software. Companiile care dețin supremația pe acest segment de piață sunt IBM, Oracle, Informix, Sybase, Teradata (NCR), Microsoft, Borland.

Analiza consistenței în date. Elemente de statistică descriptivă

Măsuri statistice pentru populații şi eşantioane

Tabelul 1. Măsuri statistice pentru caracterizarea variabilelor cantitative

Măsură	Referă	Expresie	Interpretare
Suma valorilor	Un şir de numere	$\Sigma(\cdot)$	-
Numărul de valori		$\|\cdot\|$	-
Valoarea medie		$E(\cdot) = \Sigma(\cdot)/\|\cdot\|$	Valoarea aşteptată
Moment central de ordin k, k>1		$E_k(\cdot) = E((X-E(X))^k)$	-
Media caracteristicii X	O populație	$\mu = \mu(X) = E(X)$	Tendința centrală
Media observabilei Y	Un eşantion	$m = m(Y) = E(Y)$	
Estimatorul mediei caracteristicii X	O populație	$M(Y) = m(Y)$	
Varianța caracteristicii X	O populație	$Var(X) = E((X-\mu)^2)$	Împrăştierea
Deviația standard a caracteristicii X		$\sigma = \sigma(X) = \sqrt{Var(X)}$	Dispersia
Varianța observabilei Y	Un eşantion	$var = var(Y) = E((Y-E(Y))^2)$	Împrăştierea
Deviația standard a observabilei Y		$s = s(Y) = \sqrt{Var(Y)}$	Dispersia
Estimatorul varianței caracteristicii X	O populație	$VAR(Y) = var(Y)\cdot\|Y\|/(\|Y\|-1)$	Împrăştierea
Estimatorul deviației standard a caracteristicii X		$S = S(Y) = s(Y)\cdot\|Y\|/(\|Y\|-1)$	Dispersia

Tabelul 2. Statistici pentru caracterizarea depărtării de normalitate a variabilelor cantitative

Simbol şi măsură	Referă	Expresie	Mărimi care intervin
γ_1, Asimetria caracteristicii X	O populație	$\gamma_1 = \mu_3/\mu_2^{3/2}$	$\mu_k = E_k(X)$, k>1
β_2, Boltirea caracteristicii X		$\beta_2 = \mu_4/\mu_2^2$	
γ_2, Excesul de boltire al caracteristicii X		$\gamma_2 = \beta_2-3$	
g_1, Asimetria observabilei Y	Un eşantion	$g_1 = m_3/m_2^{3/2}$	$m_k = E_k(Y)$, k>1
b_2, Boltirea observabilei Y		$b_2 = m_4/m_2^2$	
g_2, Excesul de boltire al observabilei Y		$g_2 = b_2-3$	
Estimatorul asimetriei caracteristicii X	O populație	$G_1 = \frac{\sqrt{n_Y(n_Y-1)}}{(n_Y-2)} M_3/M_2^{3/2}$	$n_Y = \|Y\|$ $M_k = \frac{n_Y}{n_Y-1} E_k(Y)$, k>1
Estimatorul boltirii caracteristicii X		$B_2 = \frac{(n_Y-1)(n_Y+1)}{(n_Y-2)(n_Y-3)} M_4/M_2^2$	
Estimatorul excesului de boltire a caracteristicii X		$G_2 = B_2 - 3\cdot\frac{(n_Y-1)^2}{(n_Y-2)(n_Y-3)}$	

Extragerea repetată de eşantioane (de volum dat) dintr-o populație face ca valorile obținute să urmeze o distribuție, numită distribuția de eşantionare. Tabelul 3 prezintă rezultatele care se obțin pentru varianța mărimilor statistice prin extragerea repetată de eşantioane dintr-o populație.

Când valorile parametrilor statistici ai populației nu sunt cunoscute, dar se poate face presupunerea că distribuția populației se comportă suficient de bine [12], aceştia pot fi aproximați cu ajutorul estimatorilor acestora (Tabelul 1). Formulele de calcul aproximativ ale mediei şi variației pentru medie şi varianță sunt redate în Tabelul 4. Dacă se pot asuma ipoteze cu privire la distribuția caracteristicii X în populație, atunci se pot obține formule de calcul pentru parametrii statistici (ai populației) şi folosind relațiile din Tabelul 1 estimatorii parametrilor statistici ai populației din măsurătorile (statisticile) efectuate asupra eşantionului.

Tabelul 3. Medii şi variațe ale mediei şi variației observabilei Y ce rezultă din distribuția de eşantionare din populația cu caracteristica X

Mărime şi notație	Valoare
Media mediei, $\mu_{\bar{Y}}$	$\mu_{\bar{Y}} = \mu(m(Y)) = \mu(X)$
Varianța mediei, $\sigma_{\bar{Y}}^2$	$\sigma_{\bar{Y}}^2 = \sigma^2(m(Y)) = \frac{\sigma^2(X)}{n_Y}$
Media variației, $\mu(s^2)$	$\mu(s^2) = \mu(s^2(Y)) = \frac{(n_Y - 1)}{n_Y}\sigma^2(X)$
Varianța variației, $\sigma^2(s^2)$	$\sigma^2(s^2) = \sigma^2(s^2(Y)) = \frac{(n_Y - 1)^2}{n_Y^3}\mu_4(X) - \frac{(n_Y - 1)(n_Y - 3)}{n_Y^3}\mu_2^2(X)$

Tabelul 4. Valori aproximative pentru mediile şi variațele mediei şi variației observabilei Y în ipotezele teoremei limită centrale

Mărime şi notație	Aproximare
Media mediei, $\mu_{\bar{Y}}$	$\mu_{\bar{Y}} \cong m(Y)$
Varianța mediei, $\sigma_{\bar{Y}}^2$	$\sigma_{\bar{Y}}^2 \cong \frac{s^2(Y)}{(n_Y - 1)}$
Media variației, $\mu(s^2)$	$\mu(s^2) \cong s^2(Y)$
Varianța variației, $\sigma^2(s^2)$	$\sigma^2(s^2) \cong \frac{(n_Y - 1)}{n_Y^2}m_4(Y) - \frac{(n_Y - 3)}{n_Y(n_Y - 1)}m_2^2(Y)$

Măsuri statistice pentru legi de distribuție

Tabelele 1-19 dau expresiile unor mărimi statistice (valabile pentru populație) în timp ce expresiile pentru estimatori se pot obține din Tabelul 1 de la 'Măsuri statistice pentru populații şi eşantioane'.

Tabelul 1. Mărimi statistice ale distribuției discrete uniforme

Mărime statistică	Expresie de calcul
Suport	$k \in \{a, a+1, ..., b-1, b\}$
Minim; Maxim	a; b
Funcția de probabilitate	$1/(b - a + 1)$
Funcția de repartiție	$([k] - a + 1)/(b - a + 1)$
Media şi mediana; varianța	$(a + b)/2$; $((b - a + 1)^2 - 1)/12$
Asimetria; excesul de boltire	0; $-\frac{6((b - a + 1)^2 + 1)}{5((b - a + 1)^2 - 1)}$

Tabelul 2. Mărimi statistice ale distribuției discrete Bernoulli

Mărime statistică	Expresie de calcul	
Suport	$k \in \{0,1\}$; $p \in (0,1)$	
Minim; Maxim	0; 1	
Funcția de probabilitate	(1-p), k = 0	p, k = 1
Funcția de repartiție	(1-p), $k \in [0,1)$	1, $1 \leq k$
Media; varianța	p; p(1-p)	
Asimetria; excesul de boltire	0; $\left(6p^2-6p+1\right)/(p(1-p))$	

Tabelul 3. Mărimi statistice ale distribuției discrete binomiale

Mărime statistică	Expresie de calcul
Suport	$k \in \{0, ..., n\}$; $p \in (0,1)$
Minim; Maxim	0; n
Funcția de probabilitate	$\frac{n!}{k!\cdot(n-k)!}p^k(1-p)^{n-k}$
Funcția de repartiție	$\sum_{i=0}^{k}\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}$
Media; varianța	np; np(1-p)
Asimetria; excesul de boltire	$(1-2p)/\sqrt{np(1-p)}$; $\frac{1-6p(1-p)}{np(1-p)}$

Tabelul 4. Mărimi statistice ale distribuției discrete Poisson

Mărime statistică	Expresie de calcul
Suport	$k = 0, 1, ...; \lambda \geq 0$
Minim; Maxim	0; ∞
Funcția de probabilitate	$e^{-\lambda}\lambda^k/k!$
Funcția de repartiție	$\sum_{i=0}^{k}e^{-\lambda}\lambda^i/i!$
Media; varianța	λ; λ
Asimetria; excesul de boltire	$1/\sqrt{\lambda}$; $1/\lambda$

Tabelul 5. Mărimi statistice ale distribuției continue uniforme

Mărime statistică	Expresie de calcul
Suport	$x \in [a, b]$
Minim; Maxim	a; b
Funcția de probabilitate	1/(b-a)
Funcția de repartiție	(x-a)/(b-a)
Media şi mediana; varianța	(a+b)/2; $(b-a)^2/12$
Asimetria; excesul de boltire	0; -6/5

Tabelul 6. Mărimi statistice ale distribuției continue Cauchy-Lorentz

Mărime statistică	Expresie de calcul
Suport	$x \in (-\infty,\infty)$; $x_0 \in (-\infty,\infty)$; $\gamma \in (0,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\frac{1}{\gamma\pi\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)}$
Funcția de repartiție	$\frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}$
Mediana şi moda	x_0

Tabelul 7. Mărimi statistice ale distribuției continue Student t

Mărime statistică	Expresie de calcul
Suport	$x \in (-\infty,\infty)$; $\nu \in (0,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\cdot\sum_{n\geq 0}\frac{\left(-x^2/\nu\right)^n}{n!}\prod_{i=0}^{n-1}\frac{(1+2i)(\nu+1+2i)}{2(3+2i)}$
Media; mediana; moda; varianța	0 ($\nu > 1$); 0; 0; $\nu/(\nu-2)$, $\nu > 2$
Asimetria; excesul de boltire	0, $\nu > 3$; $6/(\nu-4)$, $\nu > 4$

Tabelul 8. Mărimi statistice ale distribuției continue Fisher-Snedecor F

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $d_1,d_2 \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$\frac{\Gamma((d_1+d_2)/2)}{\Gamma(d_1/2)\Gamma(d_2/2)}\frac{(d_1)^{d_1/2}(d_2)^{d_2/2}x^{d_1/2-1}}{(d_1x+d_2)^{(d_1+d_2)/2}}$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$IB\left(\frac{d_1x}{d_1x+d_2},\frac{d_1}{2},\frac{d_2}{2}\right)\Big/IB\left(1,\frac{d_1}{2},\frac{d_2}{2}\right)$, $IB(z,a,b)=\int_0^z t^{a-1}(1-t)^{b-1}dt$
Media; moda	$\frac{d_2}{d_2-2}$, $d_2 > 2$; $\frac{d_1-2}{d_1}\frac{d_2}{d_2+2}$, $d_1 > 2$
Varianța; asimetria	$\frac{2d_2^{\,2}(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$, $d_2 > 4$; $\frac{(2d_1+d_2-2)\sqrt{8(d_2-4)}}{(d_2-6)\sqrt{d_1(d_1+d_2-2)}}$, $d_2 > 6$
Excesul de boltire	$\frac{3d_2^{\,3}+(5d_1-8)d_2^{\,2}+(5d_1^{\,2}-32d_1+20)d_2-22d_1^{\,2}+44d_1-16}{d_1(d_2-6)(d_2-8)(d_1+d_2-2)/12}$, $d_2 > 8$

Tabelul 9. Mărimi statistice ale distribuției continue χ^2

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $d \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$(1/2)^{d/2}x^{d/2-1}e^{-x/2}/\Gamma(d/2)$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\int_0^{x/2} t^{d/2-1}e^{-t}dt \Big/ \Gamma(d/2)$
Media; mediana; moda; varianța	d; $\cong d-2/3$; d - 2, d > 2; 2d
asimetria; excesul de boltire	$\sqrt{8/d}$; $12/d$

Tabelul 10. Mărimi statistice ale distribuției continue exponențiale

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\lambda \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$\lambda e^{-\lambda x}$
Funcția de repartiție	$1-e^{-\lambda x}$
Media; mediana; moda; varianța; asimetria; excesul de boltire	$1/\lambda$; $\ln(2)/\lambda$; 0; $1/\lambda^2$; 2; 6

Tabelul 11. Mărimi statistice ale distribuției continue Weibull

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\lambda, k \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate; funcția de repartiție	$kx^{k-1}e^{-(x/\lambda)^k}/\lambda^k$; $1-e^{-(x/\lambda)^k}$
Media; mediana; moda	$\mu=\lambda\Gamma(1+1/k)$; $\lambda(\ln(2))^{1/k}$; $\lambda((k-1)/k)^{1/k}$, $k>1$
Varianța; asimetria	$\sigma^2=\lambda^2\Gamma(1+2/k)-\mu^2$; $\gamma_1=\left(\Gamma(1+3/k)\lambda^3-3\mu\sigma^2-\mu^3\right)/\sigma^3$
Excesul de boltire	$\gamma_2=\left(\lambda^4\Gamma(1+4/k)-4\gamma_1\sigma^3\mu-6\mu^2\sigma^2-\mu^4\right)/\sigma^4$

Tabelul 12. Mărimi statistice ale distribuției continue Log-normale

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\mu \in (-\infty,\infty)$; $\sigma \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\Big/\left(x\sigma\sqrt{2\pi}\right)$
Funcția de repartiție	$\frac{1+\mathrm{erf}\left((\ln(x)-\mu)/(\sigma\sqrt{2})\right)}{2}$; $\mathrm{erf}(z)=2\int_0^z e^{-t^2}dt\Big/\sqrt{\pi}$
Media; mediana; moda; varianța	$e^{\mu+\sigma^2/2}$; e^{μ} ; $e^{\mu-\sigma^2}$; $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Asimetria; excesul de boltire	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$; $e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$

Tabelul 13. Mărimi statistice ale distribuției continue Birnbaum-Saunders (a vieții obosite)

Mărime statistică	Expresie de calcul
Suport	$\mu, \beta, \gamma \in (0,\infty)$; $x \in (\mu,\infty)$
Minim; Maxim	μ; ∞
Funcția de probabilitate	$\frac{\sqrt{\frac{x-\mu}{\beta}}+\sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)}N_{0,1}\left(\left(\sqrt{\frac{x-\mu}{\beta}}-\sqrt{\frac{\beta}{x-\mu}}\right)\Big/\gamma\right)$
Funcția de probabilitate standard	$\frac{\sqrt{x}+\sqrt{1/x}}{2\gamma(x-\mu)}N_{0,1}\left((\sqrt{x}-\sqrt{1/x})/\gamma\right)$, $N_{0,1}(z)=\int_{-\infty}^{z}\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt$
Funcția de repartiție standard	$N_{0,1}\left((\sqrt{x}-\sqrt{1/x})/\gamma\right)$
Media; varianța (standard)	$1+\gamma^2/2$; $\gamma\sqrt{1+5\gamma^2/4}$

Tabelul 14. Mărimi statistice ale distribuției continue Gamma

Mărime statistică	Expresie de calcul
Suport	$k, \theta \in (0,\infty)$; $x \in [0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$x^{k-1}e^{-x/\theta}\theta^{-k}/\Gamma(k)$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\int_0^{x/\theta}t^{k-1}e^{-t}dt\Big/\int_0^\infty t^{k-1}e^{-t}dt$
Media; moda; varianța	$k\theta$; $(k-1)\theta$, $k>1$; $k\theta^2$
Asimetria; excesul de boltire	$2/\sqrt{k}$; $6/k$

Tabelul 15. Mărimi statistice ale distribuției continue Laplace (dublu exponențială)

Mărime statistică	Expresie de calcul	
Suport	$b \in (0,\infty)$; μ, $x \in (-\infty,\infty)$	
Minim; Maxim	$-\infty$; ∞	
Funcția de probabilitate	$e^{-\lvert x-\mu\rvert/b}/2b$	
Funcția de repartiție	$e^{(x-\mu)/b}/2$, $x < \mu$	$1-e^{-(x-\mu)/b}/2$, $\mu \le x$
Media; mediana; moda; varianța	μ; μ; μ; $2b^2$	
Asimetria; excesul de boltire	0; 3	

Tabelul 16. Mărimi statistice ale distribuției continue Gumbel (log-Weibull)

Mărime statistică	Expresie de calcul
Suport	$\beta \in (0,\infty)$; μ, $x \in (-\infty,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\exp(-\exp(-(x-\mu)/\beta)/\beta)\exp(-(x-\mu)/\beta)/\beta$
Funcția de repartiție	$\exp(-\exp(-(x-\mu)/\beta))$
Media; mediana; moda; varianța	$\mu+\beta\gamma$; $\mu-\beta\cdot\ln(\ln(2))$; μ; $\pi^2\beta^2/6$
Asimetria; excesul de boltire	$\frac{12\sqrt{6}\varsigma(3)}{\pi^3} \cong 1.14$; 12/5

Tabelul 17. Mărimi statistice ale distribuției continue Beta

Mărime statistică	Expresie de calcul
Suport	$\alpha, \beta \in (0,\infty)$; $x \in [0,1]$
Minim; Maxim	0; 1
Funcția de probabilitate	$x^{\alpha-1}(1-x)^{\beta-1}/IB(1,\alpha,\beta)$; $IB(z,a,b)=\int_0^z t^{a-1}(1-t)^{b-1}dt$
Funcția de repartiție	$IB(x,\alpha,\beta)/IB(1,\alpha,\beta)$
Media; moda; varianța	$\frac{\alpha}{\alpha+\beta}$; $\frac{\alpha-1}{\alpha+\beta-2}$, $\alpha,\beta>1$; $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Asimetria; excesul de boltire	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$; $\frac{\alpha^3-(2\beta-1)\alpha^2-2\alpha\beta(\beta+2)+(\beta+1)\beta^2}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)/6}$

Tabelul 18. Mărimi statistice ale distribuției continue Gauss (normale)

Mărime statistică	Expresie de calcul
Suport	$\sigma \in (0,\infty)$; μ, $x \in (-\infty,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\exp\left(-((x-\mu)/\sigma)^2/2\right)/\left(\sigma\sqrt{2\pi}\right)$
Funcția de repartiție	$\left(1+\mathrm{erf}\left((x-\mu)/\left(\sigma\sqrt{2}\right)\right)\right)/2$; $\mathrm{erf}(z)=2\int_0^z e^{-t^2}dt \Big/ \sqrt{\pi}$
Media; moda; varianța	μ; μ; μ; σ^2
Asimetria; excesul de boltire	0; 0

Tabelul 19. Alte mărimi statistice ale distribuției continue Gauss (normale)

Mărime	Populație (finită) de volum n_X	Eșantion de volum n_Y	Estimator
Media	$\mu_{\bar{X}} = \mu$; $\sigma_{\bar{X}}{}^2 = \sigma^2/n_X$	$\mu_{\bar{Y}} = m$; $\sigma_{\bar{Y}}{}^2 = s^2/n_Y$	m; $s^2/(n_Y-1)$
Varianța	$(n_X-1)\sigma^2/n_X$ $\frac{(n_X-1)^2}{n_X{}^3\mu_4{}^{-1}} - \frac{(n_X-1)\mu_2{}^2}{n_X{}^3(n_X-3)^{-1}}$	$(n_Y-1)s^2/n_Y$ $\frac{(n_Y-1)^2}{n_Y{}^3 m_4{}^{-1}} - \frac{(n_Y-1)m_2{}^2}{n_Y{}^3(n_Y-3)^{-1}}$	s^2 $\frac{(n_Y-1)m_4}{n_Y{}^2} - \frac{(n_Y-3)m_2{}^2}{n_Y(n_Y-1)}$ $= \frac{2\sigma^4(n_Y-1)}{n_Y{}^2} \cong \frac{2s^4}{n_Y-1}$
Var γ_1	$\frac{6n_X(n_X-1)}{(n_X-2)(n_X+1)(n_X+3)}$	$\frac{6n_Y(n_Y-1)}{(n_Y-2)(n_Y+1)(n_Y+3)}$	$c_4{}^2\cdot\mathrm{var}(g_1)$ c_4 - Vezi Tabelul 29
Var γ_2	$\frac{24n_X(n_X-1)^2(n_X-3)^{-1}}{(n_X-2)(n_X+3)(n_X+5)}$	$\frac{24n_Y(n_Y-1)^2(n_Y-3)^{-1}}{(n_Y-2)(n_Y+3)(n_Y+5)}$	$c_4{}^2\cdot\mathrm{var}(g_2)$ c_4 - Vezi Tabelul 29

Statistici

Statistica Benford

Testul Benford foloseşte distribuţia Z (normală) pentru a verifica ipoteza că un şir de numere urmează distribuţia Benford, frecvenţele după care se distribuie o anume cifră a fiecărui număr din şir.

Un şir de numere urmează distribuţia Benford dacă probabilitatea de distribuţie a unei cifre (d_i) a numerelor ($d=d_0d_1\ldots$) reprezentate în baza de numeraţie b (uzual baza 10) urmează legea (Benford):

$p(d_0) = \log_b\left(1+\frac{1}{d_0}\right)$, $d_0 = 1..(b\text{-}1)$; $p(d_1) = \sum_{k=1}^{b-1}\log_b\left(1+\frac{1}{k\cdot b+d_1}\right)$, $d_1 = 0..(b\text{-}1)$ $p(d_2) = \sum_{j=1}^{b-1}\sum_{k=0}^{b-1}\log_b\left(1+\frac{1}{j\cdot b^2+k\cdot b+d_2}\right)$, $d_2 = 0..(b\text{-}1)$ …	(Benford)

Ipoteza acestei legi de distribuţie este că valorile măsurătorilor rezultate din observaţie sunt frecvent distribuite logaritmic şi astfel logaritmul setului de măsurători este distribuit uniform. Legea de distribuţie este numită după fizicianul Frank BENFORD care a formulat-o intuitiv în 1938 [13], dar demonstraţia acesteia a fost dată mult mai târziu [14].

Acest rezultat intuitiv de numărare a aparițiilor a fost găsit aplicându-se la o mare varietate de seturi de date incluzând facturile la electricitate, adresele de străzi, preţurile acţiunilor, numerele populaţiei, ratele de deces, lungimile râurilor, constante fizice şi matematice şi procesele descrise de legi putere (care sunt foarte comune în natură). Este foarte important de ştiut că rezultatul (odată observat într-o bază de numeraţie) are loc independent de baza de numeraţie în care se exprimă numerele, chiar dacă proporţiile de reprezentare se schimbă. De aici, ***acest rezultat poate fi folosit pentru a verifica datele în suspiciunea de alterare (mistificare) a acestora prin compararea frecvenţelor teoretice cu cele observate pentru prima cifră a acestora***.

Statistica Jarque-Bera

Testul Jarque-Bera [15,16] calculează şi atribuie probabilitatea statistică ca valorile unui eşantion ce provine din populaţie normal distribuită să îşi abată simultan asimetria şi excesul de boltire de la valorile teoretice corespunzătoare distribuţiei normale.

Statistica Jarque-Bera se calculează cu relaţia:

$$JB = n\frac{g_1^{\,2} + g_2^{\,2}/4}{6}$$

în care g_1 este asimetria, g_2 este excesul de boltire şi n este volumul eşantionului.

Statistica JB are o distribuţie asimptotică către $\chi^2(df=2)$.

g_1, Asimetria observabilei Y	Un eşantion	$g_1 = m_3/m_2^{3/2}$	$m_k = E_k(Y)$, $k>1$
b_2, Boltirea observabilei Y		$b_2 = m_4/m_2^{\,2}$	
g_2, Excesul de boltire al observabilei Y		$g_2 = b_2\text{-}3$	

Statistica Kolmogorov-Smirnov

Testul Kolmogorov-Smirnov [17] poate fi folosit pentru verificarea ipotezei că un eşantion de date urmează o anumită lege de distribuţie (redat în continuare), precum şi pentru compararea legilor de distribuţie ale populaţiilor din care provin două eşantioane [18].

Statistica Kolmogorov-Smirnov verifică dacă observaţiile independente $X=(X_i)_{1\le i\le n}$ provin dintr-o populaţie ce urmează legea de distribuţie dată de funcţia cumulativă de probabilitate $F_t(x)$ prin calcularea maximumului diferenţei absolute între $F_t(x)$ şi funcţia cumulativă de probabilitate observată $F_o(x)$ în toate punctele observaţiei:

$D = \max_{1\le i\le n}\left\|F_t(X_i) - F_o(X_i)\right\|$	(K-S Stat)

Distribuția Kolmogorov

Legea de distribuție Kolmogorov se obține pentru variabila aleatoare K dată de:

$K = \max_{0\le t\le 1}\lvert B(t)\rvert$, unde B este puntea Browniană condiționată de: B(0) = B(1) = 0 M(B(t)) = 0 Var(B(t)) = t(t-1) $P(K \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/8x^2}$	(K-S Dist)

Testul Kolmogorov-Smirnov

Ipoteza testului este că următoarea convergență are loc în distribuție:

$D\sqrt{n} \underset{n\to\infty}{\to} \sup_{t\in[0,1]}\lvert B(F(t))\rvert$ Ipoteza se respinge la nivelul de semnificație dacă: $D\sqrt{n} > K_\alpha$, unde K_α: $P(K \le K_\alpha) = 1-\alpha$	(K-S Test)

Pentru compararea a două distribuții observate:

$D = \max_{1\le i\le \max(n,m)}\lvert F_{o1}(X_i) - F_{o2}(X_i)\rvert$ Ipoteza se respinge la nivelul de semnificație dacă: $D\sqrt{\frac{mn}{m+n}} > K_\alpha$	(K-S Test)

Statistica Anderson-Darling

Testul Anderson-Darling [19] verifică dacă este o evidență statistică ca un eşantion să nu provină dintr-o funcție de probabilitate dată.

Statistica Anderson verifică dacă asupra observațiilor distincte ordonate crescător $(X_i)_{1\le i\le n}$, $X_i<X_{i+1}$ se poate respinge ipoteza că provin dintr-o distribuție dată de funcția cumulativă de probabilitate F calculând valoarea A dată de relația:

$$A^2 = -n - \sum_{k=1}^{n}\frac{2k-1}{n}\left(\ln(F(Y_k)) + \ln(1-F(Y_{n+1-k}))\right)$$

O aplicație de interes însă o reprezintă testul Anderson-Darling pentru mai multe eşantioane asupra cărora se poate verifica proveniența din aceeaşi populație, caz în care legea de distribuție a populației nu mai trebuie să fie specificată [20, 21]. Formulele de calcul şi interpretarea testului pentru compararea de eşantioane se găsesc la adresa [22].

În cazul comparației unei legi de distribuție discrete cunoscute cu legea de distribuție observată în eşantion varianța statisticii A^2 se calculează cu formula ([23], n - numărul de observații din eşantion, π=3.14159265358979323846264338...):

$$Var(A^2) = \frac{2(\pi^2-9)}{3} + \frac{10-\pi^2}{n}$$

În cazul verificării ipotezei de normalitate, este posibil să se aproximeze probabilitatea de observație asociată valorii statisticii A^2 [24]. Se aplică corecția de volum al eşantionului:

$$A^2{}_c = A^2(1 + 0.75/n + 2.25/n^2)$$

$$p = \begin{cases} 1-\exp\left(-13.436 + 101.14\cdot x - 223.73\cdot x^2\right), & x < 0.2 \\ 1-\exp\left(-8.318 + 42.796\cdot x - 59.938\cdot x^2\right), & 0.2 \le x < 0.34 \\ \exp\left(0.9177 - 4.279\cdot x - 1.38\cdot x^2\right), & 0.34 \le x < 0.6 \\ \exp\left(1.2937 - 5.709\cdot x + 0.0186\cdot x^2\right), & x \le 0.6 \end{cases}$$

Statistica Pearson-Fisher Chi Square

Distribuția χ^2 a fost descoperită de Karl PEARSON [25] în urma încercării de a explica varianța observată a numerelor care provin din distribuția normală.

Astfel, dacă se consideră distribuția normală standard N(0,1) şi variabila întâmplătoare X ce urmează această distribuție (Figura 1), probabilitatea (dp) de a extrage valorile -x şi x din N(0,1) sunt ambele egale şi egale cu diferențiala funcției de densitate de probabilitate a distribuției normale ($PDF_{N(0,1)}$).

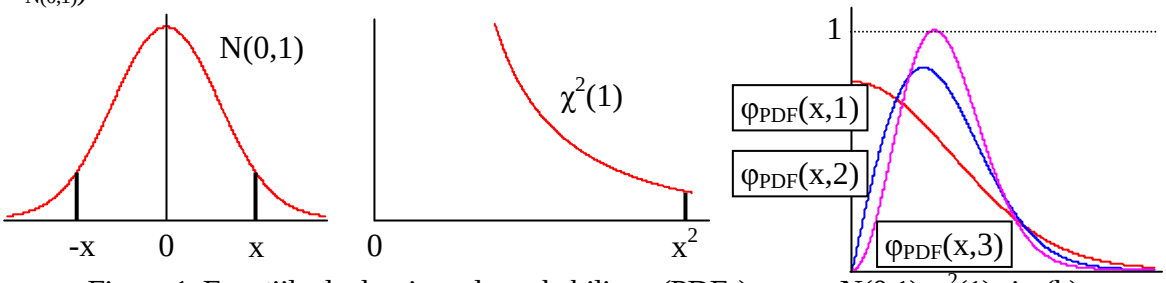


Figura 1. Funcțiile de densitate de probabilitate (PDFs) pentru N(0,1), $\chi^2(1)$ şi φ(k)

$$PDF_{N(0,1)}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{1}$$

Distribuția normală standard are media 0; astfel, pentru a exprima probabilitatea de observație pentru deviația x^2 trebuie adunate două probabilități (pentru -x şi x) date de relația (1):

$$dp(x^2) = 2 \cdot dPDF_{N(0,1)}(x) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{2}$$

Pentru a reconstitui PDF pentru x^2 trebuie să efectuăm o schimbare de variabilă $x^2 = t$; atunci $x = \sqrt{t}$ şi:

$$dp(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t}{2}\right) d\sqrt{t} = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t}{2}\right) \frac{1}{2\sqrt{t}} dt \tag{3}$$

Este uşor de verificat că (3) este un caz particular al lui (4) când k = 1:

$$\chi^2{}_{PDF}(t,k) = \frac{1}{2^{k/2}\Gamma(k/2)} t^{k/2-1} \exp\left(-\frac{t}{2}\right) \tag{4}$$

Procedura descrisă mai sus corespunde pentru distribuția Chi Square cu un grad de libertate (extragerea lui X din distribuția normală). Dacă sunt extrase mai multe valori (k valori) din distribuția normală atunci se obține distribuția Chi Square cu k grade de libertate, şi demonstrația că ecuația (4) este adevărată poate fi găsită în [26].

Calea directă de la distribuția normală la distribuția χ^2 nu este reversibilă (Figura 1); astfel, definind variabila φ ca în relația (5) - ce reprezintă o expresie modificată a coeficientului de asociere definit de LIEBETRAU [27]:

$$\varphi = \varphi(X^2, k) = \sqrt{\frac{X^2}{k}} \tag{5}$$

obținerea distribuției lui φ se poate obține pe o cale similară cu cea descrisă mai sus; notând u=φ în (5) şi substituind $t = X^2 = ku^2$ în (4) se obține ($du^2 = 2u \cdot du$):

$$d\chi^2(ku^2, k) = \frac{1}{2^{k/2}\Gamma(k/2)} (ku^2)^{k/2-1} \exp\left(-\frac{ku^2}{2}\right) d(ku^2) \tag{6}$$

După rearanjarea termenilor:

$$d\varphi_{PDF}(u,k) = \frac{2u^{k-1}}{\Gamma(k/2)} \left(\frac{k}{2}\right)^{k/2} \exp\left(-\frac{u^2}{2/k}\right) du \tag{7}$$

Pornind de la densitatea de probabilitate (PDF) a distribuției Gamma:

$$\Gamma_{PDF}(x;a,b,c)=\frac{cx^{ca-1}}{b^{ca}\Gamma(a)}\exp\left(-(x/b)^{c}\right) \quad (8)$$

este uşor de verificat că:

$$\varphi_{PDF}(x,k)=\Gamma_{PDF}(x,\frac{k}{2},\sqrt{\frac{2}{k}},2) \quad (9)$$

Relația (9) demonstrează că distribuția lui $\sqrt{\frac{X^2}{k}}$ este un caz particular al distribuției Gamma (Figura 1).

Testul χ^2 ca măsură a independenței, omogenității şi asocierii în distribuție

Distribuția χ^2 are 3 aplicații imediate:

÷ Testul Chi Square pentru verificarea independenței
- o testează asocierea între două variabile cu valori grupate pe categorii;
- o se poate aplica dacă au loc două condiții:
  - nici una din valorile aşteptate nu este mai mică decât 1;
  - nu mai mult de 20% din valorile aşteptate nu sunt mai mici de 5;
- o ipotezele de lucru sunt: nu există nici o asociere între cele două variabile (ipoteza nulă) şi este o asociere între cele două variabile (ipoteza contrară);
- o Când statistica Chi Square (X^2) este mai mare decâ valoarea funcției cumulative de probabilitate a distribuției Chi Square (χ^2) pentru numărul de grade de libaertate egal cu numărul de cazuri minus unu şi pentru riscul de a fi în eroare (nivelul de semnificație) ales, atunci există o diferență semnificativă de la ipoteza lipsei de asociere şi cele două variabile sunt asociate;

÷ Testul Chi Square pentru verificarea omogenității
- o testează dacă mai multe populații sunt similare (sau omogene sau egale) în anumite caracteristici (acele caracteristici care sunt incluse în testare);
- o ipotezele de lucru sunt: populațiile sunt similare (sau omogene sau egale) în caracteristica supusă observației (ipoteza nulă) şi populațiile sunt diferite în caracteristică (ipoteza contrară);
- o uzual caracteristica supusă observației este un moment central (ex. valoare medie, varianță);

÷ Testul Chi Square pentru verificarea asocierii în distribuție
- o testează dacă un model teoretic poate fi asociat observațiilor;
- o ipotezele de lucru sunt: datele observate urmează distribuția dată de modelul teoretic (ipoteza nulă) şi datele observate nu provin dintr-o populație ce urmează modelul teoretic (ipoteza contrară);

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a asocierii în distribuție

Testul χ^2, propus ca măsură a depărtării întâmplătoare între observație şi modelul teoretic de Karl PEARSON [25] a fost corectat în interpretare de Ronald FISHER prin reducerea numărului de grade de libertate corespunzător cu o unitate (datorită estimării frecvenței teoretice din frecvența observată, [28]), şi cu numărul parametrilor necunoscuți ai distribuției teoretice estimați din observații din măsuri ale tendinței centrale ([29]).

Testarea agrementului între observație şi ipoteză se realizează prin divizarea observațiilor într-un număr definit de intervale (n), pentru care se calculează expresia X^2 (unde s este numărul de parametrii ai distribuției teoretice estimați din momente centrale, O_i este frecvența experimentală observată în clasa de frecvență i, E_i este frecvența aşteptată calculată din legea de distribuție teoretică pentru clasa de frecvență i, X^2 este valoarea statisticii chi square iar χ^2 este valoarea parametrului statistic chi square din distribuția cu acelaşi nume):

$$X^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}\approx\chi^2(n-s-1) \quad (10)$$

Pe baza distribuției teoretice χ^2 se calculează probabilitatea de respingere a ipotezei de agrement. Uzual ipoteza de agrement este acceptată dacă probabilitatea de respingere a ipotezei de agrement ($\chi^2_{CDF}(X^2,n\text{-}s\text{-}1)$) este mai mică de 5%.

În ciuda faptului că testul χ^2 este cea mai cunoscută statistică pentru verificarea agrementului între observație și ipoteză, testarea independenței și a omogenității, definirea cadrului de aplicare al acesteia este dintre cele mai complexe [30].

O serie de probleme la compararea unei distribuții observate cu o distribuție teoretică apar în calcularea statisticii X^2 și în aplicarea testului χ^2.

O primă problemă este alegerea numărului de clase de frecvență și există mai multe soluții, dintre care două sunt:

÷ calcularea prin rotunjire a numărului de clase de frecvență din entropia Hartley [31] a observației vs. expectație: $\log_2(2N)$, unde N este numărul de observații (EasyFit [32] folosește această procedură);

÷ calcularea numărului de clase de frecvență odată cu lărgimea clasei folosind histograma ca estimator al densității [33] și alegerea pe baza acesteia a criteriului optimal pentru lărgimea clasei (Dataplot [34] generează automat clasele de frecvență folosind această regulă: lărgimea clasei de frecvență este $0.3 \cdot s$ unde s este deviația standard a eșantionului; limitele inferioară și superioară sunt date de medie $\pm 6 \cdot s$ și clasele de frecvență observată 0 marginale sunt omise;

O a doua problemă este lărgimea claselor de frecvență; și aici există cel puțin două abordări:

÷ Datele pot fi grupate în clase de frecvență de probabilitate (teoretică sau observată) egală;

÷ Datele pot fi grupate în intervale de lărgime egală;

Prima abordare (probabilitatea egală) este mai frecvent adoptată deoarece este o soluție mai bună pentru observații foarte grupate.

O altă problemă este numărul de observații din interiorul fiecărei clase de frecvență. Fiecare clasă de frecvență trebuie să conțină cel puțin 5 observații, astfel încât în practică clase de frecvență alăturate se reunesc pentru a satisface această impunere.

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a omogenității

Statistica Chi Square operează în ipoteza în care o observabilă este rezultatul suprapunerii a doi (sau mai mulți, dintre care pentru doi dintre aceștia se realizează un experiment) factori. În acest caz se constituie un experiment menit să verifice dacă se poate accepta independența între acești doi factori. Se construiește un tabel de contingență format din linii (reprezentând valorile primului factor) și coloane (reprezentând valorile celui de-al doilea factor) în care se cumulează frecvențele sau valorile medii ale variabilei observate și în care ipoteza independenței factorilor se translatează în ipoteza omogenității valorilor înregistrate în tabel.

Valoarea statisticii X^2 se calculează cu formula (unde $1 \le i \le r$ reprezintă indicii observațiilor asociate primului factor, $1 \le j \le c$ reprezintă indicii observațiilor asociate celui de-al doilea factor, $O_{i,j}$ reprezintă valori medii (pentru testul de omogenitate) sau frecvențe (pentru testul de independență) observate pentru perechea (i,j) de valori ale factorilor, $E_{i,j}$ este valoarea medie (pentru testul de omogenitate) sau frecvența (pentru testul de independență) așteptată pentru perechea (i,j) de valori ale factorilor, X^2 este valoarea statisticii chi square iar χ^2 este valoarea parametrului statistic chi square din distribuția cu același nume):

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \approx \chi^2((r-1)(c-1)) \qquad (11)$$

Testarea individuală a omogenității valorilor dintr-o clasă (linie sau coloană în tabel) și în același timp crearea unei ierarhii a iregularităților se obține descompunând expresia lui X^2 în:

$$X^2{}_c = \sum_{i=1}^{r}\frac{(O_{i,c}-E_{i,c})^2}{E_{i,c}} \approx \chi^2(r-1)\,;\; X^2{}_r = \sum_{j=1}^{c}\frac{(O_{r,j}-E_{r,j})^2}{E_{r,j}} \approx \chi^2(c-1) \qquad (12)$$

Presupunerea naturală este că observațiile $O_{i,j}$ sunt rezultatul multiplicării celor doi factori, ceea ce face ca observațiile repetate să aproximeze tot mai bine efectul de multiplicare, și de aici rezultă o formulă de exprimare pentru frecvența așteptată $E_{i,j}$ [35]:

$$E_{i,j} = \sum_{k=1}^{r} O_{i,k} \sum_{k=1}^{c} O_{k,j} \Big/ \sum_{i=1}^{r}\sum_{j=1}^{c} O_{i,j} \qquad (13)$$

În acelaşi cadru al presupunerii naturale al efectului multiplicativ al celor doi factori asupra observabilei O din punct de vedere matematic se pot formula trei presupuneri cu privire la eroarea pătratică $(O_{i,j}-E_{i,j})^2$ produsă de observație:

÷ măsurătoarea este afectată de erori absolute întâmplătoare;
÷ măsurătoarea este afectată de erori relative întâmplătoare;
÷ măsurătoarea este afectată de erori întâmplătoare pe o scară intermediară între erori absolute şi erori relative;

Prima dintre ipoteze (erori absolute întâmplătoare) conduce din punct de vedere matematic la minimizarea varianței între model şi observație (relația 14), a doua dintre ipoteze conduce la minimizarea pătratului coeficientului de variație (relația 15) iar o soluție (una din mai multe soluții posibile) la cea de-a treia dintre ipoteze o reprezintă minimizarea statisticii X^2 (relația 16).

$$S^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(O_{i,j} - a_i b_j)^2 \qquad = \text{min. (14)}$$

$$CV^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - a_i b_j)^2}{(a_i b_j)^2} \qquad = \text{min. (15)}$$

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - a_i b_j)^2}{a_i b_j} \qquad = \text{min. (16)}$$

În relațiile (14)-(16) apar exprimați cei doi factori a căror independență se verifică prin intermediul efectului multiplicativ (a_i, $1\le i\le r$ reprezintă contribuția primului factor la valoarea aşteptată $E_{i,j}$ iar b_j, $1\le j\le c$ reprezintă contribuția celui de-al doilea factor la valoarea aşteptată $E_{i,j}$ şi expresia valorii aşteptate $E_{i,j}$ este dată, aşa cum presupunerea naturală a fost făcută de produsul celor două contribuții: $E_{i,j}=a_i \cdot b_j$).

Minimizarea cantităților date de relațiile (14)-(16) în scopul determinării contribuțiilor factorilor A ($A=(a_i)_{1\le i\le r}$) şi B ($B=(b_j)_{1\le j\le c}$) se face pe aceeaşi cale, dată generic de relația (17):

$$\left(\frac{\partial\cdot(a_i,b_j)}{\partial a_i}=0\right)_{1\le i\le r};\quad \left(\frac{\partial\cdot(a_i,b_j)}{\partial b_j}=0\right)_{1\le j\le c} \qquad (17)$$

unde expresia de derivat $\cdot(a_i,b_j)$ este una din expresiile S^2, CV^2 şi X^2 date de relațiile (14)-(16). În urma calculului se obține că relația (14) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (18), relația (15) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (19), iar relația (16) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (20):

$$a_i = \sum_{j=1}^{c} b_j O_{i,j} \Big/ \sum_{j=1}^{c} b_j^{\,2},\ i=1..r;\quad b_j = \sum_{i=1}^{r} a_i O_{i,j} \Big/ \sum_{i=1}^{r} a_i^{\,2},\ j=1..c \qquad (18)$$

$$a_i = \sum_{j=1}^{c} \frac{O_{i,j}^{\,2}}{b_j^{\,2}} \Big/ \sum_{j=1}^{c} \frac{O_{i,j}}{b_j},\ i=1..r;\ b_j = \sum_{i=1}^{r} \frac{O_{i,j}^{\,2}}{a_i^{\,2}} \Big/ \sum_{i=1}^{r} \frac{O_{i,j}}{a_i},\ j=1..c \qquad (19)$$

$$a_i^{\,2} = \sum_{j=1}^{c} \frac{O_{i,j}^{\,2}}{b_j} \Big/ \sum_{j=1}^{c} b_j,\ i=1..r;\ b_j^{\,2} = \sum_{i=1}^{r} \frac{O_{i,j}^{\,2}}{a_i} \Big/ \sum_{i=1}^{r} a_i,\ j=1..c \qquad (20)$$

Se poate de asemenea arăta matematic că relațiile (18)-(20) admit o infinitate de soluții şi că familiile de soluții ale relațiilor (18)-(20) se află în vecinătatea familiei de soluții date de relația (13), re-scrisă aici ca relația (21), exprimând explicit cei doi factori A şi B:

$$a_i \cdot b_j = \sum_{k=1}^{r} O_{i,k} \sum_{k=1}^{c} O_{k,j} \Big/ \sum_{i=1}^{r}\sum_{j=1}^{c} O_{i,j} \qquad (21)$$

Calea directă de rezolvare a ecuațiilor (18)-(20) fără a face apel la ecuația (21) este ineficientă. De exemplu pentru r=2, c=3 substituțiile în relația (18) duc la:

$$\left(\frac{a_2}{a_1}\right)^2 + \frac{(O_{1,1}^{\,2}+O_{1,2}^{\,2}+O_{1,3}^{\,2})-(O_{2,1}^{\,2}+O_{2,2}^{\,2}+O_{2,3}^{\,2})}{(O_{1,1}O_{2,1}+O_{1,2}O_{2,2}+O_{1,3}O_{2,3})}\left(\frac{a_2}{a_1}\right)-1=0 \qquad (22)$$

care este rezolvabilă în (a_2/a_1) care dovedeşte că există o infinitate de soluții (pentru orice valoare

nenulă a lui a_1 există o valoare a_2 care să verifice ecuația 22) și gradul ecuației (22) este dat de min(r,c). Ecuațiile ce se obțin pe calea substituției directe devin din ce în ce mai complicate cu creșterea lui `r` și `c` și cu coborârea dinspre relația (18) către relația (20). Astfel, de exemplu pentru același r=2 și c=3 substituțiile în relația (20) conduc la:

$$\begin{aligned} &O_{1,1}^{\ 2}O_{1,2}^{\ 2}(O_{1,1}^{\ 2}-O_{1,2}^{\ 2})(a_2/a_1)^5+(O_{1,1}^{\ 4}O_{2,2}^{\ 2}-O_{1,2}^{\ 4}O_{2,1}^{\ 2})(a_2/a_1)^4+\\ &+2O_{1,1}^{\ 2}O_{1,2}^{\ 2}(O_{2,2}^{\ 2}-O_{2,1}^{\ 2})(a_2/a_1)^3+2O_{2,1}^{\ 2}O_{2,2}^{\ 2}(O_{1,2}^{\ 2}-O_{1,1}^{\ 2})(a_2/a_1)^2 \qquad (23)\\ &+(O_{1,2}^{\ 2}O_{2,1}^{\ 4}-O_{1,1}^{\ 2}O_{2,2}^{\ 4})(a_2/a_1)+O_{2,2}^{\ 2}O_{2,1}^{\ 2}(O_{2,1}^{\ 2}-O_{2,2}^{\ 2})=0 \end{aligned}$$

care este o ecuație de gradul 5 (r+c).

Calea indirectă de rezolvare a relațiilor (18)-(20) este prin aproximații succesive făcând apel la soluția aproximativă oferită de (21). Astfel, se foloseşte relația (21) pentru a obține prima aproximație (aproximația inițială) a soluției după care în fiecare succesiune de aproximații se înlocuiesc vechile valori ale aproximației în partea dreaptă a relațiilor (18)-(20) pentru a obține noile aproximații.

Metoda aproximațiilor succesive converge rapid către soluția optimală. Astfel pentru relația (18) trei iterații sunt suficiente pentru a obține (vezi Tabelul 1) o valoare reziduală de 282.11735 și de la această iterație încolo valoarea reziduală își schimbă cifrele dincolo de a 5-a zecimală, în timp ce pentru relația (20) aceeași calitate a reprezentării soluției optimale este obținută după 4 iterații.

Folosind datele din [35] redate în Tabelul 1, valorile sugerate de ecuațiile (21) pentru produsele $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ sunt redate în Tabelul 2, valorile ce rezultă după rezolvarea iterativă a relațiilor (18)-(20) sunt redate în Tabelele 3-5, în timp ce Tabelul 6 centralizează rezultatele obținute pe cele 4 căi.

Tabelul 1. Valori experimentale în tratamentul cartofilor

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY	Suma
DS	25.3	28	23.3	20	22.9	20.8	22.3	21.9	18.3	14.7	13.8	10	**241.3**
DC	26	27	24.4	19	20.6	24.4	16.8	20.9	20.3	15.6	11	11.8	**237.8**
DB	26.5	23.8	14.2	20	20.1	21.8	21.7	20.6	16	14.3	11.1	13.3	**223.4**
US	23	20.4	18.2	20.2	15.8	15.8	12.7	12.8	11.8	12.5	12.5	8.2	**183.9**
UC	18.5	17	20.8	18.1	17.5	14.4	19.6	13.7	13	12	12.7	8.3	**185.6**
UB	9.5	6.5	4.9	7.7	4.4	2.3	4.2	6.6	1.6	2.2	2.2	1.6	**53.7**
Suma	**128.8**	**122.7**	**105.8**	**105**	**101.3**	**99.5**	**97.3**	**96.5**	**81**	**71.3**	**63.3**	**53.2**	**1125.7**

Legendă:

÷ T_V: Tratament vs. Varietate

÷ UD, KK, KP, TP, ID, GS, AJ, BQ, ND, EP, AC, DY: varietăți de cartofi (UD: Up to Date; KK: K of K; KP: Kerr's Pink; TP: Tinwald Perfection; ID: Iron Duke; GS: Great Scott; AJ: Ajax; BQ: British Queen; ND: Nithsdale; EP: Epicure; AC: Arran Comrade; DY: Duke of York)

÷ DS, DC, DB, US, UC, UB: tratamente (D* - cu fertilizant natural; U* - fără; S - sol fertilizat cu sulfat; C - sol fertilizat cu cloruri; B - sol fertilizat cu baze)

Tabelul 2. Valorile produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ calculate cu relația (21)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.61	26.30	22.68	22.51	21.71	21.33	20.86	20.69	17.36	15.28	13.57	11.40
DC	27.21	25.92	22.35	22.18	21.40	21.02	20.55	20.39	17.11	15.06	13.37	11.24
DB	25.56	24.35	21.00	20.84	20.10	19.75	19.31	19.15	16.07	14.15	12.56	10.56
US	21.04	20.04	17.28	17.15	16.55	16.25	15.90	15.76	13.23	11.65	10.34	8.69
UC	21.24	20.23	17.44	17.31	16.70	16.41	16.04	15.91	13.35	11.76	10.44	8.77
UB	6.14	5.85	5.05	5.01	4.83	4.75	4.64	4.60	3.86	3.40	3.02	2.54

Tabelul 3. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (18)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.07	26.42	22.64	21.85	21.85	21.94	20.94	20.63	17.93	15.48	13.54	11.61
DC	26.66	26.02	22.29	21.52	21.52	21.60	20.62	20.32	17.66	15.24	13.33	11.43
DB	24.91	24.32	20.83	20.11	20.11	20.19	19.27	18.99	16.50	14.25	12.46	10.69
US	20.64	20.15	17.26	16.66	16.66	16.73	15.96	15.73	13.67	11.80	10.32	8.85
UC	20.58	20.09	17.21	16.61	16.61	16.68	15.92	15.69	13.63	11.77	10.29	8.83
UB	6.29	6.14	5.26	5.08	5.08	5.10	4.86	4.79	4.17	3.60	3.14	2.70

Tabelul 4. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (19)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.57	26.08	23.04	22.61	21.48	21.61	21.13	20.69	17.66	15.23	13.79	11.56
DC	27.38	25.9	22.88	22.45	21.34	21.46	20.99	20.55	17.54	15.13	13.69	11.48
DB	25.84	24.44	21.59	21.19	20.14	20.26	19.8	19.4	16.56	14.28	12.92	10.83
US	21.23	20.08	17.74	17.4	16.54	16.64	16.27	15.93	13.6	11.73	10.62	8.9
UC	21.47	20.31	17.94	17.61	16.73	16.83	16.46	16.12	13.76	11.86	10.74	9
UB	7.02	6.64	5.87	5.76	5.47	5.51	5.38	5.27	4.5	3.88	3.51	2.94

Tabelul 5. Valorile optimizate ale produselor $(a_i b_j)_{1\le i\le 6; 1\le j\le 12}$ folosind relațiile (20)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.64	26.19	22.85	22.60	21.59	21.44	20.98	20.71	17.49	15.24	13.67	11.47
DC	27.35	25.91	22.61	22.36	21.36	21.22	20.76	20.50	17.30	15.08	13.52	11.35
DB	25.74	24.40	21.28	21.05	20.11	19.97	19.55	19.29	16.29	14.20	12.73	10.68
US	21.17	20.06	17.50	17.31	16.53	16.42	16.07	15.87	13.39	11.68	10.47	8.78
UC	21.40	20.28	17.69	17.50	16.71	16.60	16.25	16.04	13.54	11.80	10.58	8.88
UB	6.57	6.23	5.43	5.37	5.13	5.10	4.99	4.93	4.16	3.63	3.25	2.73

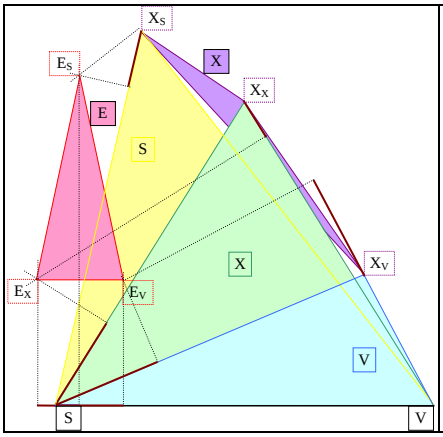
După cum se observă în Tabelul 6, fiecare dintre metodele definite de relațiile (18)-(20) îmbunătățește valoarea sumei obiectiv în raport cu expresia definită de formula aproximativă (21) și reprezintă corecții ale acesteia. Astfel, relația (18) îmbunătățește soluția propusă de relația (21) în ipoteza erorii experimentale uniform distribuite între clase (eroarea experimentală absolută), relația (19) îmbunătățește soluția propusă de relația (21) în ipoteza erorii experimentale proporționale cu magnitudinea fenomenului observat (eroarea experimentală relativă) în timp ce relația (20) îmbunătățește soluția propusă de relația (21) minimizând statistica Pearson-Fisher X^2 (a cărei expresie este o Pearsoniană de tipul III [28]).

Tabelul 6. Valori comparative pentru eroarea experimentală întâmplătoare

	S^2				X^2				CV^2			
Cat	eq(21)	eq(18)	eq(20)	eq(19)	eq(21)	eq(18)	eq(20)	eq(19)	eq(21)	eq(18)	eq(20)	eq(19)
DS	23.4	18.76	24.12	57.97	1.10	0.937	1.127	2.308	0.056	0.0515	0.0573	0.0971
DC	59.7	48.48	59.86	104.95	3.08	2.497	3.052	4.847	0.164	0.133	0.1611	0.2365
DB	69.8	66.77	71.47	95.21	3.78	3.596	3.796	4.803	0.221	0.2078	0.2167	0.2633
US	41.6	49.03	41.66	35.34	2.72	3.19	2.709	2.358	0.186	0.2158	0.183	0.1635
UC	57.6	59.01	56.53	82.16	3.46	3.66	3.339	4.367	0.218	0.2375	0.2065	0.2444
UB	37.5	40.1	37.13	28.26	7.89	8.295	7.659	5.956	1.751	1.8018	1.6696	1.3512
UD	30.3	26.3	28.2	78.9	2.66	2.35	2.15	3.58	0.335	0.293	0.235	0.232
KK	15.3	13.5	15.8	18.7	0.76	0.64	0.73	0.88	0.045	0.033	0.035	0.044
KP	63	62.7	64	67.5	3.11	3.15	3.13	3.19	0.155	0.162	0.159	0.155
TP	34.3	31.4	33.3	76.5	2.79	2.69	2.37	3.67	0.357	0.340	0.256	0.242
ID	3.4	3.9	4	4.5	0.21	0.27	0.28	0.26	0.017	0.028	0.029	0.021
GS	26.2	25.6	26.9	28.6	2.29	2.45	2.52	2.42	0.319	0.349	0.352	0.327
AJ	45	47	45.3	43.4	2.56	2.71	2.60	2.44	0.152	0.168	0.164	0.148
BQ	21.5	20.4	21	31.8	1.93	1.71	1.67	2.19	0.253	0.205	0.182	0.193
ND	18.3	17.9	19.1	20.5	2.13	2.29	2.35	2.27	0.393	0.424	0.427	0.403
EP	2.9	3.2	3.3	3.8	0.53	0.64	0.66	0.62	0.133	0.158	0.163	0.142
AC	18.2	18.8	18.7	19.3	1.76	1.87	1.84	1.83	0.209	0.232	0.233	0.221
DY	11.1	11.5	11.2	10.6	1.31	1.40	1.39	1.27	0.228	0.255	0.258	0.227
Σ	289.5	**282.2**	290.8	404.1	22.04	22.17	**21.69**	24.62	2.596	2.647	2.493	**2.355**

Valorile obținute în Tabelul 6 pentru eroarea experimentală în cele 3 forme ale sale (pătratică absolută S^2, pătratică relativă CV^2, și Pearsoniană X^2) pentru cele 4 cazuri (frecvență teoretică estimată din contingență - eq. 21, frecvență teoretică estimată din minimizarea erorii pătratice absolute - eq. 18, frecvență teoretică estimată din minimizarea erorii pătratice relative - eq. 19, frecvență teoretică estimată din minimizarea statisticii Pearson-Fisher - eq. 20 sunt valori obținute într-un design de experiment în care există exact doi factori independenți (tip tratament și tip sol sau factor A și factor B) ceea ce permite o reprezentare în plan a distanțelor Euclidiene între rezultate.

În Figura 2 au fost reprezentate distanțele Euclidiene între erorile experimentale estimate de fiecare formulă (18)-(20) folosind triunghiuri Snyder [36] (diagrame frecvent folosite în cromatografie pentru a reprezenta 3 sau mai mulți parametrii ce depind de doi factori).



Legendă:
S - coordonata erorii experimentale în ipoteza S^2=min. (ec. 9);
V - coordonata erorii experimentale în ipoteza CV^2=min. (ec. 10);
X_S - coordonata erorii experimentale pătratice absolute (S^2) în ipoteza X^2=min. (ec.11);
X_V - coordonata erorii experimentale pătratice relative (CV^2) în ipoteza X^2=min. (ec.11);
X_X - coordonata statisticii X^2 în ipoteza X^2=min. (ec.11);
E_S - coordonata erorii experimentale pătratice absolute (S^2) în ipoteza contingenței (ec.4);
E_V - coordonata erorii experimentale pătratice relative (CV^2) în ipoteza contingenței (ec.4);
E_X - coordonata statisticii Pearson-Fisher (X^2) în ipoteza contingenței (ec.4);
S - triunghiul erorilor pătratice absolute (S^2); V - triunghiul erorilor pătratice relative (CV^2);
X - triunghiul statisticilor X^2; E - triunghiul de contingență; X - triunghiul de variație a statisticii X^2;

Figura 2. Distanțe Euclidiene între estimările erorilor experimentale

Figura 2 a fost realizată impunând reprezentarea la aceeași scară a ariei de eroare în raport cu cei doi factori (prin fixarea distanței dintre coordonata erorii experimentale în ipoteza S^2 = min. şi coordonata erorii experimentale în ipoteza CV^2 = min.) când coordonata în ipoteza X^2=min. s-a obținut prin maximizarea ariei de eroare (maximizarea ariilor triunghiurilor S, V şi X). Coordonatele contingenței s-au obținut astfel încât proiecțiile contingenței pe laturile triunghiurilor să împartă laturile în rapoartele observate între diferențele din Tabelul 6.

Construcția din FiguraA1F02 2 permite aprecieri calitative cu privire la modelul de contingență definit de ec. (21) şi la statistica Pearson-Fisher în raport cu natura erorii experimentale. Astfel, se observă (în Figura 2) că singura intersecție între aria de contingență şi ariile de eroare se realizează cu eroarea pătratică absolută, deci contingența definită de ecuația (21) asigură agrementul între observație şi model numai pentru acest tip de erori din cele 3 cuprinse în studiu. De asemenea, singura intersecție a triunghiului de variație a statisticii X^2 este cu triunghiul statisticii X^2 ceea pe de o parte recomandă folosirea optimizării definite de ec. (14) [35] sau de ec. (16) [29] şi pe de altă parte demonstrează de ce testul Chi Square este mai expus [37] decât alte teste cum ar fi Kolmogorov-Smirnov ([38, 39]) şi Anderson-Darling ([40, 41]) la erori de tip I respingând ipoteza nulă că variabila linie nu este în relație cu variabila coloană (asocierea este întâmplătoare) chiar când de fapt ipoteza este adevărată.

Se poate reprezenta poziția relativă a soluției propuse de relația (21) în raport cu valorile optimale propuse de relațiile (18)-(20). Pentru aceasta datele din Tabelul 6 au fost transformate cum arată Tabelul 7.

Tabelul 7. Transformarea valorilor reziduale din Tabelul 6 în valori relative la minim

Valori absolute	S^2	X^2	CV^2
E	289.5	22.04	2.596
S^2=min.	**282.2**	22.17	2.647
X^2=min.	290.8	**21.69**	2.493
CV^2=min.	404.1	24.62	**2.355**
Valori relative	S^2	X^2	CV^2
E	1.026	1.016	1.102
S^2=min.	**1**	1.022	1.124
X^2=min.	1.030	**1**	1.059
CV^2=min.	1.432	1.135	**1**

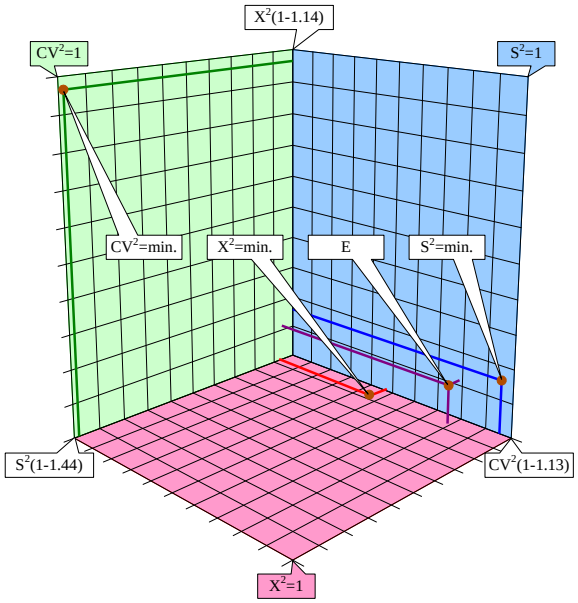


Figura 3. Poziția estimării empirice (21) în spațiul erorilor minime relative (18)-(19)-(20)

În Figura 3 s-a reprezentat în coordonatele definite de valorile pentru S^2, CV^2 şi X^2 valorile relative ale erorii (excesul de eroare) pentru rezultatele obținute prin estimarea simplă (E, relația 21),

minimizarea erorii pătratice absolute (S^2=min., relația 18), minimizarea erorii pătratice relative (CV^2=min., relația 19) şi minimizarea statisticii X^2 (X^2=min., relația 20).

Rezultatul reprezentării din Figura 3 este consistent cu rezultatul proiecțiilor în plan din Figura 2. Figura 3 evidențiază că soluția propusă de (21) este foarte aproape de soluția propusă de (18) şi (20) fiind intermediară acestora şi este foarte departe de soluția propusă de (19).

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a independenței

Nici aplicarea testului χ^2 pentru verificarea independenței nu este scutită de dificultăți în practică [42]. Astfel, FISHER a propus ca alternativă la testul χ^2 [43] testul care astăzi îi poartă numele (Fisher Exact Test, [44]), care se bazează pe calculul probabilităților marginale.

Pentru o tabelă de contingență 2X2, se cunoaşte că există exact un singur grad de libertate. Tabelul de mai jos (Tabelul 8) ilustrează această situație, în care impunerile sunt date de sumele observațiilor.

Tabelul 8. O tabelă de contingență 2X2 are un sigur grad de libertate (x)

X^2	Clasa A	Clasa Ω_1\A	Total Ω_1
Clasa B	x	n_1-x	n_1
Clasa Ω_2\B	n_2-x	n_3-n_1+x	n_2+n_3-n_1
Total Ω_2	n_2	n_3	n_2+n_3

Probabilitatea de a observa configurația din Tabelul 8 este dată de distribuția multinomială (relația 24), în timp ce valoarea statisticii Chi Square (X^2) este dată de relația (25):

$$p_{MN}(x;n_1,n_2,n_3)=\frac{n_1!\cdot n_2!\cdot n_3!\cdot(n_2+n_3-n_1)!}{x!\cdot(n_1-x)!\cdot(n_2-x)!\cdot(n_3-n_1+x)!\cdot(n_2+n_3)!} \quad (24)$$

$$X^2(x;n_1,n_2,n_3)=\frac{(xn_2+xn_3-n_1n_2)^2(n_2+n_3)}{n_1n_2n_3(n_2+n_3-n_1)} \quad (25)$$

Intervalul pe care observabila x poate lua valori este [0..min(n_1,n_2)].

Pentru exemplificarea problematicii s-au folosit datele din [43] (n_1 = 13, n_2 = 12, n_3 = 18) când intervalul de variație al lui x este [0..12] în timp ce valoarea observată a fost 10. Valoarea statisticii X^2 (relația 25) a fost reprezentată în Figura 4.

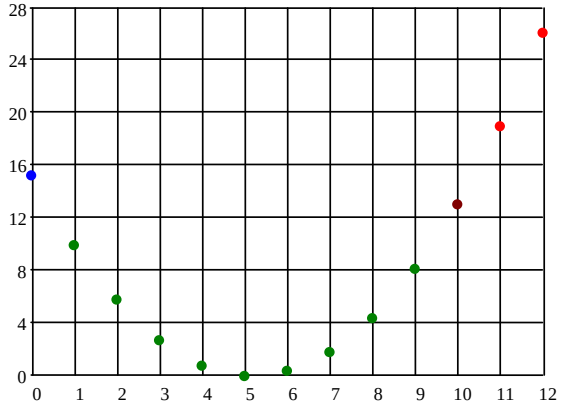


Figura 4. Valoarea statisticii X^2 în funcție de observabila independentă x

Aşa cum se evidențiază în Figura 4, spațiul observațiilor posibile cu privire la valoarea statisticii X^2 în funcție de observabila independentă x este discret. Valoarea observată (x=10) este situată într-o vecinătate a unei margini (x=12) având două observații mai defavorabile decât ea (cu o valoare X^2 mai mare) în aceeaşi vecinătate (x=11 şi x=10) şi o observație mai defavorabilă în vecinătatea opusă (x=0).

Din acest moment există două abordări, corespunzător cu obiectivul comparației din tabela de contingență. Dacă obiectivul observației este evidențierea probabilității ca să se observe depărtări mai mari de la omogenitate decât depărtarea observată, atunci probabilitatea asociată observației se obține din cumularea probabilităților în x=0, x=10, x=11 şi x=12. Dacă obiectivul observației este evidențierea probabilității ca să se observe depărtări mai mari de la omogenitate în sensul depărtării observate, atunci probabilitatea asociată observației se obține din cumularea probabilităților în x=10, x=11 şi x=12.

În Figura 5 a fost reprezentată probabilitatea observației (relația 24).

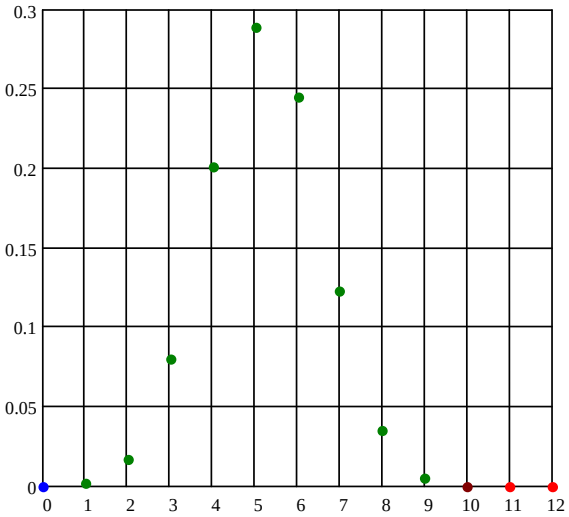


Figura 5. Valoarea statisticii probabilității observației în funcție de observabilă

Tabelul 9 prezintă pentru comparație valorile a trei probabilități: din distribuția χ^2 (p_{X2}), a probabilității de observare a unei depărtări de la omogenitate mai mari în sensul celei observate (p_{O2}) şi respectiv a unei depărtări mai mari în orice sens (p_{D2}). În această construcție probabilitatea din distribuția χ^2 (p_{X2}) este un estimator al unei depărtări mai mari în orice sens (p_{D2}).

Tabelul 9. Probabilități de observare

Probabilitate	Expresie de calcul	Valoare
p_{X2}	$\chi^2_{CDF}(X^2=13.03,df=1)$	$3.063\cdot10^{-4}$
$p_{O2}\,(x^2\geq X^2)$	$p_{MN}(10,13,12,18)+p_{MN}(11,13,12,18)+p_{MN}(12,13,12,18)$	$4.625\cdot10^{-4}$
$p_{O2}\,(x^2>X^2)$	$p_{MN}(11,13,12,18)+p_{MN}(12,13,12,18)$	$1.548\cdot10^{-5}$
$p_{D2}\,(x^2\geq X^2)$	$p_{O2}(x^2\geq X^2)+p_{MN}(0,13,12,18)$	$5.367\cdot10^{-4}$
$p_{D2}\,(x^2>X^2)$	$p_{O2}(x^2>X^2)+p_{MN}(0,13,12,18)$	$8.702\cdot10^{-5}$

Tabelul 9 arată cum testul χ^2 este în eroare atunci când valorile din tabelul de contingență se abat de la condițiile impuse asupra frecvențelor observate (cel mult 20% dintre celulele contingenței să conțină valori mai mici decât 5). Tabelul 9 mai arată cum în aceste cazuri testul Chi Square este expus la erori de tipul I (acordând o probabilitate mai mică decât cea reală evenimentului de a se produce observația observată, se află în riscul de a accepta ipoteza contrară chiar dacă ea nu este adevărată, ceea ce este totuna cu a respinge ipoteza nulă chiar dacă ea este adevărată).

Pentru a corecta semnificația statistică pentru tabele de contingență (sau frecvență) cu puține observații, Frank YATES a propus [45] o corecție la continuitate în care în expresia ecuației statisticii (relațiile (10), (11) şi (12)) din modulul diferenței între frecvența observată şi frecvența estimată în ipoteza independenței estimare se scade 0.5 simbolizând mijlocul intervalului de frecvență în timp ce MANTEL şi HAENSZEL au propus [46] ponderarea (împărțirea) statisticii X^2 cu df/(df-1), unde df este numărul de grade de libertate ale asocierii. Nici una dintre aceste ajustări însă nu este o alternativă decât la χ^2, testul Fisher Exact reprezentând testul de aur (Golden Test) pentru valoarea adevărată a probabilității de apariție a evenimentului observat.

Analiza asocierilor liniare. Regresii liniare multiple

Cel mai cunoscut model matematic de estimare a parametrilor ecuațiilor de regresie este cel fundamentat de Kolmogorov prin *minimizarea riscului*, un model cunoscut sub denumirea de *metoda celor mai mici pătrate*:

$$K(X,Y,B)=\sum\left(\hat{y}-y\right)^2=\sum\left(b_0+b_1x-y\right)^2$$

unde X,Y,B sunt vectorii coloană ai variabilei independente, variabilei dependente respectiv a coeficienților.

Au fost dezvoltate şi alte metode de estimare a parametrilor, bazate pe alte funcții de pierdere (sume de reziduuri) după cum urmează:

1. R. Fisher, 1912, *metoda verosimilității maxime*:

$$F(X,Y,B)=\sum\left(1-\exp(-(\hat{y}-y)^2/2)\right)=\sum\left(1-\exp(-(b_0+b_1x-y)^2/2)\right)$$

2. J. Newman, A. Wald, *metoda minimax*:

$$NW(X,Y,B)=\sum\left|\hat{y}-y\right|$$

3. Bayes, 1750, *metoda probabilității aposteriori maxime*:

$$NW(X,Y,B)=\sum\begin{cases}0, & \hat{y}-y<D(\hat{Y}-Y)/2\\ 1, & \hat{y}-y\geq D(\hat{Y}-Y)/2\end{cases}$$

În cazul multidimensional se fac convențiile: $x^T=(x^0,x^1,...,x^p)$, $x^0=1$; $X=(x_1,x_2,...,x_N)$; $Y=(y_1,y_2,...,y_N)$; $\hat{Y}=(\hat{y}_1,\hat{y}_2,...,\hat{y}_N)$; $B^T=(b^0,b^1,...,b^p)$ iar valoarea estimată este:

$$\hat{y}=\sum_{i=0}^{p}b^i\cdot x^i$$

Minimizând pătratele erorilor K(X,Y,B)=min avem:

$$K(X,Y,B)=\sum\left(\hat{y}-y\right)^2=\sum_{j=1}^{N}\left(\sum_{i=0}^{p}b^ix_j^i-y_j\right)^2=\min$$

În cazul de mai sus, soluția dată de algebra liniară sistemului de ecuații:

$$\frac{\partial}{\partial b^k}\sum_{j=1}^{N}\left(\sum_{i=0}^{p}b^ix_j^i-y_j\right)^2=0,\quad k=\overline{0,p}$$

este, după aranjarea sumelor:

$$\sum_{i=0}^{p}b^i\left(\sum_{j=1}^{N}x_j^kx_j^i\right)=\sum_{j=1}^{N}x_j^ky_j\quad k=\overline{0,p}$$

dată de ecuația:

$$B=CZ^{-1}$$

unde:

$$Z=(z_k^i)_{\substack{0\le i\le p\\0\le k\le p}}=(\sum_{j=1}^{N}x_j^kx_j^i)_{\substack{0\le i\le p\\0\le k\le p}}\text{ şi }C^T=(c^k)_{0\le k\le p}=(\sum_{j=1}^{N}x_j^ky_j)_{0\le k\le p}$$

Mai concret, dacă în urma unei determinări prin analiza spectrală [47] dacă avem p probe, fiecare având câte r constituenți şi determinăm semnalele pe q canale (de exemplu lungimi de undă diferite), semnalele depinzând liniar de concentrații, vor duce la ecuația: $R=CS^T+E$, unde:

R - matricea semnalelor (răspunsurilor) pe canalele considerate în număr de q pentru fiecare din cele p probe (dimensiune p×q);

C - matricea concentrațiilor celor r componenți în probe (dimensiune p×r);

S - matricea sensibilităților (dimensiune q×r);

E - matricea erorilor - cu aceleaşi dimensiuni cu R (p×q).

Deoarece în ultimul timp achiziția datelor se face în laboratoarele de analize aproape exclusiv cu ajutorul calculatoarelor, pentru analiza chimică cantitativă metodele bazate pe

algebra liniară multidimensională şi statistica multiliniară au devenit aplicații curente.

Odată stabiliți, coeficienții şi erorile ce afectează rezultatele semnalelor pe baza ecuațiilor de regresie, în analiza chimică se parcurge drumul invers, ecuațiile de regresie devenind ecuații de calibrare (corespondentul multidimensional al curbei de calibrare în două dimensiuni).

Tot ecuații de regresie se obțin şi prin implementarea modelelor de decizie multiliniare din domeniul inteligenței artificiale. Ecuațiile şi modelele de regresie au căpătat o utilizare tot mai frecventă odată cu dezvoltarea instrumentației analitice computerizate. În acest domeniu sunt nelipsite curbele de calibrare.

O noutate în analiza de regresie multiliniară (multifactorială) este ***analiza componentelor principale***. Deşi aceasta se apropie mai mult de analiza factorială, se înrudeşte foarte mult cu regresia multiliniară. Ca principiu al metodei, este o regresie liniară repetată de un număr de ori egal cu numărul de componente principale considerat. La fiecare iterație se determină coeficienții componentei considerate având ca date de intrare X_K: caracteristica principală K, Y_K: reziduul provenit din iterația pentru componenta principală (K-1) şi ca date de ieşire Y_{K+1} , reziduul provenit de la regresia Y_K după X_K şi vectorul de coeficienți B_K al componentei principale K.

Este de preferat analiza componentelor principale în locul regresiei multiliniare atât din considerente teoretice [48] cât şi practice.

Dintre considerentele teoretice, cel mai important este că vectorii B_K, K = 1, 2, ... sunt ortogonali în spațiul multidimensional al componentelor principale.

Dintre considerentele de natură practică [49], (1) nu este obligatoriu precizat la început numărul componentelor principale, numărul acestora putând să se modifice fără ca componentele principale deja calculate să fie afectate de acest lucru; (2) este mult mai uşor de interpretat fiecare componentă în parte, prin proiecția sa în planul corespunzător; (3) nu sunt afectate corelațiile de serie între şirurile de date prin aplicarea regresiei liniare repetate în locul regresiei liniare multiple.

În optimizare, atunci când numărul seturilor de date depăşeşte numărul coeficienților, modelul de optimizare ne conduce la un sistem de ecuații de regresie. În acest caz se minimizează suma erorilor generate de fiecare ecuație în parte pentru a obține un sistem determinat de ecuații, de unde, pe acelaşi principiu algebric enunțat la regresia multiliniară, se deduc coeficienții. În continuare, ecuația de regresie obținută este folosită pentru a da interpretări cantitative ale fenomenului studiat prin intermediul parametrului optimizat.

O ecuație de regresie liniară multiplă este o ecuație de forma:

$$b_0 + b_1X_1 + ... + b_nX_n = \hat{Y} \sim Y \qquad (1)$$

sau

$$b_1X_1 + ... + b_nX_n = \hat{Y} \sim Y \qquad (2)$$

unde Y este un şir de observații experimentale supuse erorii experimentale întâmplătoare iar $\{X_1, ..., X_n\}$ reprezintă o mulțime de descriptori $\{X_i\}_{1\le i\le n}$ asupra cărora se formulează ipoteza că o asociere liniară a acestora explică observațiile experimentale efectuate, iar şirul $(b_i)_{i\le n}$ reprezintă parametrii modelului (şi în acelaşi timp coeficienții ecuației).

Următoarele caracteristici definesc ecuațiile (1) şi (2):

÷ numărul de variabile independente: n = |X|;
÷ numărul de observații experimentale: m = |Y| = |X_1| = ... = |X_n|;
÷ numărul de parametrii ai modelului: |b| = n+1 pentru (1) şi |b| = n pentru (2).

În obținerea parametrilor ecuației de regresie (1) sau (2) se asumă următoarele ipoteze:

÷ valorile variabilei Y sunt normal distribuite; eroarea de măsură a lui Y este întâmplătoare şi de asemenea distribuită normal;
÷ variabilele X_1, ..., X_n au valori distribuite normal şi nu sunt afectate de erori.

Obținerea parametrilor unei ecuații de regresie $(b_i)_{i\le n}$ din observații este întotdeauna însoțită de un risc de a fi în eroare, iar în ipoteza că există relația liniară definită de (1) sau (2) folosind distribuția Student t se poate aprecia semnificația statistică şi intervalul de încredere al acestora.

Pentru ca ecuația (1) sau (2) să admită soluție unică este necesar (nu însă şi suficient) ca n ≤ m-

1. Pentru ca parametrii ecuației de regresie $(b_i)_{0\le i\le n}$ să aibă şi semnificație statistică este necesar (nu însă şi suficient) ca $n \le m-6$.

În cazul absenței semnificației statistice pentru coeficientul b_0, ecuația (1) se poate restrânge la ecuația (2).

Absența semnificației statistice pentru un coeficient b_i al unei variabile X_i $(1\le i\le n)$ în ecuația de regresie (1) asociată cu absența semnificației statistice a acestuia şi în ecuația de regresie (2) impune respingerea ipotezei legăturii liniare între observabila Y şi variabila X_i.

În aceste ipoteze problema determinării coeficienților (b_i) ale ecuației se rezolvă prin minimizarea sumei erorilor observat vs. cunoscut:

$$\Sigma_{1\le i\le m}(\hat{Y}_i - Y_i)^2 \to \min. \qquad (3)$$

Rezolvarea ecuației de minimizare presupune rezolvarea unui sistem de ecuații liniar şi omogen ale cărei necunoscute sunt coeficienții (b_i).

Rezolvarea ecuației de regresie (1) prin minimizarea pătratelor erorilor (LSE - least squares error) dată de relația (3) implică:

÷ exprimarea matriceală a sistemului de ecuații liniare şi omogene date de (3):

$$a=\begin{array}{c} \begin{matrix} 0 & 1 & \dots & n \end{matrix} \\ \begin{pmatrix} 1 & M(X_1) & \dots & M(X_n) \\ M(X_1) & M(X_1X_1) & \dots & M(X_1X_n) \\ \dots & \dots & \dots & \dots \\ M(X_n) & M(X_nX_1) & \dots & M(X_nX_n) \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix};$$

$$b=\begin{array}{c} \begin{matrix} 0 \end{matrix} \\ \begin{pmatrix} M(Y) \\ M(X_1Y) \\ \dots \\ M(X_nY) \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix}; \quad c=\begin{array}{c} \begin{matrix} 0 & 1 & \dots & n \end{matrix} \\ \begin{pmatrix} 1/m & 0 & \dots & 0 \\ 0 & 1/m & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/m \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix} \qquad (4)$$

÷ construcția matricei extinse a sistemului:

$$\begin{array}{c} \begin{matrix} -1 & 0 & 1 & \dots & n & n+1 & n+2 & \dots & 2n+1 \end{matrix} \\ \begin{pmatrix} M(Y) & 1 & M(X_1) & \dots & M(X_n) & 1/m & 0 & \dots & 0 \\ M(X_1Y) & M(X_1) & M(X_1X_1) & \dots & M(X_1X_n) & 0 & 1/m & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ M(X_nY) & M(X_n) & M(X_nX_1) & \dots & M(X_nX_n) & 0 & 0 & \dots & 1/m \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix}$$

÷ transformarea matricei extinse a sistemului folosind algoritmul Gauss-Jordan (prin operații elementare efectuate asupra liniilor matricei) având ca obiectiv (şi până când) se obține matricea unitară în spațiul matricei a şi când se obțin coeficienții $(b_i)_{0\le i\le n}$ şi erorile standard ale acestora $(s(b_i))_{0\le i\le n}$:

$$\begin{array}{c} \begin{matrix} -1 & 0 & 1 & \dots & n & n+1 & n+2 & \dots & 2n+1 \end{matrix} \\ \begin{pmatrix} b_0 & 1 & 0 & \dots & 0 & s(b_0) & 0 & \dots & 0 \\ b_1 & 0 & 1 & \dots & 0 & 0 & s(b_1) & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_n & 0 & 0 & \dots & 1 & 0 & 0 & \dots & s(b_n) \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix} \qquad (5)$$

Rezolvarea ecuației de regresie (2) prin minimizarea pătratelor erorilor (LSE - least squares error) dată de relația (3) implică:

÷ exprimarea matriceală a sistemului de ecuații liniare şi omogene date de (3):

$$b=\begin{matrix}0\\1\\...\\n\end{matrix}\begin{pmatrix}M(X_1Y)\\...\\M(X_nY)\end{pmatrix};\ a=\begin{matrix}1 & ... & n\\1\\...\\n\end{matrix}\begin{pmatrix}M(X_1X_1) & ... & M(X_1X_n)\\... & ... & ...\\M(X_nX_1) & ... & M(X_nX_n)\end{pmatrix};\ c=\begin{matrix}1 & ... & n\\1\\...\\n\end{matrix}\begin{pmatrix}1/m & ... & 0\\... & ... & ...\\0 & ... & 1/m\end{pmatrix} \quad (6)$$

÷ construcția matricei extinse a sistemului:

$$\begin{matrix}0 & 1 & ... & n & n+1 & ... & 2n\end{matrix}\begin{pmatrix}M(X_1Y) & M(X_1X_1) & ... & M(X_1X_n) & 1/m & ... & 0\\... & ... & ... & ... & ... & ... & ...\\M(X_nY) & M(X_nX_1) & ... & M(X_nX_n) & 0 & ... & 1/m\end{pmatrix}\begin{matrix}1\\...\\n\end{matrix}$$

÷ transformarea matricei extinse a sistemului folosind algoritmul Gauss-Jordan (prin operații elementare efectuate asupra liniilor matricei) având ca obiectiv (și până când) se obține matricea unitară în spațiul matricei a și când se obțin coeficienții $(b_i)_{0\le i\le n}$ și erorile standard ale acestora $(s(b_i))_{0\le i\le n}$:

$$\begin{matrix}0 & 1 & ... & n & n+1 & ... & 2n\end{matrix}\begin{pmatrix}b_1 & 1 & ... & 0 & s(b_1) & ... & 0\\... & ... & ... & ... & ... & ... & ...\\b_n & 0 & ... & 1 & 0 & ... & s(b_n)\end{pmatrix}\begin{matrix}1\\...\\n\end{matrix} \quad (7)$$

Coeficientul de corelație oferă o măsură a legăturii liniare între cele două variabile (Y și $\hat{Y}$) și se calculează pe baza formulei (unde M este valoarea medie):

$$r(Y,\hat{Y})=\frac{cov(Y,\hat{Y})}{s(Y)\cdot s(\hat{Y})}=\frac{M(Y\hat{Y})-M(Y)\cdot M(\hat{Y})}{\sqrt{M(Y^2)-M^2(Y)}\sqrt{M(\hat{Y}^2)-M^2(\hat{Y})}} \quad (8)$$

Semnificația statistică a legăturii liniare caracterizate de corelația dată de relația (8) este obținută din statistica Fisher F (unde |b| este numărul de coeficienți folosiți în estimare), iar probabilitatea asociată respingerii modelului liniar din funcția cumulativă de probabilitate (CDF) a distribuției Fisher:

$$F(r)=\frac{r^2}{1-r^2}\cdot\frac{m-|b|}{n};\ p_F=F_{CDF}\left(F(r),n,m-|b|\right) \quad (9)$$

În ipoteza că sistemul de ecuații admite o soluție unică pentru ecuația de regresie, ipotezele asumate permit și obținerea semnificațiilor statistice ale parametrilor $t(b_i)$ și a probabilităților asociate valorilor semnificativ statistic nenule ale acestora folosind distribuția Student t (unde $s(b_i)$ este dat de (5) pentru (1) și de (7) pentru (2)):

$$t(b_i)=\frac{b_i}{s(b_i)}\sqrt{\frac{m-|b|}{\sum_{i=1}^{m}(Y_i-\hat{Y}_i)^2}};\ p_t=t_{CDF}\left(t(b_i),m-|b|\right) \quad (10)$$

Dezvoltarea softurilor a dus la o explozie pe piața de programe specializate de prelucrări statistice. Majoritatea acestor programe au implementate rutine pentru calculul regresiilor de diferite feluri:

÷ GraFit, Data Analysis and Graphics Program, Erithacus Software Ltd.
÷ Slide, Slide Write Plus for Windows, Advanced Graphics Software Inc.
÷ MathCad, MathSoft Inc., Collabra Software Inc.
÷ Excell, Microsoft Corporation, Soft Art Dictionary and Program.
÷ Statistica, Statistica for Windows, StatSoft Inc.
÷ Surfer for Windows, Software Package, Golden Software.

Inginerie Mecanică
Managementul energiei termice

Managementul resurselor si gestiunea datelor

Curs:

- Probleme de managementul resurselor şi euristici
- Analiza calitativă şi cantitativă şi procedeul analitic
- Nivele de măsură şi scale de măsură
- Algoritmi genetici şi decizia asistată
- Baze de date şi sisteme de gestiune a bazelor de date
- Analiza consistenței în date. Elemente de statistică descriptivă
- Analiza asocierilor în date. Elemente de statistică inferențială
- Analiza asocierilor liniare. Regresii liniare multiple

Aplicații:

- Elemente de bază în utilizarea Excel. Tabele şi foi de calcul
- Elemente de bază în utilizarea SQL. Utilizarea MySQL & PHPMyAdmin
- Operații elementare asupra datelor. Utilizarea funcțiilor Excel
- Analiza factorilor & metoda Fisher. Utilizarea calculului tabelar Excel
- Analiza de regresie liniară multiplă. Utilizarea modulului de regresii Excel
- Modelarea proceselor. Iterația desfăşurării proceselor cu Excel
- Achiziția şi gestiunea datelor. Aplicațiile http://l.academicdirect.org/Engineering/

Baze de date:

- Date: http://www.epa.gov
- Standarde: http://www.nist.gov
- Patente: http://www.uspto.gov
- Aplicații: http://l.academicdirect.org
- Studii: http://lori.academicdirect.org

Cerință examen: Prezentarea (prezentare PowerPoint) unui studiu care să folosească metode discutate pe parcursul cursului. Metoda sau metodele folosite sunt la alegere. Obligatorie este structura prezentării:

- Introducere
- Scop
- Material şi/sau instrumente
- Metodă sau metode
- Rezultate
- Discuții
- Concluzii
- Referințe

Referințe curs:

---

[1] Gabelnick Aaron M., Capitano Adam T., Kane Sean M., Gland John L., and Fischer Daniel A., *Propylene Oxidation Mechanisms and Intermediates Using in Situ Soft X-ray Fluorescence Methods on the Pt(III) Surface*, Journal of the American Chemical Society, p. 143-149, Volume 122, Issue 1, January 12, 2000.

[2] Solak H. H. et al., *Measurement of strain in Al–Cu interconnect lines with x-ray microdiffraction*, Journal of Applied Physics, 86, 884, 15 July 1999.

[3] Steger-Hartmann T., Länge R., Schweinfurth H., *Environmental Risk Assessment for the Widely Used Iodinated X-Ray Contrast Agent Iopromide (Ultravist)*, American Society, EESA, p. 274-281, Volume 42, Issue 3.

[4] Chapman Wendy Webber, Fizman Marcelo, Chapman Brian E., Haug Peter J., *A Comparison of Classification Algorithms to Automatically Identify Chest X-Ray Reports That Support Pneumonia*, American Society, JBIN, p.

4-14, Volume 34, Issue 1.
[5] Venezia A. M., Liotta L. F., Deganello G., Schay Z., Guczi L., *Characterization of Pumice-Supported Ag-Pd and Cu-Pd Bimetallic Catalysts by X-Ray Photoelectron Spectroscopy and X-Ray Diffraction*; American Society, JCAT, p. 449-455, Volume 182, Issue 2.
[6] Ohno Youichi, *The Scanning-Tunneling Microscopy, the X-Ray Photoelectron Spectroscopy, the Inner-Shell-Electron Energy-Loss Spectroscopy Studies of M $Te_2$and M_3SiTe_6(M=Nb and Ta)*, American Society, JSSC, p. 63-73, Volume 142, Issue 1.
[7] BOOLE George, 1854. An Investigation of the Laws of Thought. (Reprinted 2003 as Laws of Thought. New York: Prometheus Books. ISBN 1-59102-089-1), p. 430.
[8] FISHER Ronald A, 1922. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1):87-94. DOI:10.2307/2340521
[9] Ralph V.L. HARTLEY, 1928. Transmission of Information. Bell Syst Tech J 1928:535-563.
[10] Maddison DR, Maddison WP. 2000. MacClade v4.0. http://macclade.org/
[11]Maddison WP, Maddison DR. 2006. Mesquite v1.1. http://mesquiteproject.org/
[12] Teorema Limită Centrală

÷ Cronologia contribuţiilor majore:
- Abraham DE MOIVRE. 1733. Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi. In: The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play (Abraham DE MOIVRE). W. Pearforn 1738: 235-243.
- Joseph L. LAGRANGE. 1776. Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relat ifs à cette matière. Miscellanea Taurinensia 5:167-232.
- Pierre S. LAPLACE. 1812. Théorie Analytique des Probabilités. Courcier, 465 p.
- Aleksandr M. LIAPUNOV. 1901. Nouvelle forme du théoreme sur la limite des probabilités. Mémoires de l'Académie Impériale des Sciences de St. Pétersbourg 12(5):1-24.

÷ Enunţul teoremei (fie $(X_n)_{n\geq 1}$ variabile independente şi $\exists\delta>0$ a.î. $\mu_{2+\delta}(X_n)<\infty$):

- dacă $\lim_{n\to\infty}\frac{\sum_{k=1}^{n}\mu_{2+\delta}(X_k)}{\left(\sum_{k=1}^{n}\sigma_k^{\,2}\right)^{(2+\delta)/2}}=0$ atunci $\frac{\sum_{i=1}^{n}(X_n-\mu_1(X_n))}{\sqrt{\sum_{k=1}^{n}\sigma_k^{\,2}}}\xrightarrow[n\to\infty]{}N(0,1)$

[13] BENFORD Frank. 1938. The law of anomalous numbers. Proceedings of the American Philosophical Society 78(4):551-572.
[14] HILL Theodore P. 1995. Base invariance implies Benford's Law. Proceedings of the American Mathematical Society 123(3):887-895.
[15] Carlos M JARQUE, Anil K BERA. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ Lett* 6(3):255-259.
[16] Carlos M JARQUE, Anil K BERA. 1981. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Econ Lett* 7(4):313-318.
[17] KOLMOGOROV Andrey. 1941. Confidence Limits for an Unknown Distribution Function. The Annals of Mathematical Statistics 12(4):461-463.
[18] SMIRNOV Nikolay V. 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19(2):279-281.
[19] Theodore W ANDERSON, Donald A DARLING. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23(2):193-212.
[20] Fritz W SCHOLZ, Michael A STEPHENS. 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association 82(399):918-924.
[21] Department of Defense Handbook. 2002. Composite Materials Handbook. Volume 1. Polymer Matrix Composites Guidelines for Characterization of Structural Materials. Chapter 8. Statistical Methods. 8.3.2.2 The k-sample Anderson-Darling test MIL-HDBK-17-1F:8-17.
[22] Lorentz JÄNTSCHI. 2009. http://l.academicdirect.org/Statistics/tests/kAD/, k-sample Anderson-Darling.
[23] Fritz W SCHOLZ, Michael A STEPHENS. 1986. K-Sample Anderson-Darling Tests of Fit, for Continuous and Discrete Cases. Technical Report. University of Washington. GN-22:81.
[24] A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, A. Castro-Perez. 2007. AnDartest:Anderson-Darling test for assessing normality of a sample data. http://mathworks.com/matlabcentral/fileexchange/14807

[25] PEARSON Karl. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 5th Ser 50:157-175.
[26] FISHER Ronald A. 1935. The Mathematical Distributions Used in the Common Tests of Significance. Econometrica 3:353-365.
[27] LIEBETRAU Albert M. 1983. Measures of association. Newbury Park, CA: Sage Publications. Quantitative Applications in the Social Sciences 32:1-96 (p.13).
[28] FISHER Ronald A. 1922. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85:87-94.
[29] FISHER Ronald A. 1924. The Conditions Under Which χ^2 Measures the Discrepancy Between Observation and Hypothesis. Journal of the Royal Statistical Society 87:442-450.
[30] SNEDECOR George W. and COCHRAN William G. 1989. Statistical Methods, Eighth Edition, Iowa State University Press.
[31] HARTLEY Ralph V L. 1928. Transmission of Information. Bell System Technical Journal 1928:535-563.
[32] Software. 2008. EasyFit v.5. MathWave Technologies. http://www.mathwave.com
[33] SCOTT David. 1992. Multivariate Density Estimation. John Wiley, Chapter 3.
[34] Software. 2005. Dataplot. National Institute for Standards and Technology. http://www.itl.nist.gov/div898/software/dataplot.html
[35] FISHER Ronald A. 1923. Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties. Journal of Agricultural Science 13:311-320.
[36] SNYDER Lloyd R. 1974. Classification of the solvent properties of common liquids. Journal of Chromatography A 92(2):223-230.
[37] STEELE Mike, CHASELING Janet, HURST Cameron. 2005. Simulated Power of the Discrete Cramer-von Mises Goodness-of-Fit Tests. International Congress on Modelling and Simulation. Advances and Applications for management and decision making. MODSIM 2005:1300-1304.
[38] KOLMOGOROV Andrey. 1941. Confidence Limits for an Unknown Distribution Function. The Annals of Mathematical Statistics 12(4):461-463.
[39] SMIRNOV Nikolay V. 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19(2):279-281.
[40] ANDERSON Theodore W, DARLING Donald A. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23(2):193-212.
[41] SCHOLZ Fritz W, STEPHENS Michael A. 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association 82(399):918-924.
[42] FISHER Ronald A. 1934. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
[43] FISHER Ronald A. 1935. The Logic of Inductive Inference. Journal of the Royal Statistical Society 98:39-54.
[44] AGRESTI Alan. 1992. A Survey of Exact Inference for Contingency Tables. Statistical Science 7(1):131-177.
[45] YATES Frank. 1934. Contingency table involving small numbers and the χ^2 test. Journal of the Royal Statistical Society (Supplement) 1: 217-235.
[46] MANTEL Nathan, HAENSZEL William. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4):719-748.
[47] D. Lorber; K. Faber and R. Kowalski, Anal. Chem., 1983, **55**, 643
[48] V. Centner, şi colab., Anal. Chem.; 1996, 68, 4851-4858.
D. Jouan-Rimbaud, B. Walczak, R.J. Poppi, O.E. de Noard and D.L.Massart; *Application of Wavelet Transform to Extract the Relevant From Spectral Data for Multivariate Calibration*, Anal. Chem., 1997, 69, 4317-4323. O.Stainback, S.Newmann, B.Cage, J.Saltiel, S.C.Miller, N.S.Dalal; Anal. Chem.; 1997; 69; 3708-3713.
[49] Massart D.L., Vandeginste B.G.M., Deming, S.N., Michotte Y., Kaufman L., *Chemometrics: a Textbook*, Elsevier, Amsterdam, 1988. Brereton R.G., *Chemometrics: Applications of Mathematics and Statistics to the Laboratory*; Ellis Horwood; Chichester; 1990. Jalliffe I.T., *Principal Component Analysis*, Springer-Verlag; New York, 1986. Meloun M., Mlitky J., Forina M., *Chemometrics for Analytical Chemistry, vol I: PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester, 1992.