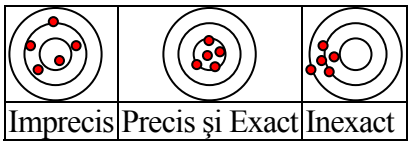
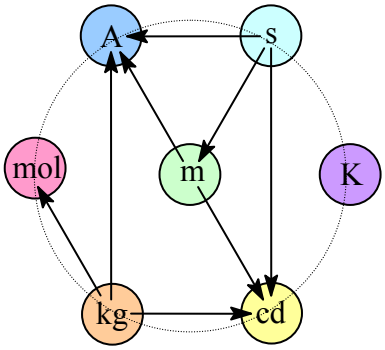
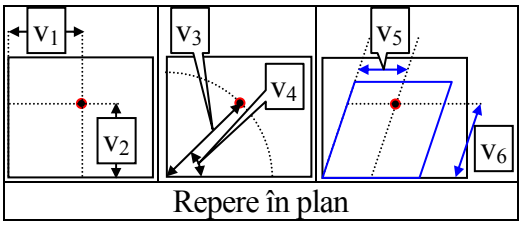
Lorentz JÄNTSCHI

Prezentarea şi prelucrarea datelor experimentale







U.T.Press
2013
ISBN 978-973-662-912-9

Prezentarea şi prelucrarea datelor experimentale

Mărimi şi măsurarea lor 2
Mărimi
Procesul de măsurare a unei mărimi
Valoarea a unei mărimi. Unitate de măsură
Sisteme de mărimi şi sisteme de unităţi de măsură 9
Cantitatea de substanţă
Analiza dimensională
Sisteme de unităţi de măsură; sistemul internaţional
Evaluarea numerică a expresiilor matematice 15
Ecuaţii algebrice şi transcedentale
Rezolvarea ecuaţiilor polinomiale
Rezolvarea numerică a ecuaţiilor transcedentale
Rezolvarea sistemelor de ecuaţii liniare
Metode iterative
Sisteme de ecuaţii neliniare
Integrarea numerică
Regula 1/3 a lui Simpson de integrare numerică
Elemente de teoria probabilităţilor 24
Măsuri statistice pentru populaţii şi eşantioane
Distribuţii statistice din măsurători repetate
Legi de distribuţie şi statistici ale acestora
Agrementul între observaţie şi model: statistici
Minimizarea erorii de agrement
Utilizarea momentelor
Utilizarea momentelor centrale
Folosind statisticile referitoare la populaţie
Estimare pe baza şansei maxime
Statistica Benford
Statistica Jarque-Bera
Statistica Kolmogorov-Smirnov
Distribuţia Kolmogorov
Testul Kolmogorov-Smirnov
Statistica Anderson-Darling
Statistica Pearson-Fisher Chi Square
Testul χ^2 independenţă, omogenitate şi asociere
Probleme frecvente în aplicarea testului χ^2
Măsuri ale tendinţei centrale
Binary and multinomial variables analysis: binomial confidence intervals 38
(binomial design and model)
(binomial confidence intervals)
(contingency of binomials)
(multinomial distribution)
Ordinal variables analysis: ranks statistic 43
(order statistics)
(Fisher-Yates correlation coefficient)
(generalized correlation coefficient)
(Jackknife method)
(Bootstrap method)
(tied values and fractional ranking)
(Spearman ρ)
(Kendall τ)
(Goodman and Kruskal's γ)
(Hoeffding D)
(other statistics)
(ordered contingencies)
(polychoric correlation)
Linear regression analysis: linear models 51
Aplicaţii 55
Referinţe 66

Mărimi şi măsurarea lor

Mărimi

O ***mărime*** este rezultatul unei măsurători efectuate asupra unei observabile cu scopul de a colecta valoarea unei proprietăți.

Se poate imagina ***spațiul de observare*** ca având o structură de arbore (a se vedea *Structura spațiului de observare*) care exprimă relațiile de apartenență dintre observabile în care la bază se află Universul (ca întreg spațiul de observare) iar la suprafață (aproape de noi în calitate de observatori) se află compuşii chimici - ca formă de reprezentare a materiei cu compoziție (de atomi) şi relații (între aceştia) bine definite.

Structură	Proprietate
- Univers	Întreg spațiul de observare
- Energie Radiantă	Viteza comparabilă cu cea a luminii
+ Radiații ca β, γ	Diferențiate prin intermediul proprietăților
- Materie	Întreg spațiul de observare nerelativistic
- Corp	Viteza mult mai mică decât a luminii
- Ansamblu de materiale	Compoziție (chimică) variabilă şi discontinuă
- Materiale	Compoziție (chimică) variabilă şi continuă
- Amestec de substanțe	Compoziție (chimică) bine definită
+ Substanța eterogenă	Compoziție (chimică) variabilă
- Soluție	Stare de agregare bine definită
+ Aliaj	Amestec de metale în stare lichidă sau solidă
- Substanța omogenă	Compoziție (chimică) constantă
+ Compus chimic	Structură chimică bine definită şi unică

Fig. 1. Structura spațiului de observare

Procesul de măsurare a unei mărimi

Procesul de observare este o activitate de colectare a cunoştințelor cu ajutorul simțurilor sau instrumentelor. Se presupune existența unui observator şi a unei observabile. Procesul de observare transferă o formă abstractă a cunoaşterii de la observabilă la observator (ca de exemplu, sub formă de numere sau imagini).

Măsurarea cuprinde două operațiuni serializate: observarea şi înregistrarea rezultatelor observației. Măsurarea depinde de natura obiectului observat (material) sau fenomene (imateriale), de metoda de măsurare şi de modul de înregistrare a rezultatelor observării.

Măsurarea presupune identificarea anterioară a elementului sau a elementelor care fac obiectul investigației şi rezultatul măsurării este o proprietate a elementului observat. O serie de măsurători presupune existența unei colecții de elemente distincte - ***mulțime*** - în care ordinea poate să nu fie relevantă. Mulțimea vidă ($\varnothing$) este mulțimea cu nici un element în ea.

Proprietatea (ca urmare a unei serii de observații) înregistrată cu exact una din exact două valori numite nefavorabil (şi scris ca F sau 0) şi favorabil (şi scrise ca T sau 1), respectiv, dă o ***valoare de adevăr***. Mulțimea de valori de adevăr ({0,1} sau {T, F}) este o mulțime în care elementele sunt ordonate în mod convențional (0 <1, F <T). Negația logică (!) este operațiunea (informațională) ce schimbă valoarea de adevăr, în timp ce identitatea logică (≡) lasă valoarea de adevăr neschimbată şi se exprimă faptul că rezultatul unei operații de măsurare pe două elemente este acelaşi. Folosind proprietatea 'valoare de adevăr' pe elementele unei mulțimi conceptul de ***submulțime*** este raționalizat. Apartenența este o proprietate a unui element de a fi ($\in$) sau a nu fi ($\notin$) într-o mulțime.

Asocierea totală scrisă ca S1×S2 şi definită prin $S1 \times S2 = \{(e1, e2) \mid E1 \in S1, S2 \in e2\}$ este produsul cartezian al mulțimilor S1 şi S2, iar o submulțime a S1×S2 este numită relație binară. Dacă S1=S2 relațiile sunt numite endo-relații. Următorul tabel redă (endo)relații (binare) cu proprietăți speciale (a se vedea *Relații binare: proprietăți şi reprezentanți*).

Id	Nume	Definiție	Reprezentanți
RE	Reflexive	$\forall$a: (a,a) $\in$ RE	=, $\subseteq$, |, $\leq$
CR	Co-reflexive	$\forall$a,b: (a,b) $\in$ CR then a$\equiv$b	=
QR	Cvasi-reflexive	(a,b) $\in$ QR atunci (a,a), (b,b) $\in$ QR	lim
IR	Ireflexive	$\forall$a: (a,a) $\notin$ IR	$\neq$, $\perp$, <
SY	Simetrice	(a,b) $\in$ SY atunci (b,a) $\in$ SY	=, CD, CM
NS	Anti-simetrice	(a,b), (b,a) $\in$ NS atunci a$\equiv$b	$\leq$
AS	Asimetrice	(a,b) $\in$ AS atunci (b,a) $\notin$ AS	IH, <
TS	Tranzitive	(a,b), (b,c) $\in$ TS atunci (a,c) $\in$ TS	=, $\leq$, <, $\subseteq$, |, $\Rightarrow$, IH
TL	Totale	$\forall$a,b: (a,b) $\in$ TL sau (b,a) $\in$ TL	$\leq$
TC	Tri-hotome	exact una dintre (a,b) $\in$ TL, (b,a) $\in$ TL, a$\equiv$b	<
ED	Euclidiene	(a,b), (a,c) $\in$ ED atunci (b,c) $\in$ ED	=
SE	Seriale	$\exists$b : (a,b) $\in$ SE	$\leq$
UQ	Unicitate	(a,b), (a,c) $\in$ UQ atunci b$\equiv$c	f(·)
EQ	Echivalență	atunci RE, SY, TS	=, ~, $\equiv$, CM, CD, ||
PO	Ordine parțială	atunci RE, NS, TS	|
TO	Ordine totală	atunci PO, TL	Alfabet, $\leq$
WO	Bine ordonate	atunci TO, SE	
$\perp$	Co-prime	cel mai mare divizor comun este 1	
VT	Adevărul vacuos	`daca A atunci B` când A = Fals	
=	Egal	atunci RE, CR, SY, NS, TS, ED, EQ	
$\leq$	Mai mic sau egal	atunci RE, NS, TS, TL, SE, PO, TO	
<	Mai mic	atunci IR, NS, AS, TS, TC, SE	
$\subseteq$	Submulțime	RE, NS, TS, SE, PO	
$\neq$	Diferit	IR, SI	
DI	Distanță euclidiană	RE, SI, TS, ED, SE, EQ	
IH	Moștenire	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 $\neq$ b atunci f(a) $\neq$ f(b)	
srj	Funcție surjectivă	$\exists$x : b=f(a)	
bij	Funcție bijectivă	INJ, SRJ	

Tab. 1. Relații binare: proprietăți și reprezentanți

Similaritatea între conceptul de ***funcție matematică*** și ***funcția de măsurare*** este evidentă când analizăm proprietățile relațiilor care se stabilesc între mulțimea observabilelor și mulțimea valorilor asociate din spațiul informațional. Ca și în cazul funcțiilor matematice, atunci când sunt efectuate măsurători experimentale sunt asigurate două proprietăți între elemente observate și proprietățile lor înregistrate. Și anume, pentru toate elementele observate avem înregistrări ale proprietăților lor atunci când facem măsurători - fiind asigurată *serializarea* (SE, v. *Relații binare: proprietăți și reprezentanți*). O măsură ne oferă (într-un anumit moment de timp și spațiu), o piesă informațională (o înregistrare) și unicitatea (UQ, v. *Relații binare: proprietăți și reprezentanți*) fiind asigurată de asemenea. Nici o altă proprietate cunoscută (matematică) a relațiilor nu este, în general, valabil pentru funcții matematice și nici pentru funcția de măsurare, așa încât putem spune că ceea funcția de măsurare face prin intermediul informațiilor este expresia unei funcții matematice (vezi *Colectarea datelor experimentale este o funcție matematică*).

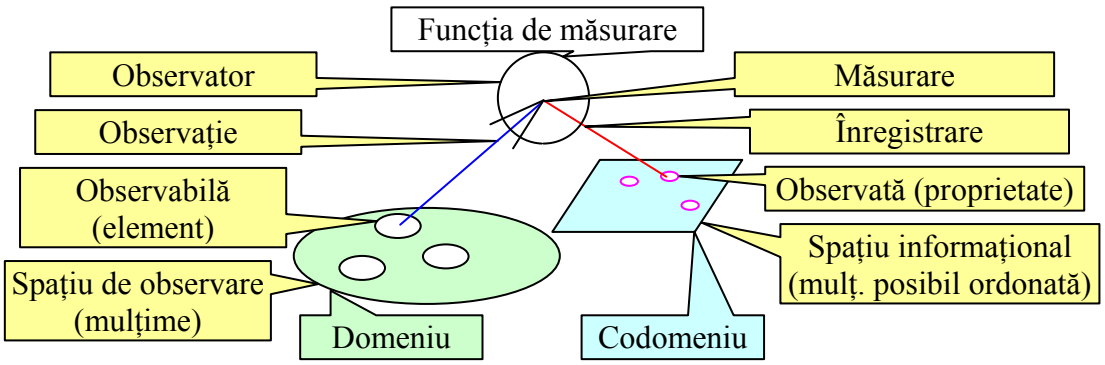


Fig. 2. Colectarea datelor experimentale este o funcție matematică

Pentru o mulțime finită S o ***funcție de numerotare*** poate fi definită iterativ după cum urmează: $S_0=S$; $S_1=S\setminus\{s_1\}$; ...; $S_i=S\setminus\{s_i\}$; ...; (etc.). Funcția $f(i)=s_i$ este o funcție de numerotare pe mulțimea S, şi arată că orice mulțime finită este numărabilă. Alegerea elementelor s_1, ..., s_i, ... din mulțimea S este instrumentul specific de măsurare. Este implicită observației, înregistrării, şi are ca efect construirea unei submulțimi reunind elementele rămase.

S-a arătat mai sus că conceptul de funcție matematică este legat de conceptul de măsurare. Mai departe, funcția de numerotare este instrumentul specific, cu care se face ordonarea în spațiul informațional. Mai mult, în cazul în care mulțimea S conține *n* elemente (desigur, ar trebui numărate mai întâi), atunci există exact n! posibilități de a enumera elementele sale prin intermediul funcției de numerotare. În afară de numerotarea implicită, funcția de măsurare aduce în spațiul informațional valoarea unei proprietăți observate.

Pentru două (presupus) finite mulțimi A (spațiul nostru de observare) şi B (spațiul nostru informațional) sunt exact $|B|^{|A|}$ posibilități de a defini (construi) funcții matematice $f:A\rightarrow B$ (să ne amintim, posibilități de măsurare) care asociază elementele din A cu elemente din B.

Pentru o observație cu 0 şi 1 ($|B|=2$) asupra unei mulțimi cu *n* elemente ($|A|=n$) avem un rezultat al numărării ($|A|=n$), un rezultat al posibilităților de enumerare ($|A|!=n!$) şi un rezultat al posibilităților de observare ($|B|^{|A|}=2^n$). Se poate verifica imediat că $n<n!$ Pentru $n>3$ şi mai mult, $n<2^n<n!$ pentru $n>4$. Chiar mai mult decât atât, pentru $n\rightarrow\infty$ $n<<2^n<<n!$, adică $\lim_{n\rightarrow\infty}(n/2^n) = \lim_{n\rightarrow\infty}(2^n/n!) = 0$.

Dacă o observație cu 0 şi 1 este 'cel mai simplu tip de observație' atunci o observație ce înregistrează în spațiul informațional numere reale este cel mai complex tip de observație.

Presupunând că rezultatul observației este un număr real, putem folosi o pereche formată dintr-un bit (0 sau 1) consemnând semnul şi un număr real pozitiv pentru a echivala conținutul din spațiul informațional (numărul real cu semn). Mai mult, se poate construi o funcție matematică bijectivă (care aduce o corespondență 1:1) între orice număr real pozitiv $[0,\infty)$ şi un număr real din intervalul $[0,1)$: $f:[0,1)\rightarrow[0,\infty)$, $f(x)=1+1/(1-x)$. Verificarea că este o funcție bijectivă pentru domeniul de definiție se poate face verificând că $f'(x)=1/(x-1)^2>0$. Rezultă deci că o codificare formată dintr-un semn (un bit) şi o succesiune de 0 şi 1 (reprezentarea în baza 2 a oricărui număr real subunitar) reflectă în totalitate orice număr real. Trecând însă din nou la limită dimensiunea spațiului observațional ($n\rightarrow\infty$) "puterea" reprezentării prin numere reale ($f:A\rightarrow\Re$) este de aceeaşi cardinalitate cu cea a reprezentării cu numere întregi ($f:A\rightarrow№$) sau binar ($f:A\rightarrow\{0,1\}$), $2^{\aleph_0}$, unde $\aleph_0$ este cardinalitatea mulțimii numerelor naturale. Acest simplu fapt ne arată că chiar dacă se măreşte calitatea reprezentării prin numere reale, rezoluția reprezentării este în continuare insuficientă pentru a egala calitatea enumerării ($\aleph_0!$).

Valoarea unei mărimi. Valoarea numerică. Unitate de măsură

O primă consecință imediată a calității reprezentării din spațiul informațional este existența degenerării. Degenerarea este reprezentarea prin intermediul aceleiaşi valori a rezultatului observației asupra a două elemente distincte (diferite).

Această degenerare este uneori un avantaj (când se pun în evidență similitudinile între proprietățile a două elemente) alteori un dezavantaj (când măsurarea care a avut ca scop evidențierea diferențelor între cele două elemente a eşuat în a-şi atinge scopul).

O a doua consecință imediată a calității reprezentării din spațiul informațional este că dacă degenerarea nu poate fi evitată prin funcția de măsurare, încă poate fi diminuată prin scala de măsurare. Ar trebui remarcat faptul că nu toate scalele de măsurare induc relații de ordine în spațiul informațional. Exemple naturale sunt grupa de sânge şi aminoacizi care constituie codul genetic, şi anume sunt situații când codificarea din spațiul informațional nu exprimă o relație de ordine (naturale) între valorile măsurate.

Fie o mulțime cu două elemente (C={a,b}) şi forțăm ipoteza că ordinea nu este relevantă între ele. Mulțimea submulțimilor lui C este S_C={{},{a},{b},{a,b}}. Un ordinea naturală în mulțimea S_C este definită prin cardinalitatea submulțimii. Cardinalitatea ca relație de ordine nu este strictă, pentru că există două submulțimi cu același număr de elemente: 0=|{}|<|{a}|=1=|{b}|<|{a,b}|=2. S-ar putea întreba: "Ce tip de scală de măsură definește cardinalitatea?" - Pentru a oferi un răspuns util trebuie să ne întoarcem la măsurare şi noi ar trebui să întrebăm mai întâi: "Ce caracteristici se doresc a fi evaluate?". În cazul în care răspunsul la a doua întrebare este numărul de elemente în subgrupul observat, atunci cardinalitatea este bine definită a fi ***cantitativă*** - fiind dotată cu o relație de ordine. În cazul în care diferențierea între submulțimile lui C este scopul dorit, atunci cardinalitatea submulțimii nu este suficientă. S-ar putea construi în continuare un alt experiment menit să diferențieze submulțimile pentru care apare degenerarea (în cazul de mai sus pentru submulțimile cu un element) şi o nouă funcție de măsurare ar da răspunsul la întrebarea "Submulțimea conține elementul 'a'?" (complementar cu răspunsul la întrebarea "Submulțimea conține elementul b?"). Aceasta este o măsurătoare tipic ***calitativă*** - căutăm potriviri.

O altă consecință derivată din căutarea după submulțimile unei mulțimi este că ***scala de măsură*** care se intenționează a se aplica ar trebui să fie de cel puțin verificată din punct de vedere al consistenței cu scopul propus.

Mai mult, chiar şi atunci când nu există relații de ordine, pot exista alte relații (cum ar fi complementul logic {a}={a,b}\{b} în mulțimea submulțimilor mulțimii {a,b}), care aduce în spațiul informațional faptul că nu întotdeauna rezultatele măsurătorilor sunt independente unul față de altul.

Mergând mai departe, tabelul de mai jos clasifică după complexitate (definită de către operațiile permise între valorile înregistrate) scalele de măsură (a se vedea *Scale de măsură*).

Scală	Tip	Operații	Structură	Statistici	Exemple
Binomială	Logic	"=", "!"	Algebră Booleană [1]	Moda, Fisher Exact [2]	Dead/Alive Fețele unei monezi
(multi) Nomi(n)ală	Discret	"="	Mulțime standard	Moda, Chi squared	Sistemul de grupe de sânge ABO Clasificarea organismelor vii
Ordinală	Discret	"=", "<"	Algebră comutativă	Mediana, Rangul	Numărul de atomi în molecule
Interval	Continuu	"≤", "-"	Spațiu afin (uni-dimensional)	Media, StDev, Corelația, Regresia, ANOVA	Scala de temperatură
Raport	Continuu	"≤", "-", "*"	Spațiu vectorial (uni-dimensional)	GeoMean, HarMean, CV, Logaritm	Dulceața relativă la sucroză pH Scala distanțelor Time scale Energy scale

Tab.2. Scale de măsură

O scală de măsurare este ***nominală*** dacă între valorile sale o relație de ordine nu poate fi definită. De obicei scala nominală de măsurare este destinată să fie utilizată pentru măsuri calitative. Scala ***binară*** (sau binomială) este cu doar două valori posibile (între care există o 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 nominală cu mai mult de două valori posibile este numit multinomială. Scara multinomială de măsurare are un număr finit de valori posibile şi independent de numărul lor, operează relaţia de complementaritate. Astfel, pentru {0,A,B,AB} grupe sanguine o valoare diferită de oricare dintre cele trei, sigur este cea de a patra. O serie finită de valori poate fi considerată o scală ***ordinală*** dacă între valorile lor posibile se poate defini o relaţie de ordine (naturală). Dacă presupunem că "Absent"<"Prezent", "Fals"<"Adevărat", "0"<"1" , "Negativ"<"Nenegativ", "Nepozitiv"<"Pozitiv", atunci toate aceste scale de măsură sunt ordinale. Mai mult, un exemplu de scală ordinală cu trei valori este: "Negativ"<"Zero"<"Pozitiv". Un alt lucru important cu privire la scalele ordinale este că nu sunt necesare, cu o cardinalitate finită. Dar este necesară existenţa unei relaţii de ordine definită prin "Succesorul unui element" (al unei valori) şi complementul acesteia "Predecesorul unui element" (al unei valori). În scala ***interval*** distanţa (sau diferenţa) între valorile posibile are un sens. De exemplu diferenţa între 30° şi 40° pe scala de temperatură are aceeaşi semnificaţie cu diferenţa între 70° şi 80°. Intervalul între două valori este interpretabil (are un sens fizic). Acesta este motivul pentru care are sens calcularea valorii medii a unei variabile de tip interval, ceea ce însă nu are sens pentru valorile unei scale ordinale. În acelaşi timp (vezi *Termometrul cu mercur şi scale de temperatură*) cum ar fi 80° nu este de două ori mai fierbinte decât 40° (aşa cum 2m sunt de 2 ori mai mulţi decât 1m), pentru scalele interval raportul dintre două valori nu are nici un sens.

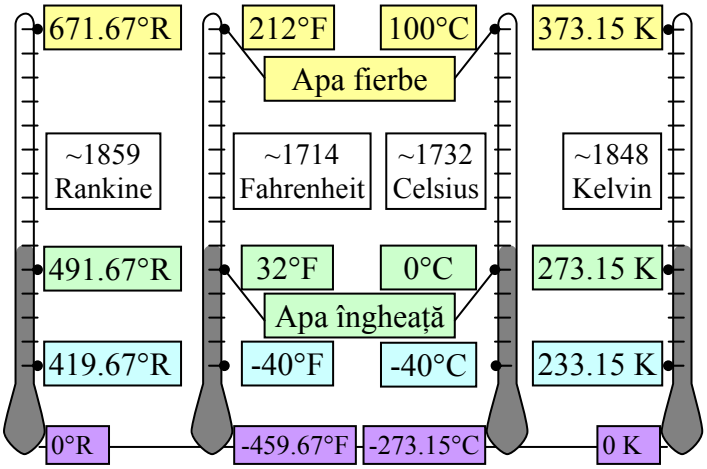
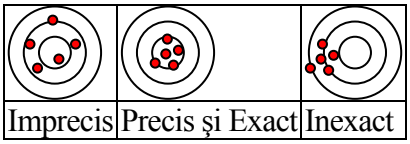


Fig. 3. Termometrul cu mercur şi scale de temperatură

În cele din urmă, pe scalele de tip ***raport***, valorile 0 şi/sau 1 au întotdeauna o semnificaţie. Ipoteza este că cea mai mică valoare observabilă este 0. Rezultă prin urmare faptul că dacă două valori sunt luate pe o scală raport, putem calcula raportul lor, şi, de asemenea, această măsură posedă o scală de măsurare de tip raport.

În cele din urmă, trebuie să remarcăm că mărimile măsurate pe o scală raport sunt ***aditive***, în timp ce mărimile măsurate pe celelalte scale sunt ***neaditive***.

Trebuie menţionat faptul că scala de măsurare nu dă precizia de măsurare, şi din acest punct de vedere este ilustrativ exemplul ţintei (v. *Precizie şi exactitate*).



Tab. 3. Precizie şi exactitate

De scala pe care a fost măsurată proprietatea depinde şi modul în care datele se pot prelucra şi interpreta. Aşa cum s-a ilustrat (v. Precizie şi exactitate) ***precizia*** şi ***exactitatea*** unei măsurători sunt la fel de importante ca şi valoarea măsurată înseşi. Acesta este motivul pentru care se obişnuieşte să se exprime valoarea unei măsurători împreună cu precizia sa de măsurare.

Desigur că există mai multe modalităţi de a exprima precizia unei măsurători. De exemplu, o măsurătoare cu un aparat de măsură nu poate depăşi precizia pentru care aparatul a fost construit.

Din punctul de vedere al scalei de măsură, o variabilă (din spațiul informațional) care numără moleculele dintr-un spațiu (fizic) dat este "la fel" de tip raport ca şi o variabilă ce măsoară temperatura mediului (fizic) în care aceste molecule sunt localizate, chiar dacă rezultatul acestor două operații de măsurare nu are aceeaşi precizie sau precizie comparabilă (sau nu pare a avea).

Este de dorit în mod evident ca scala de măsurare să încorporeze cât mai multe caracteristici ale variabilei măsurate, precum şi să posede numai caracteristicile găsite în variabilă măsurată, pentru că, în caz contrar, scala de măsurare devine o sursă de eroare.

Pe de cealaltă parte, există limite. Pentru a arăta aceasta, este suficient să facem apel la o serie de probleme nerezolvate în fizică[3]:

÷ Este 'spațiu-timp-ul' fundamental continuu sau discret?
÷ Există în natură mai mult de 4 dimensiuni 'spațiu-timp'?
÷ Sunt motive (fizice) să ne aşteptăm la existența altor universuri care să fie fundamental neobservabile?
÷ Prin ce diferă spațiul de timp?
÷ Există particule purtătoare de sarcină magnetică cum sunt electronii pentru sarcina electrică?
÷ Care este cel mai greu posibil nucleu atomic stabil sau instabil?

Simpla enumerare de mai sus ne arată că noi trebuie să operăm cu incertitudini.

În acest sens, este extrem de util să ne amintim de existența ***principiului incertitudinii*** [4]: Principiul lui Heisenberg al incertitudinii stabileşte prin inegalități (precise) că anumite perechi de proprietăți (fizice) cum sunt poziția şi momentul nu pot fi simultan cunoscute (măsurate) cu o precizie mare arbitrară. Cu cât mai precis o proprietate este măsurată, cu atât mai puțin precis cea de-a doua proprietate poate fi furnizată de un experiment de măsură. Principiul incertitudinii stabileşte că minimul pentru produsul incertitudinilor celor două proprietăți este egal cu jumătate din constanta lui Planck redusă ($\hbar = h/2\pi$).

În sensul celor de mai sus, este perfect justificat să se definească starea unei observabile prin intermediul unei ***funcții de undă*** având ca domeniu un spațiu-timp real iar ca codomeniu o coordonată complexă a cărei amplitudine să semnifice probabilitatea unei configurații a sistemului.

Într-adevăr, în 1926 Schrödinger [5] formulează ecuația ondulatorie a mecanicii cuantice a cărei soluție este o funcție de probabilitate (Ecuația lui Schrödinger): $i\hbar\partial\Psi / \partial t = \hat{H}\Psi$, unde Ψ este funcția de undă ce dă amplitudinea probabilității pentru diferite configurații ale sistemului la diferite momente de timp ($|\Psi(x,y,z,t)|^2$ este densitatea de probabilitate de a găsi particula la coordonata (x,y,z) şi momentul de timp t); $i\hbar\partial/\partial t$ este operatorul energiei; i este unitatea imaginară ($i = \sqrt{-1}$); $\hbar$ este constanta lui Planck redusă ($\hbar = h/2\pi$); $h = 6.62606 \cdot 10^{-34}$ J·s; $\hat{H}$: operatorul Hamilton ($\hat{H} = -\hbar^2\nabla^2/2m$); ∇^2: operatorul Laplace ($\nabla^2 = \partial^2/\partial x^2+\partial^2/\partial y^2+\partial^2/\partial z^2$).

Este dificil de înțeles acest lucru pentru o stare fizică, ceea ce a făcut ca autorul să explice plastic acest fapt într-o corespondență cu un coleg. Exemplul a devenit faimos şi a rămas sub numele de "pisica lui Schrödinger" (v. *Pisica lui Schrödinger*).

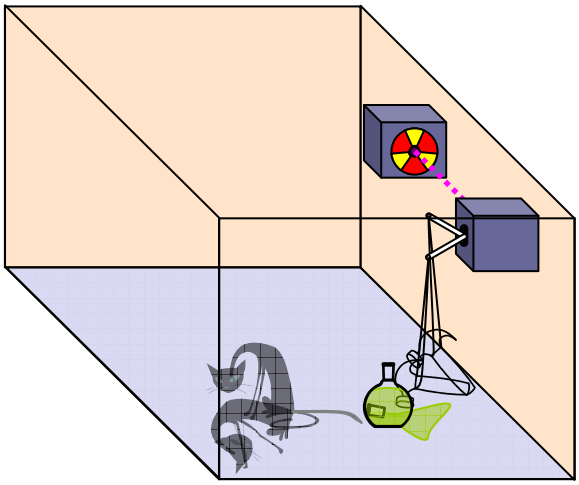


Fig. 4. Pisica lui Schrödinger

În esență, ceea ce experimentul denumit "pisica lui Schrödinger" exprimă este că odată supus un sistem unei protecții de decoerență cuantică este imposibil să se evalueze starea sistemului fără ca simultan cu evaluarea stării acestuia să se strice înseşi starea de decoereță în care se află.

Structura spațiului observabil aşa cum a fost ea prezentată (v. *Structura spațiului de observare*) desfăşoară structura materialelor până la nivelul de compus chimic, însă acesta nu este ultimul nivel de structurare intrinsecă. La rândul său, compusul chimic posedă o structură şi este alcătuit din atomi. Nici atomul nu este ultimul nivel de structură, fiind la rândul său alcătuit din nucleoni şi electroni.

Privind problema din alt unghi, următorul nivel de rafinament (celui ilustrat de *Structura spațiului de observare*) sunt compuşii chimici (v. *Nivele de rafinament ale conceptului de compus chimic*) definiți în sensul unei structuri chimice definite şi unice. Raționalizarea structurii chimice se face prin intermediul formulelor chimice. În acest sens, formele de reprezentare ale structurii chimice se pot desfăşura în continuare astfel:

Structură	Proprietate
Compus chimic	Structură moleculară definită şi unică
Formulă brută	Numărul de atomi din fiecare element în raport cu unul dintre elemente
Formulă moleculară	Numărul de atomi ai fiecărui element cuprinşi într-o moleculă
Formulă rațională	Exprimă grupele structurale din moleculă
Formulă geometrică	Exprimă geometria moleculei

Tab. 4. Nivele de rafinament ale conceptului de compus chimic

Nici măcar ultimul nivel de rafinament (formula geometrică) nu este întotdeauna suficient pentru a reda fidel structura moleculară. În acest sens, un exemplu simplu în care cunoscând distanțele între atomi şi unghiurile pe care legăturile între aceştia le formează nu este suficient pentru a accepta că referim o structură moleculară definită şi unică este butanul şi anume conformerii acestuia "Gauche g-" şi "Gauche g+" care au proprietatea de a răsuci diferit lumina polarizată (v. *Conformerii butanului*).

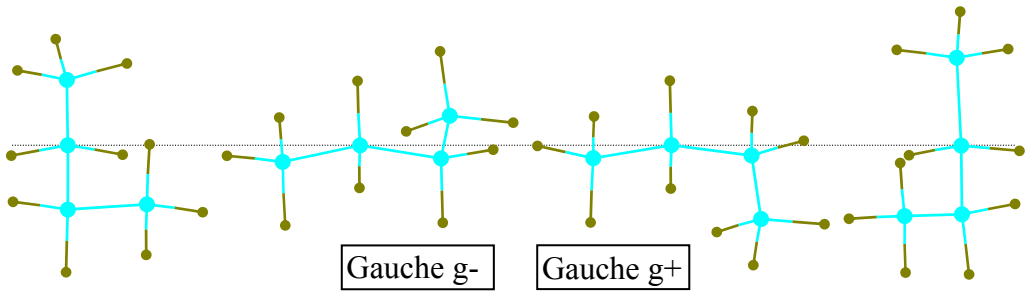


Fig. 5. Conformerii butanului

Sisteme de mărimi şi sisteme de unităţi de măsură

Cantitatea de substanţă

Indiferent de nivelul de structură (moleculară, atomică, subatomică) la care ne referim numărul de particule (compuşi chimici, atomi, electroni) la nivel macroscopic (observabil cu ochiul liber sau cu instrumente de mărire) cuprins într-un spaţiu de volum definit este imens. Din acest motiv pentru a face referire la macrocantităţi este nevoie de o unitate de măsură corespunzătoare. Aceasta este molul.

Molul este cantitatea de particule (molecule, ioni, atomi, electroni, altele asemenea sau grupuri ale acestora) al căror tip trebuie specificat şi al căror număr este egal cu numărul de atomi de carbon existenţi în 0.012 kg (12g) din izotopul acestuia ^{12}C.

Astfel, cantitatea de particule (impropriu spus "cantitate de substanţă") se poate exprima prin intermediul numărului de particule (N) sau prin intermediul numărului de moli (n) iar între aceste două modalităţi de exprimare există relaţia:

$$n = N/N_A$$

în care N_A este numărul lui Avogadro şi exprimă valoarea aproximativă a numărului de atomi de carbon existenţi în 0.012 kg (12g) din izotopul acestuia ^{12}C: $N_A = 6.02214 \cdot 10^{23}$ mol^{-1}.

Prin intermediul cantităţii de substanţă o serie de proprietăţi observate au caracter intensiv şi extensiv:

$$X_m = \frac{X}{n}$$

în care X nominalizează oricare ***proprietate extensivă*** (care depinde de cantitatea de substanţă) iar X_m nominalizează ***proprietatea intensivă*** corespondentă (care nu mai depinde de cantitatea de substanţă).

Energia ca atribut al unei substanţe şi consecinţă a structurii sale atomice, moleculare sau agregate este o mărime intensivă în timp ce energia specifică este corespondentul intensiv al energiei. Similar, energia liberă - eliberată sau absorbită într-un proces este o mărime extensivă în timp ce potenţialul chimic este mărimea intensivă asociată. Capacitatea calorică este cantitatea de căldură ce produce schimbarea temperaturii cu 1K şi este o mărime extensivă, şi capacitatea calorică specifică este mărimea intensivă asociată.

Masa este proprietate extensivă (M), iar masa molară (M_m) este proprietate intensivă. Volumul (V) este o proprietate extensivă în timp ce volumul molar (V_m) este o proprietate intensivă. Concentraţia (molară, molală, procentuală) este o mărime intensivă:

$$n = \frac{M}{M_m} = \frac{V}{V_m},\ c_M = \frac{n}{V_S},\ c_m = \frac{n}{m_S},\ c_{\%m} = \frac{m_d}{m_S},\ c_{\%v} = \frac{V_d}{V_S}$$

Se poate remarca că concentraţia molară variază cu temperatura, deoarece volumul variază cu temperatura, în timp ce molalitatea este o mărime independentă de temperatură. Se numeşte o soluţie diluată, o soluţie ce conţine cel mult 10^{-2} mol·l^{-1} de solut. În soluţiile diluate ionii de solut sunt separaţi de cel puţin 10 molecule de solvent. O altă mărime frecvent utilizată la amestecuri este fracţia molară x_j (a componentului j) din amestecul cu J ($j \in J$) componenţi:

$$x_j = \frac{n_j}{\Sigma_j n_j}$$

Se poate demonstra că fracţia molară este o mărime intensivă. Astfel, fie un amestec P cu compoziţia exprimată prin raportul numărului de molecule din fiecare component j în amestec $\alpha_1:\alpha_2:\ldots:\alpha_J$ (cum ar fi pentru $C_2O_4H_2$, $\alpha_1:\alpha_2:\alpha_3 = 2:4:2 = 1:2:1$), şi numărul de moli n.

Din cele $N = n \cdot N_A$ molecule ale amestecului, pentru a respecta proporţia, numărul de molecule din componentul j este $N_j = N \cdot \alpha_j / \Sigma_j \alpha_j$. Fracţia molară a amestecului este:

$$x_j = \frac{n_j}{\Sigma_j n_j} = \frac{N_j}{N_A} \Big/ \Sigma_j \frac{N_j}{N_A} = \frac{N_j}{\Sigma_j N_j} = \frac{N \cdot \alpha_j / \Sigma_j \alpha_j}{\Sigma_j N \cdot \alpha_j / \Sigma_j \alpha_j} = \frac{N \cdot \alpha_j}{\Sigma_j N \cdot \alpha_j} = \frac{\alpha_j}{\Sigma_j \alpha_j}$$

Expresia rezultată nu depinde decât de compoziţie şi nu depinde de numărul de moli sau molecule implicate aşa că este o mărime intensivă.

Densitatea este o mărime intensivă. În cazul unui amestec cu J componenţi:

$$\rho = \frac{\Sigma_j m_j}{\Sigma_j V_j} = \frac{\Sigma_j n_j M_j}{\Sigma_j V_j} = \frac{\Sigma_j n \cdot x_j M_j}{\Sigma_j V_j} = \frac{n \cdot \Sigma_j x_j M_j}{\Sigma_j V_j} = \frac{\Sigma_j x_j M_j}{\Sigma_j V_j / n} = \frac{\Sigma_j x_j M_j}{V_m}$$

În formula de mai sus intervin numai mărimi intensive (x_j, M_j şi V_m) şi astfel defineşte o mărime intensivă.

Presiunea, într-un sistem în echilibru, este o mărime intensivă atâta timp cât valoarea acesteia în sistem este egală cu valoarea acesteia în orice parte a acestuia.

Temperatura, într-un sistem în echilibru, este o mărime intensivă atâta timp cât valoarea acesteia în sistem este egală cu valoarea acesteia în orice parte a acestuia.

În final trebuie făcută remarca că conceptul de mărime intensivă referă macrocantităţi şi îşi pierde sensul la nivel microscopic. Luând doar temperatura ca exemplu, în spaţiu este de câteva grade Kelvin, în timp ce obiectele care se deplasează (cum ar fi o rachetă sau un meteorit) pot ajunge la temperaturi de câteva mii de grade Kelvin, aşa cum rezultă din teoria cinetico-moleculară.

Analiza dimensională

În geometrie, pentru a identifica în mod unic poziţia unui punct în plan avem nevoie de un reper. Faţă de acest reper, punctul de probă are două grade de libertate, ceea ce ne arată că poziţia sa este în mod unic determinată de fixarea valorilor pentru două proprietăţi geometrice ale acestuia. Alegerea reperului nu schimbă condiţionarea anterioară. În figura de mai jos (v. *Repere în plan*), reperul poate fi de exemplu unul din colţurile dreptunghiului. Cea mai evidentă modalitate este de a fixa valorile proiecţiilor din punctul de probă pe laturile dreptunghiului (fig. a) însă nu este necesar ca ambele valori să provină din acelaşi tip de măsurătoare (în fig. b o măsurătoare este de distanţă, alta este de unghi) şi nici nu este necesar să fie ortogonale (în fig. c oricare din laturile paralelogramului are o proiecţie nenulă pe laturile învecinate).

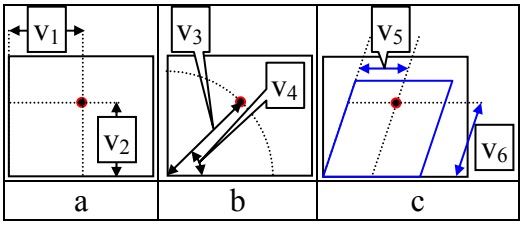


Fig.6. Repere în plan

Exemplul *Repere în plan* ne arată că modalităţile de a exprima prin valori proprietăţile observate pot fi extrem de diferite. În acelaşi timp ne arată că ar putea exista, în fiecare caz în parte, o soluţie de tipul celei ilustrate în fig. a, în care variabilele asociate celor două valori să fie ortogonale. Chiar însă şi în acest caz, nimic nu ne opreşte să considerăm că şi cazul ilustrat de fig. b este de asemenea ortogonal, deci putem avea chiar mai multe soluţii. În toate cazurile însă, am remarcat necesitatea impunerii a două valori fixe, ceea ce ne arată că dimensionalitatea sistemului este 2.

Problema poate fi extinsă la mărimi şi relaţii altele decât geometrice. În general în ştiinţe operăm cu mărimi şi unităţi de măsură, şi o ***analiză dimensională*** poate fi condusă dacă se alege un set redus de mărimi şi unităţi pe baza cărora să se poată exprima toate celelalte. Aşa cum s-a arătat mai sus, problema nu are o singură soluţie, însă unele soluţii sunt preferate.

Analiza dimensională asupra mărimilor şi relaţiilor poate servi la verificarea corectitudinii definirii acestora.

În analiza dimensională a mărimilor şi relaţiilor fizice, se preferă ca mărimi de bază lungimea, masa, timpul şi sarcina electrică. Chiar şi aici însă, o remarcă poate fi făcută, şi anume că folosind constanta vitezei luminii în vid, se poate exprima o relaţie de dependenţă între lungime şi timp, iar în ceea ce priveşte sarcina electrică, definiţia sa utilizează lungimea, masa, şi timpul.

Aşa cum în *Repere în plan* se poate remarca, existenţa unor valori constante în proprietăţile observate reduce dimensionalitatea sistemului. Dacă în oricare dintre cele trei cazuri se fixează o valoare din cele două, dimensionalitatea sistemului se reduce la 1, chiar dacă în fiecare caz în parte manifestarea reducerii dimensionalităţii produce efecte observabile diferite. Rezultă de aici că, indiferent de modalitatea de exprimare a valorii (deci a unităţii de măsură alese) de importanţă sunt

valorile constantelor universale (v. *Constante universale*).

Constantă	Explicație	Valoare
Viteza luminii în vid	Distanța parcursă de lumină, în vid, în timp de 1 secundă; în fapt metrul este definit pe baza acestei valori	c_0=299792458 $m \cdot s^{-1}$
Impedanța vidului	Este raportul între magnitudinea câmpului electric şi magnetic al radiațiilor electromagnetice ce traversează vidul	Z_0 = 376.730313461... Ω
Constanta magnetică	Este permitivitatea magnetică a vidului	$\mu_0 = 4\pi \cdot 10^{-7}$ $N \cdot A^{-2}$
Constanta electrică	Este permitivitatea electrică a vidului	$\varepsilon_0 = \frac{1}{\mu_0 c^2}$
Constanta gravitațională	Constanta care intervine în legea atracției universale	G = 6.673...·10^{-11} $m^3 \cdot kg^{-1} \cdot s^{-2}$
Constanta lui Planck	Este cuanta de acțiune în mecanica cuantică exprimând raportul între energia şi frecvența unei radiații electromagnetice	h = 6.626069...·10^{-34} J·s
Lungimea Plank	Este aproximativ 10^{-20} mai mică decât diametrul unui proton şi e considerată cea mai mică dimensiune posibilă; nu există însă semnificație fizică dovedită a acesteia	$l_P = \sqrt{\frac{\hbar G}{c^3}}$
Masa Planck	Este masa unei găuri negre ipotetice a cărei rază este egală cu lungimea Planck; raza găurii negre ipotetice este astfel încât viteza necesară pentru a evada de pe suprafață este egală cu viteza luminii	$m_P = \sqrt{\frac{\hbar c}{G}}$
Timpul Planck	Este timpul necesar luminii să traverseze în vid o distanță egală cu l_P	$t_P = \sqrt{\frac{\hbar G}{c^5}}$
Temperatura Planck	Este temperatura de la care teoria fizică actuală încetează să mai opereze, deoarece nu avem cunoştințe despre gravitația cuantică; punctul în care se formează o 'gaură neagră' de energie	$T_P = \frac{m_P c^2}{k_B}$

Tab. 5. Constante universale

Constantă	Explicație	Valoare
Sarcina electrică elementară	Sarcina electrică transportată de un singur electron sau proton	e = 1.6021765...·10^{-19} C
Energia Hartree	Energia electrică potențială a unui atom de hidrogen în starea sa fundamentală şi aproximativ dublul energiei sale de ionizare	E_h = 4.359743…·10^{-18} J
Numărul lui Avogadro	Numărul de particule dintr-un mol de substanță	N_A = 6.02214X·10^{23} mol^{-1}
Constanta atomică de masă	A 12-a parte din masa unui atom de carbon 12	$m_u = 10^{-3}/N_A$
Constanta gazelor	Constanta de proporționalitate din legea gazului ideal (pV=nRT)	R=8.31446... J·mol^{-1}·K^{-1}
Constanta Boltzmann	Stabileşte proporționalitatea între energie şi temperatură în teoria cinetico-moleculară (ε/J=k_B·T/2, ε fiind energia unei particule, iar J numărul de componente ale energiei	$k_B = \frac{R}{N_A}$

Tab. 6. Alte constante frecvent utilizate

Pe lângă constantele universale o serie de alte constante sunt cunoscute şi frecvent utilizate (v. *Alte constante frecvent utilizate*).

În tabelele 6 şi 7 se observă că sunt relativ puține constante universale independente, celelalte, mult mai multe fiind obținute din cele 'de bază'.

Un element important al analizei dimensionale îl reprezintă alegerea setului de mărimi de bază. De exemplu în cinematică un set de mărimi de bază este (masa, distanța, timpul) sau (M, L, T). Pe baza acestuia se poate exprima dimensionalitatea vitezei ($L \cdot T^{-1}$), accelerației ($L \cdot T^{-2}$), impulsului ($M \cdot L \cdot T^{-1}$) şi a forței ($M \cdot L \cdot T^{-2}$). Setul de mărimi (distanța, viteza, timpul) nu este un set de bază pentru cinematică pentru că masa nu se poate exprima pe baza acestuia şi de asemenea cele trei nu sunt independente ($V=L \cdot T^{-1}$).

Cu ajutorul analizei dimensionale se poate stabili corectitudinea unei ecuații din punctul de vedere al omogenității dimensionale şi în cazul în care sunt implicate o serie de mărimi diferite într-o expresie simplă se poate chiar stabili forma ecuației din analiza dimensională.

Sisteme de unități de măsură; sistemul internațional

Un sistem de unități de măsură este un set de unități care poate fi folosit pentru a specifica orice care poate fi măsurat.

Sistemele actuale de unități de măsură, MKS (metru, kilogram, secundă), CGS (centimetru, gram, secundă), FPS (picior, livră, secundă) diferă unul de celălalt doar prin alegerea unităților de măsură, mărimile măsurate fiind aceleaşi (distanță, masă, timp).

La acestea trei se adaugă şi temperatura (măsurată în Kelvin, grade Celsius sau Fahrenheit), însă cum definiția actuală a distanței este făcută prin intermediul constantei vitezei luminii în vid ($d = c \cdot t$, c - constanta vitezei luminii în vid, v. *Constante universale*) practic tot prin intermediul a trei unități fundamentale (temperatura, timpul, masa) sunt construite şi definițiile celorlalte de bază (v. *Relații între definițiile mărimilor sistemului fundamental de unități de măsură*).

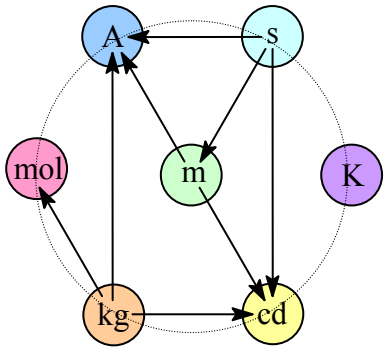


Fig. 7. Relații între definițiile mărimilor sistemului fundamental de unități de măsură

Revenind asupra mărimilor de bază, sistemul internațional de unități recunoaşte un număr de 7 ca fiind mărimi de bază [6].

Mărime			Unitate	
Nume	Dimensiune	Simbol	Nume	Simbol
masa	M	m	kilogram	kg
cantitatea de substanță	N	n	mol	mol
temperatura termodinamică	Θ	T	kelvin	K
timpul	T	t	secundă	s
lungimea	L	l, x, r	metru	m
curentul electric	I	I, i	amper	A
intensitatea luminoasă	J	I_V	candela	cd

Tab. 7. Mărimi şi unități de bază ale sistemului internațional

Următoarele sunt definițiile mărimilor de bază din sistemul internațional de unități:

÷ ***Kilogramul*** este masa prototipului internațional pentru kilogram (materialul folosit fiind un aliaj de Platină cu 10±0.0001% Iridiu);

÷ ***Molul*** este cantitatea de substanță a unui sistem ce conține tot atâtea entități elementare câți atomi sunt în 0.0012 kg de Carbon cu masa atomică 12 (când molul este utilizat, entitățile elementare

trebuie specificate şi pot fi atomi, molecule, ioni, electroni, alte particule sau grupuri specificate de astfel de particule);

÷ ***Kelvinul*** este 1/273.16 din temperatura termodinamică a punctului triplu al apei (definiția referă apa cu următoarea compoziție în izotopi: $^{2}H:^{1}H=0.00015576$, $^{17}O:^{16}O=0.0003799$, $^{18}O:^{16}O=0.0020052$);

÷ ***Secunda*** este durata corespunzătoare unui număr de 9192631770 perioade ale radiației corespunzătoare tranziției între cele două nivele hiperfine ale stării fundamentale a atomului de Cesiu cu masa atomică 133 (definiția referă atomul de cesiu la temperatura de repaus de 0K); ***Hetzul*** ([Hz], pentru ***frecvență***) este definit implicit în definiția secundei, frecvența radiației corespunzătoare tranziției între cele două nivele hiperfine ale stării fundamentale a atomului de Cesiu cu masa atomică 133 fiind de 9192631770 Hz;

÷ ***Metrul*** este lungimea drumului parcurs de lumină în vid într-un interval de timp de 1/299792458 secunde (rezultă că viteza luminii în vid este exact $c_0=299792458\ m\cdot s^{-1}$);

÷ ***Amperul*** este curentul electric constant care, dacă este menținut între doi conductori paraleli de lungime infinită, de grosime neglijabilă, plasați la 1m unul de celălalt în vid vor produce între aceşti conductori o forță egală cu $2\cdot10^{-7}$ N ($kg\cdot m\cdot s^{-2}$) pentru fiecare metru de lungime;

÷ ***Candela*** este intensitatea luminoasă, într-o direcție dată, a unei surse care emite radiație monocromatică cu frecvența de $540\cdot10^{12}$ Hz (s^{-1}) şi care are o intensitate radiantă în acea direcție de 1/683 W/sr (sr este simbolul pentru steradian, unitatea SI de măsură a unghiului solid; sfera are 4π sr; 1sr este aria unei calote sferice cu aria 1 pe o sferă cu raza 1); ***Lumenul*** ([lm], pentru ***flux luminos***) este definit implicit în definiția candelei, fiind fluxul luminos produs de sursa care emite radiație monocromatică cu frecvența de $540\cdot10^{12}$ Hz (s^{-1}) în unghiul solid de 1sr;

Pe baza unităților de bază sunt construite unitățile derivate:

÷ ***Radianul*** ([rad], pentru ***unghi***) este unghiul la centru dat de o deschidere de 1m pe un cerc cu o rază de 1m;

÷ ***Steradianul*** ([sr], pentru ***unghi solid***) este unghiul solid la centru dat de o calotă sferică cu aria de $1m^2$ pe o sferă cu o rază de 1m;

÷ ***Newtonul*** ([N], pentru ***forță***) este forța necesară să accelereze 1kg de masă cu $1ms^{-2}$.

÷ ***Pascalul*** ([Pa], pentru ***presiune***) este presiunea exercitată de o forță de 1N pe o suprafață de $1m^2$;

÷ ***Jouleul*** ([J], pentru ***energie***) este energia cheltuită în aplicarea unei forțe de 1N în deplasarea pe distanța a 1m, sau în traversarea curentului electric de 1A printr-o rezistență de 1Ω pentru 1s;

÷ ***Wattul*** ([W], pentru ***putere***) este puterea necesară pentru a produce o energie de 1J pe durata a 1s;

÷ ***Voltul*** ([V], pentru ***potențial electric***) este diferența în potențialul electric pe lungimea unui fir când un curent electric de 1A disipă 1W;

÷ ***Ohmul*** ([Ω], pentru ***rezistența electrică***) este rezistența între 2 puncte ale unui conductor când o diferență de potențial constantă de 1V aplicată între aceste puncte produce în conductor un curent de 1A; conductanța electrică este inversul rezistenței electrice; ***Siemensul*** ([S], pentru ***conductanța electrică***) este conductanța între 2 puncte ale unui conductor când o diferență de potențial constantă de 1V aplicată între aceste puncte produce în conductor un curent de 1A;

÷ ***Coulombul*** ([C], pentru ***sarcina electrică***) este sarcina electrică transportată de un curent constant de 1A pe durata a 1s;

÷ ***Faradul*** ([F], pentru ***capacitatea electrică***) este valoarea care produce o diferență de 1V când este încărcată cu 1C;

÷ ***Luxul*** ([lx], pentru ***iluminare***) este iluminarea corespunzătoare unui flux luminos de 1lm care traversează o suprafață de $1m^2$;

÷ ***Becquerelul*** ([Bq], pentru ***radioactivitate***) este activitatea unui material radioactiv în care are loc dezintegrarea a 1 nucleu atomic pe durata a 1s;

÷ ***Grayul*** ([Gy], pentru ***doza absorbită de radiație ionizantă***) este doza absorbită când se transferă o energie de 1J unei cantități de materie de 1kg;

÷ ***Sievertul*** ([Sv], pentru ***doza echivalentă de radiație ionizantă***) este doza absorbită când se observă apariția cancerului în 5.5% din cazuri;

÷ ***Katalul*** ([kat], pentru ***activitatea catalitică***) este valoarea catalitică ce converteşte 1mol de reactant pe durata a 1s;

Multiplii şi submultiplii unităților de măsură sunt standardizate (v. *Multiplii şi submultiplii ai unităților de măsură*).

Divizor	10^0	10^1	10^2	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
Simbol	-	d	c	m	μ	n	p	f	a	z	y
Prefix	-	deci	centi	mili	micro	nano	pico	femto	atto	zepto	yocto
Multiplicator	10^0	10^1	10^2	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
Simbol	-	da	h	k	M	G	T	P	E	Z	Y
Prefix	-	deca	hecto	kilo	mega	giga	tera	peta	exa	zetta	yotta

Tab. 8. Multiplii şi submultiplii ai unităților de măsură

Evaluarea numerică a expresiilor matematice

Ecuații algebrice și transcedentale

O ***funcție transcendentală*** este o funcție care nu satisface o ecuație polinomială ai cărei coeficienți sunt ei înșiși polinoame, în contrast cu o funcție algebrică, care satisface o asemenea ecuație. Cu alte cuvinte, o funcție transcendentală este o funcție care "transcende" algebra, în sensul că aceasta nu poate fi exprimată în termenii unei secvențe finite de operații algebrice de adunare, înmulțire și extragerea rădăcinii. Exemple de funcții transcedentale sunt funcția exponențială, logaritmul și funcțiile trigonometrice.

O ***ecuație transcedentală*** este o ecuație care conține o funcție transcedentală.

Un ***număr algebric*** este un număr o rădăcină a unui polinom (nenul) într-o variabilă cu coeficienți raționali. Numerele care nu sunt algebrice sunt ***numere transcedentale***. Exemplele includ pe π și e.

Inversa funcției $f(w)=w\cdot e^w$ se numește ***funcția W a lui Lambert***. Derivata funcției f, $f'(w) = e^w(1+w)$ ne arată că w=-1 este un punct de extrem pentru funcția f (de minim, $f(-1)=-e^{-1}$) așa încât față de acest punct de extrem de o parte și de alta a acestuia există exact două soluții (w_1 și w_2) pentru ecuația $y=w\cdot e^w$. Din acest motiv inversa funcției f, funcția W este fie multivalorică (redând pentru un y două valori, w_1 și w_2) fie se restrânge domeniul de definiție al funcției astfel încât să devină (atât f cât și W) funcție bijectivă. Funcția Lambert este o sursă de numere transcedentale, ceea ce înseamnă că soluțiile w asociate ecuației $y=w\cdot e^w$ unde y este un număr algebric exceptând pe y=0 sunt toate transcedentale.

Funcția W a lui Lambert este de importanță teoretică, întrucât multe ecuații implicând exponențiale se 'rezolvă' oferind soluțiile (indirecte) ca expresii implicând soluțiile funcției W.

Exemplul 1. Pentru a, b > 0 să se rezolve ecuația: $a^t=b\cdot t$. *Rezolvare*. $a^t=b\cdot t \rightarrow 1 = b\cdot t/a^t \rightarrow 1 = b\cdot t\cdot e^{-t\cdot\ln(a)} \rightarrow 1/b = t\cdot e^{-t\cdot\ln(a)} \rightarrow -\ln(a)/b = -t\cdot\ln(a)\cdot e^{-t\cdot\ln(a)} \rightarrow$ Dacă se obține w ca o soluție a ecuației $-\ln(a)/b = w\cdot e^w$ atunci $t = -w/\ln(a)$, care se exprimă formal astfel: $t = -W(-\ln(a)/b)/\ln(a)$.

Exemplul 2. Mai general, pentru $a \neq 0$ și $q>0$, să se rezolve ecuația: $q^{a\cdot x+b} = c\cdot x+d$. *Indicație*. Se substituie $-t = a\cdot x + a\cdot d/c$ și $R = t\cdot q^t = -(a/c)\cdot q^{b-a\cdot d/c}$.

Exemplul 3. Să se rezolve ecuația $x^x=a$. *Rezolvare*. $x^x=a \rightarrow x\cdot\ln(x) = \ln(a) \rightarrow e^{\ln(x)}\cdot\ln(x) = \ln(a) \rightarrow$ Dacă se obține w ca o soluție a ecuației $\ln(a) = w\cdot e^w$ atunci $x = e^w$, care se exprimă formal astfel: $x = e^{W(\ln(a))}$. *Remarcă*. Dacă $y=w\cdot e^w$, sau scris parametrizat $y=w(y)\cdot e^{w(y)}$ pentru $y=\ln(z)$ atunci și $\ln(z)=w(\ln(z))\cdot e^{w(\ln(z))}$ are loc, și $e^{w(\ln(z))}=\ln(z)/w(\ln(z)$ astfel încât soluția ecuației $x^x=a$ se mai poate scrie ca: $x=e^{W(\ln(a))}=\ln(a)/W(\ln(a))$.

Alte exemple de utilizare a funcției W a lui Lambert pentru a obține soluțiile ecuațiilor transcedentale includ ecuațiile în forma $x\cdot\log_b(x)=a$ (cu soluția $x=e^{W(a\cdot\ln(b))}$), valorile explicite ale curentului într-un circuit în care sunt în serie o diodă și o rezistență (relația între curent și tensiune în cazul diodei este dat de o lege exponențială [7] în timp ce la rezistor este o dependență liniară), încetinitorul atomic Zeeman [8], curgerile și depunerile granulelor și suspensiilor, curgerea fluidelor vâscoase [9], și imagistica creierului [10].

În toate aceste cazuri, ca și în altele de asemeni, când este posibilă obținerea unei soluții analitice sau, în cazul soluțiilor oferite prin intermediul funcției W a lui Lambert, obținerea unor soluții pseudo-analitice, aceste soluții analitice sunt de preferat în favoarea celor obținute prin ***evaluare numerică***.

Rezolvarea ecuațiilor polinomiale

Un caz aparte de ecuații sunt cele în forma $y = a_nx^n+...+a_1x+a_0$, numite ecuații polinomiale. Din fericire, astăzi există un algoritm care găsește toate rădăcinile oricărui polinom de orice grad [11], algoritm disponibil online [12] și care este implementat în majoritatea programelor profesionale care necesită rezolvarea de astfel de ecuații. Precizia în care acest algoritm rezolvă ecuațiile polinomiale este limitată numai de precizia de calcul a implementării. De exemplu o implementare C/C++ poate da o precizie de 10^{-14}, o implementare în FreePascal o precizie de 10^{-19} în timp ce o implementare în Fortran poate da o precizie de 10^{-23}.

Rezolvarea numerică a ecuațiilor transcedentale

Pentru ecuațiile la care literatura de specialitate nu ne pune la dispoziție metode deterministe care să ne conducă spre soluție, singura modalitate de abordare rămasă este de a căuta soluțiile. Avantajul pe care îl are utilizarea calculatorului este imens: stocarea expresiei ecuației de rezolvat în calculator oferă practic posibilitatea evaluării rezultatului funcției asociate în orice moment. Fie astfel ecuația de rezolvat y=f(x), în care y este o valoare care trebuie precizată de fiecare dată când facem apel la o rezolvare numerică. Se construiește funcția asociată g(x) = f(x) - y. Se poate evalua funcția g(x) în orice valoare a domeniului său de definiție prin simpla sa implementare într-o rutină de calcul.

Rezolvarea numerică totdeauna presupune ***cunoaşterea domeniului*** în care se află soluția (sau soluțiile) ecuației. Dacă nu avem nici o indicație asupra domeniului, o bună idee este să se încerce ***reprezentarea grafică*** a dependenței într-un program specializat (cum este MathCad). Odată identificat domeniul (fie acesta [a,b]) în care se află soluția sau soluțiile, următorul pas necesar este separarea soluțiilor, şi anume găsirea unui subdomeniu (fie acesta [c,d], a≤c, d≤b) în care există o singură soluție. Practic nu există metode standard care să ne ducă la acest subdomeniu, însă din nou o reprezentare grafică ne poate ajuta.

Condiția ca ***într-un subdomeniu să existe cel puțin o soluție*** este ca g(c)·g(d)<0. Condiția g(c)·g(d)<0 este ***suficientă***, în sensul în care dacă g(c)·g(d)<0 atunci există cel puțin o soluție g(z)=0 în intervalul [c,d] însă ***nu este şi necesară***, existența unui număr par de soluții în domeniul [c,d] făcând ca g(c)·g(d)>0 şi totuşi să existe soluții în acest interval.

Odată identificat un domeniu [c,d] pentru care g(c)·g(d)<0, căutarea soluției ecuației g(z)=0 se poate face prin ***căutări succesive***, şi totdeauna un astfel de algoritm va produce cel puțin o soluție.

Cea simplă metodă de căutare succesivă este prin înjumătățirea intervalului (v. *Căutarea soluției g(z)=0 când g(c)·g(d)<0*).

```
Let a0=c; b0=d;
Repeat
        If(a0-b0<eps)then stop;//solution is any of a0 and b0
        If(b0-a0<eps)then stop; //solution is any of a0 and b0
        c0=(a0+b0)/2;  if(g(c0)<eps)then stop;//solution is c0
        If(g(a0)*g(c0)<0)then b0=c0; else a0=c0;
Until(false);
```

Fig.8. Căutarea soluției g(z)=0 când g(c)·g(d)<0

Aşa cum se observă în figura de mai sus (v. *Căutarea soluției g(z)=0 când g(c)·g(d)<0*) căutarea succesivă a soluției g(z)=0 impune definirea unei toleranțe (eps) a cărei valoare minimă este stabilită în funcție de ***cerințele problemei de rezolvat*** (o precizie de 4 cifre semnificative este foarte rar depăşită de instrumentația de analiză fizică şi chimică) şi de ***capacitatea procesorului matematic*** al platformei de calcul (s-a arătat mai sus că limita de precizie poate merge până la 23 de cifre semnificative fără a face apel la calculul matematic în precizie arbitrară).

Rezolvarea sistemelor de ecuații liniare

Din fericire, şi pentru acest caz sunt elaborate metode, unele dintre ele foarte eficiente, de rezolvare exactă. Diferențele între o metodă şi o altă metodă în acest caz o reprezintă precizia cu care este oferită soluția. Pentru a putea rezolva un sistem de ecuații liniare, sunt necesare un număr mare de operații de înmulțire, împărțire, adunare şi scădere, iar cea mai eficientă rezolvare are ordinul de complexitate polinomial, fiind de ordinul puterii a 3-a a numărului de ecuații, $O(n^3)$. Astfel, pentru un număr foarte mare de ecuații, sunt implicate şi un număr mare de operații (pentru n = 1000, $n^3=10^9$), ceea ce face ca precizia maximă posibilă să scadă şi ea (folosind o implementare C/C++, din precizia de 14 cifre semnificative, scăzând cele 9 datorate operațiilor aritmetice, ajungem la pragul de 5 cifre semnificative ca precizie maximă a evaluării). Morala este că pentru cazurile în care se operează cu sisteme de ecuații foarte mari, utilizarea celor mai precise metode de rezolvare este obligatorie.

În figura următoare (v. *Procedura Gauss-Jordan pentru regresii liniare multiple*) este redată

procedura Gauss-Jordan pentru rezolvarea sistemelor de ecuații liniare, aplicată pentru modele de regresie liniară, când soluția este furnizată simultan cu calculul de varianță [13].

df = m+1	$y \sim \hat{y} = a_0 \cdot 1 + a_1 \cdot x_1 + \ldots + a_m \cdot x_m$ pentru $(y_i, x_{1,i}, \ldots, x_{m,i})_{1 \le i \le n}$								
i(linii)\j(coloane)	-1	0	1	…	m	m+1	m+2	…	2m+1
0	Σy_i	n	$\Sigma x_{1,i}$	…	$\Sigma x_{m,i}$	1	0	…	0
1	$\Sigma y_i x_{1,i}$	$\Sigma x_{1,i}$	$\Sigma x_{1,i}^2$	…	$\Sigma x_{1,i} x_{m,i}$	0	1	…	0
…	…	…	…	…	…	…	…	…	…
m	$\Sigma y_i x_{m,i}$	$\Sigma x_{m,i}$	$\Sigma x_{1,i} x_{m,i}$	…	$\Sigma x_{m,i} x_{m,i}$	0	0	…	1
Inițial	S_B	S_A				I_{m+1}			
Operații elementare pe linii	↓	↓				↓			
Final	$A = S_A^{-1} S_B$	I_{m+1}				S_A^{-1}			
0	a_0	1	0	…	0	$(S_A)^{-1}_{0,0}$	…	…	…
1	a_1	0	1	…	0		$(S_A)^{-1}_{1,1}$	…	…
…	…	…	…	…	…	…	…	…	…
m	a_m	0	0	…	1	…	…	…	$(S_A)^{-1}_{m,m}$
df = m	$y \sim \hat{y} = a_1 \cdot x_1 + \ldots + a_m \cdot x_m$ pentru $(y_i, x_{1,i}, \ldots, x_{m,i})_{1 \le i \le n}$								
i(linii)\j(coloane)	-1	0	1	…	m	m+1	m+2	…	2m+1
0									
1	$\Sigma y_i x_{1,i}$		$\Sigma x_{1,i} x_{1,i}$	…	$\Sigma x_{1,i} x_{m,i}$		1	…	0
…	…		…	…	…		…	…	…
m	$\Sigma y_i x_{m,i}$		$\Sigma x_{1,i} x_{m,i}$	…	$\Sigma x_{m,i} x_{m,i}$		0	…	1
Inițial	S_B		S_A				I_m		
Operații elementare pe linii	↓		↓				↓		
Final	$A = S_A^{-1} S_B$		I_m				S_A^{-1}		
0									
1	a_1		1	…	0		$(S_A)^{-1}_{1,1}$	…	…
…	…		…	…	…		…	…	…
m	a_m		0	…	1		…	…	$(S_A)^{-1}_{m,m}$
Coeficienți și varianțe	$\varepsilon_i = \hat{y}_i - y_i = a_0 + a_1 x_{1,i} + \ldots + a_m x_{m,i} - y_i$					$s^2(a_i) = (S_A)^{-1}_{i,i} \cdot \Sigma_i \varepsilon_i^2 / (n-df)$			
Semnificația coeficienților	$t_i = t(a_i) = a_i / s(a_i)$					$p_i = p(t_i) = CDF_t(t_i, n-df)$			
Semnificația modelului	$r = \frac{n \cdot \Sigma_i y_i \hat{y}_i - \Sigma_i y_i \cdot \Sigma_i \hat{y}_i}{\sqrt{n \cdot \Sigma_i y_i^2 - \Sigma_i y_i \cdot \Sigma_i y_i} \sqrt{n \cdot \Sigma_i \hat{y}_i^2 - \Sigma_i \hat{y}_i \cdot \Sigma_i \hat{y}_i}}$ $F(r) = \frac{r^2}{1-r^2} \cdot \frac{n-df}{m}, p_r = CDF_F(F(r), m, n-df)$								

Fig.9. Procedura Gauss-Jordan pentru regresii liniare multiple

Metode iterative

Pentru ecuații neliniare cu expresii complicate, şi cu atât mai mult pentru ecuații transcedentale, cel puțin o soluție poate rezulta prin aproximații succesive. Fie ecuația f(x)=0 pentru care căutăm o rădăcină x.

O metodă iterativă totdeauna porneşte de la o valoare inițială x_0 pe care căutăm să o îmbunătățim astfel încât la final să obținem o bună aproximație $|f(x_0)|<\varepsilon$ unde ε este o valoare arbitrară impusă de cerințele problemei de rezolvat.

Ecuația f(x)=0 se poate transforma astfel încât să se obțină o ecuație echivalentă cu aceasta în forma x=g(x). Succesul metodei iterative depinde de modalitatea de exprimare a ecuației echivalente.

Exemplu. Fie $f(x)=x^2-2x-3$. În tabelul următor (v. *Aplicarea metodelor iterative*) sunt ilustrate mai multe modalități de exprimare a ecuației echivalente.

Ecuație echivalentă	$x = x^2 - x - 3$	$x = (2x+3)/x$	$x = \pm\sqrt{2x+3}$	$x = \pm\sqrt{2x+3}$
Ecuație iterativă	$x_{i+1} = x_i^{\,2} - x_i - 3$	$x_{i+1} = (2x_i+3)/x_i$	$x_{i+1} = \sqrt{2x_i+3}$	$x_{i+1} = -\sqrt{2x_i+3}$
i=0	1	1	9	9
i=1	5	-3	3.872983	-3.87298
i=2	2.6	9	2.178524	#NUM!
i=3	3.153846	69	1.164924	#NUM!
i=4	2.95122	4689	#NUM!	#NUM!
i=5	3.016529	21982029	#NUM!	#NUM!

Tab.9. Aplicarea metodelor iterative

Aşa cum se observă din tabelul anterior (v. *Aplicarea metodelor iterative*) nu orice alegere a metodei iterative conduce la convergența către soluție (f(x)=0).

Sisteme de ecuații neliniare

Metoda aproximațiilor succesive poate fi o soluție comodă în anumite cazuri când se cere rezolvarea de sisteme de ecuații neliniare.

Un exemplu tipic de aplicare este pentru maximizarea agrementului între observație şi model. Fie exemplul ilustrat mai jos (v. *Observații la contingența de doi factori multiplicativi*).

Factor "B" / Factor "A"	Nivel b_1	Nivel b_2	Nivel b_3	Nivel b_4
Nivel a_1	Obs_{11}	Obs_{12}	Obs_{13}	Obs_{14}
Nivel a_2	Obs_{21}	Obs_{22}	Obs_{23}	Obs_{24}
Nivel a_3	Obs_{31}	Obs_{32}	Obs_{33}	Obs_{34}
Nivel a_4	Obs_{41}	Obs_{42}	Obs_{43}	Obs_{44}

Tab.10. Observații la contingența de doi factori multiplicativi

În cazul observațiilor la contingența de doi factori multiplicativi (v. *Observații la contingența de doi factori multiplicativi*) ipoteza este că $Obs_{ij} \sim a_i \cdot b_j$ şi valoarea aşteptată în urma observației este $Exp_{i,j}=a_i \cdot b_j$. Datorită unor factori aleatori necunoscuți, observația e afectată de erori, astfel încât aproape niciodată $Exp_{i,j}=Obs_{i,j}$. Dificultatea însă constă în faptul că nu se cunosc valorile nivelelor factorilor (a_1, ...,a_4; b_1, ...,b_4) şi nici tipul erorii (eroare proporțională cu valoarea observată, eroare în magnitudine uniformă, etc.). Orice tentativă de a rezolva analitic sistemul (prin identificarea valorilor nivelelor) e sortită eşecului, deoarece sistemul are cel puțin o nedeterminare (putând fi astfel rezolvat doar cel mult parametric).

Avem însă posibilitatea să obținem un set de valori inițiale pentru valorile aşteptate cu formula:

$$Exp_{i,j} = \frac{\sum_{k=1}^{r} Obs_{i,k} \sum_{k=1}^{c} Obs_{k,j}}{\sum_{i=1}^{r}\sum_{j=1}^{c} Obs_{i,j}}$$

În acelaşi cadru al presupunerii naturale al efectului multiplicativ al celor doi factori asupra observabilei Obs din punct de vedere matematic se pot formula trei presupuneri cu privire la eroarea pătratică $(Obs_{i,j}-Exp_{i,j})^2$ 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 (S^2), a doua dintre ipoteze conduce la minimizarea pătratului coeficientului de variație (CV^2) iar o soluție (una din mai multe soluții posibile) la cea de-a

treia dintre ipoteze (X^2) o reprezintă minimizarea statisticii X^2 (v. *Minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model*).

S^2	CV^2	X^2
$\sum_{i=1}^{r}\sum_{j=1}^{c}(Obs_{i,j}-a_ib_j)^2=\min.$	$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(Obs_{i,j}-a_ib_j)^2}{(a_ib_j)^2}=\min.$	$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(Obs_{i,j}-a_ib_j)^2}{a_ib_j}=\min.$

Tab.11. Minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model

În relaţiile de mai sus (v. *Minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model*) apar exprimaţi cei doi factori ("A" şi "B") a căror independenţă se verifică prin intermediul efectului multiplicativ (a_i, $1\le i\le r$ reprezintă contribuţia primului factor la valoarea aşteptată $E_{i,j}$ iar b_j, $1\le j\le c$ reprezintă contribuţia 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 contribuţii: $E_{i,j}=a_i\cdot b_j$).

Minimizarea cantităţilor date de relaţiile de mai sus în scopul determinării contribuţiilor 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 relaţia:

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

unde expresia de derivat $\cdot(a_i,b_j)$ este una din expresiile date de S^2, CV^2 şi X^2. În urma calculului (derivare) se poate obţine că condiţiile de minimizare a diferitelor tipuri de erori pentru agrementul între observaţie sunt echivalente cu ecuaţiile următoare (v. *Ecuaţii în minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model*).

S^2	CV^2	X^2	Domeniu
$a_i=\sum_{j=1}^{c}b_jO_{i,j}\Big/\sum_{j=1}^{c}b_j^2$	$a_i=\sum_{j=1}^{c}\frac{O_{i,j}^2}{b_j^2}\Big/\sum_{j=1}^{c}\frac{O_{i,j}}{b_j}$	$a_i^2=\sum_{j=1}^{c}\frac{O_{i,j}^2}{b_j}\Big/\sum_{j=1}^{c}b_j$	$i=1..r$
$b_j=\sum_{i=1}^{r}a_iO_{i,j}\Big/\sum_{i=1}^{r}a_i^2$	$b_j=\sum_{i=1}^{r}\frac{O_{i,j}^2}{a_i^2}\Big/\sum_{i=1}^{r}\frac{O_{i,j}}{a_i}$	$b_j^2=\sum_{i=1}^{r}\frac{O_{i,j}^2}{a_i}\Big/\sum_{i=1}^{r}a_i$	$j=1..c$

Tab.12. Ecuaţii în minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model

Se poate de asemenea arăta matematic că relaţiile de mai sus (v. *Ecuaţii în minimizarea diferitelor tipuri de erori pentru agrementul între observaţie şi model*) admit o infinitate de soluţii şi că familiile de soluţii ale relaţiilor se află în vecinătatea familiei de soluţii date valorile iniţiale ale valorilor aşteptate ($Exp_{i,j}=a_i\cdot b_j$):

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

Calea directă de rezolvare a ecuaţiilor de minimizare fără a face apel la ecuaţia de aproximare este ineficientă. De exemplu pentru r=2, c=3 substituţiile în relaţia pentru S^2 duc la:

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

care este rezolvabilă în (a_2/a_1) care dovedeşte că există o infinitate de soluţii (pentru orice valoare nenulă a lui a_1 există o valoare a_2 care să verifice ecuaţia şi gradul ecuaţiei este dat de min(r,c). Ecuaţiile ce se obţin pe calea substituţiei directe devin din ce în ce mai complicate cu creşterea lui `r` şi `c` şi cu coborârea dinspre relaţia (S^2) către relaţia (X^2).

Calea indirectă de rezolvare a ecuaţiilor de minimizare este prin aproximaţii succesive făcând apel la soluţia aproximativă oferită de estimarea iniţială. Astfel, se foloseşte relaţia de estimare iniţială pentru a obţine prima aproximaţie (aproximaţia iniţială) a soluţiei după care în fiecare succesiune de aproximaţii se înlocuiesc vechile valori ale aproximaţiei în partea dreaptă a relaţiilor pentru a obţine

noile aproximații.

În acest caz metoda aproximațiilor succesive ***converge rapid către soluția optimală***. Astfel pentru relația (S^2) trei iterații sunt suficiente pentru a obține (vezi Tabelul 13) o valoare reziduală de 282.11735 şi de la această iterație încolo valoarea reziduală îşi schimbă cifrele dincolo de a 5-a zecimală, în timp ce pentru relația (20) aceeaşi calitate a reprezentării soluției optimale este obținută după 4 iterații.

Folosind datele din [14] redate în Tabelul 13, valorile sugerate de ecuațiile de estimare inițială pentru produsele $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ sunt redate în Tabelul 14, valorile ce rezultă după rezolvarea iterativă a relațiilor (18)-(20) sunt redate în Tabelele 15-17.

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:

÷ TV: 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)

Tab.13. Valori experimentale în tratamentul cartofilor

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

Tab.14. Valorile produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$

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

Tab.15 Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (S^2)

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

Tab.16. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (CV^2)

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

Tab.17. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (X^2)

Tabelul 18 centralizează rezultatele obținute pe cele 4 căi.

	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**

Tabelul 18. Valori comparative pentru eroarea experimentală întâmplătoare

După cum se observă în Tabelul 18, fiecare dintre metodele definite de relațiile (S^2)-(X^2) îmbunătățeşte valoarea sumei obiectiv în raport cu expresia definită de formula aproximativă de estimare inițială şi reprezintă corecții ale acesteia.

Integrarea numerică

O foarte largă categorie de probleme din fizică şi chimie operează cu ecuații în care este necesar calculul unei integrale care nu poate fi evaluată formal, prin intermediul formulelor şi metodelor de integrare cunoscute sub numele de ***primitive***. În aceste situații se face apel la metodele de ***integrare numerică***. Metodele de integrare numerică sunt metode aproximative de evaluare a valorii integralei definite. Rezultatul integrării numerice poate fi evaluat aplicând un număr definit (în preambulul integrării) de paşi sau un număr de paşi care rezultă din aplicarea înseşi a algoritmului numeric de integrare.

Semnificația fizică a integralei definite este aria subgraficului funcției de integrat. În acest sens, figura următoare ilustrează două modalități de aproximare în care domeniul de integrare este împărțit în intervale de dimensiune egală.

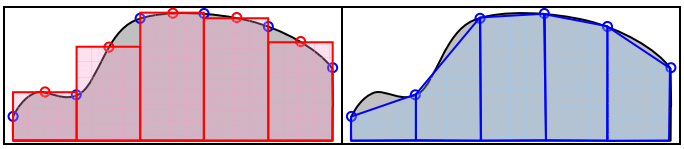


Fig.10. Aproximarea ariei prin dreptunghiuri şi trapeze

În integrarea numerică este totdeauna un compromis între timpul de calcul (mai multe diviziuni necesită mai multe evaluări şi deci mai mult timp) precizia dorită (față de care însă criteriul de apreciere a preciziei este uneori foarte dificil de apreciat având în vedere că valoarea reală a ariei de sub graficul funcției este necunoscut), limita preciziei de calcul (eroarea cea mai mare se face în jurul valorii de 1 prin eroarea de evaluare a acestuia cu 0.9...99 şi respectiv 1.0...01 şi nu în jurul valorii de 0 unde modalitatea de stocare în virgulă flotantă 1.2e-4 ne asigură o precizie foarte ridicată) şi propagarea erorii din operații de calcul succesive (dacă a=1.6 şi b=3.0 şi calculăm a/b cu o precizie de 2 cifre obținem 0.53 pierzând astfel restul cifrelor de 3 ale numărului rațional 0.5(3) şi calculele care urmează propagă şi amplifică această pierdere).

Metodele de cuadratură (sau Newton-Cotes [15, 16]) aproximează funcția de integrat cu un polinom. Din acest punct de vedere:

÷ aproximarea ariei prin dreptunghiuri reprezintă aproximarea funcției cu un polinom constant; pentru $x_i \le x \le x_{i+1}$ $f_i(x)=c_i$ ($c_i=f(x_i)$, $c_i=f(x_{i+1})$ sau $c_i=f(\varepsilon_i)$, $\varepsilon_i=(x_i+x_{i+1})/2$)

÷ aproximarea ariei prin trapeze reprezintă aproximarea funcției cu o funcție liniară; pentru $x_i \le x \le x_{i+1}$ $f_i(x)=\alpha \cdot x+\beta$, $\alpha=(f(x_{i+1})-f(x_i))/(x_{i+1}-x_i)$, $\beta=(f(x_i)\cdot x_{i+1}-f(x_{i+1})\cdot x_i)/(x_{i+1}-x_i)$, relații ce rezultă din condițiile $f_i(x_i)=\alpha \cdot x_i+\beta=f(x_i)$ şi $f_i(x_{i+1})=\alpha \cdot x_{i+1}+\beta=f(x_{i+1})$

Este de asemenea posibilă aproximarea prin polinoame de ordinul 2, 3 şi superior. În acest sens este exemplificat în continuare pentru polinoamele de ordinul 2 (regula 1/3 a lui Simpson [17]), existând însă şi regula 3/8 a lui Simpson de interpolare cu polinoame de ordinul 3 (regula 3/8 a lui Simpson) şi formule polinomiale de grad egal cu numărul de intervale.

Regula 1/3 a lui Simpson de integrare numerică

Fie un şir de sub-intervale în care se evaluează funcția f(x): $a=x_0, x_1, \ldots, x_n=b$, de lungimi egale: $x_{i+1}-x_i=(b-a)/n$ (pentru $0 \le i \le n-1$). Pe o pereche de intervale oarecare consecutive $[x_{i-1},x_i]$ şi $[x_i,x_{i+1}]$ expresia funcției se aproximează cu o funcție pătratică $g(x)=a\cdot x^2+b\cdot x+c$. Se pot identifica coeficienților a, b şi c direct în expresia integrată:

$$\int_{x_{i-1}}^{x_{i+1}} g(x)dx = \int_{x_{i-1}}^{x_{i+1}} (ax^2+bx+c)dx = \frac{a(x_{i+1}-x_{i-1})^3}{3} + \frac{b(x_{i+1}-x_{i-1})^2}{2} + \frac{c(x_{i+1}-x_{i-1})}{1}$$

dacă se foloseşte valoarea funcției în punctul x_i:

$$g(x_i) = g\left(\frac{x_{i-1}+x_{i+1}}{2}\right) = a\left(\frac{x_{i-1}+x_{i+1}}{2}\right)^2 + b\left(\frac{x_{i-1}+x_{i+1}}{2}\right) + c = \frac{a(x_{i-1}+x_{i+1})^2 + 2b(x_{i-1}+x_{i+1}) + 4c}{4}$$

Succesiunea de calcule este următoarea (unde G(x) este primitiva lui g(x)):

$$\int_{x_{i-1}}^{x_{i+1}} g(x)dx = \frac{x_{i+1}-x_{i-1}}{6}\left(2a({x_{i+1}}^2 + x_{i+1}x_{i-1} + {x_{i-1}}^2) + 3b(x_{i+1}+x_{i-1}) + 6c\right)$$

$$= \frac{x_{i+1}-x_{i-1}}{6}\left(a({x_{i+1}}^2 + 2x_{i+1}x_{i-1} + {x_{i-1}}^2) + 2b(x_{i+1}+x_{i-1}) + 4c + a({x_{i+1}}^2 + {x_{i-1}}^2) + b(x_{i+1}+x_{i-1}) + 2c\right)$$

$$= \frac{x_{i+1}-x_{i-1}}{6}\left(4g(x_i) + g(x_{i+1}) + g(x_{i-1})\right)$$

Exprimând valorile funcției g cu ajutorul funcției f în punctele x_{i-1}, x_i, x_{i+1} rezultă că aproximarea integralei funcției f cu integralele funcțiilor g se realizează aşadar cu ajutorul formulelor:

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx \cong \frac{x_{i+1}-x_{i-1}}{6}\left(f(x_{i-1}) + 4f(x_i) + f(x_{i+1})\right)$$

Însumând integralele de pe fiecare domeniu $[x_{i-1},x_{i+1}]$ se obține o expresie în care capetele

$[x_0,x_1]$ şi $[x_{n-1},x_n]$ sunt luate o singură dată iar celelalte intervale $[x_i,x_{i+1}]$, pentru $0<i<n-2$ sunt toate luate de 2 ori, adică:

$$\sum_{i=1}^{n-1}\int_{x_{i-1}}^{x_{i+1}} f(x)dx = 2\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx - \int_{x_0}^{x_1} f(x)dx - \int_{x_{n-1}}^{x_n} f(x)dx = 2\int_a^b f(x)dx - \int_{x_0}^{x_1} f(x)dx - \int_{x_{n-1}}^{x_n} f(x)dx$$

Expresia integralei funcției f pe întreg domeniul [a,b] se obține ca:

$$2\int_a^b f(x)dx = \sum_{i=1}^{n-1}\int_{x_{i-1}}^{x_{i+1}} f(x)dx - \int_{x_0}^{x_1} f(x)dx - \int_{x_{n-1}}^{x_n} f(x)dx$$

Relația de mai sus arată că evaluarea completă a integralei funcției f mai necesită o aproximare, şi anume a integralelor intervalelor de la capete ($[x_0,x_1]$ şi $[x_{n-1},x_n]$), care se poate face de exemplu cu metoda trapezelor:

$$\int_a^b f(x)dx \cong \frac{1}{2}\sum_{i=1}^{n-1}\frac{x_{i+1}-x_{i-1}}{6}\left(f(x_{i-1})+4f(x_i)+f(x_{i+1})\right) - \frac{1}{2}\int_{x_0}^{x_1} f(x)dx - \frac{1}{2}\int_{x_{n-1}}^{x_n} f(x)dx$$

$$\int_{x_0}^{x_1} f(x)dx \cong \frac{1}{2}\frac{f(x_1)+f(x_0)}{2}(x_1-x_0),\quad \int_{x_{n-1}}^{x_n} f(x)dx \cong \frac{1}{2}\frac{f(x_n)+f(x_{n-1})}{2}(x_n-x_{n-1})$$

Ținând cont că $x_{i+1}-x_i=(b-a)/n$,

$$\int_a^b f(x)dx \cong \frac{b-a}{6n}\left(\sum_{i=1}^{n-1} f(x_{i-1}) + 4\sum_{i=1}^{n-1} f(x_i) + \sum_{i=1}^{n-1} f(x_{i+1})\right) - \frac{1}{4}\frac{b-a}{n}\left(f(x_0)+f(x_1)+f(x_{n-1})+f(x_n)\right)$$

în care se ține seama acum că:

$$\sum_{i=1}^{n-1} f(x_{i-1}) = \sum_{i=0}^{n-2} f(x_i) = \sum_{i=0}^{n} f(x_i) - f(x_n) - f(x_{n-1})$$

$$\sum_{i=1}^{n-1} f(x_i) = \sum_{i=0}^{n} f(x_i) - f(x_0) - f(x_n)$$

$$\sum_{i=1}^{n-1} f(x_{i+1}) = \sum_{i=2}^{n} f(x_i) = \sum_{i=0}^{n} f(x_i) - f(x_1) - f(x_0)$$

$$\sum_{i=0}^{n} f(x_i) \overset{def}{=} S,\quad \int_a^b f(x)dx \overset{def}{=} I$$

de unde:

$$I \cong \frac{b-a}{n}\left(\frac{6S - f(x_n) - f(x_{n-1}) - 4f(x_0) - 4f(x_n) - f(x_1) - f(x_0)}{6} - \frac{f(x_0)+f(x_1)+f(x_{n-1})+f(x_n)}{4}\right)$$

$$I \cong \frac{(b-a)S}{n} - \frac{(b-a)}{12n}\left(13f(x_0)+5f(x_1)+5f(x_{n-1})+13f(x_n)\right)$$

Elemente de teoria probabilităților

Măsuri statistice pentru populații şi eşantioane

Frecvent operăm cu date provenite din măsurători repetate ale aceluiaşi fenomen. Aproape niciodată însă, două măsurători independente asupra aceleiaşi observabile, produce acelaşi rezultat. Din acest punct de vedere, odată încheiată observația fenomenului fizic, se pune problema prelucrării şi interpretării informației din spațiul informațional şi un alt tip de experiment trebuie derulat aici: ***experimentul statistic***.

Într-un experiment statistic operăm cu variabile aleatoare - o ***variabilă aleatoare*** fiind asociată unei măsurabile ale cărei valori sunt colectate în spațiul informațional. Variabila aleatoare are o ***serie de valori*** care corespund rezultatelor măsurătorilor înregistrate. Totdeauna în spațiul informațional vom poseda serii finite de valori în timp ce în spațiul fizic s-ar putea repeta experimentul de un număr infinit de ori. Din acest motiv, în statistică se diferențiază ***eşantionul*** - seria de valori colectată în spațiul informațional, de ***populație*** - întreaga mulțime infinită de valori care ar putea fi adusă din spațiul fizic printr-un proces continuu de măsurare.

Din punctul de vedere al reprezentării prin valori în spațiul informațional, avem ***variabile aleatoare discrete***, ale căror valori aparțin unei mulțimi finite (înzestrate sau nu cu o relație de ordine) sau infinite dar atunci înzestrate cu o relație de ordine şi numărabile (cum este mulțimea numerelor naturale, sau a celor întregi, sau a celor raționale) şi ***variabile aleatoare continue***, ale căror valori aparțin mulțimii numerelor reale (de puterea continuului) - şi asta chiar dacă, în spațiul informațional le stocăm tot folosind o mulțime de valori discrete şi ordonate (folosind o anumită precizie de exprimare a numerelor). O altă clasificare a tipului variabilelor aleatoare este în ***cantitative*** - dacă scala de măsură a fenomenului observat permite operația de adiție ("+") şi ***calitative*** - dacă nu e permisă (nu are sens).

Distribuții statistice în urma unor măsurători repetate

Aşa cum s-a anticipat la începutul secțiunii anterioare, o problemă la care statistica are o soluție de exprimare a rezultatului este cu privire la rezultatul urmând o serie de măsurători repetate. În tabelul următor (v. *Măsuri statistice pentru caracterizarea variabilelor cantitative*) sunt prezentate măsurile statistice ce rezultă din aplicarea pe populații şi respectiv pe eşantioane, pentru cazul variabilelor cantitative.

Măsură	Referă	Expresie	Interpretare
Suma valorilor	Un şir de numere	$\Sigma(\cdot)$	-
Numărul de valori		$\lvert\cdot\rvert$	-
Valoarea medie		$E(\cdot) = \Sigma(\cdot)/\lvert\cdot\rvert$	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	$VAR(Y) = \frac{\lvert Y\rvert}{\lvert Y\rvert - 1} var(Y)$	Împrăştierea
Estimatorul deviației standard a caracteristicii X		$S = S(Y) = \sqrt{\frac{\lvert Y\rvert}{\lvert Y\rvert - 1}} s(Y)$	Dispersia

Tabelul 19. Măsuri statistice pentru caracterizarea variabilelor cantitative

Un caz foarte frecvent în măsurătorile repetate este când valorile se distribuie 'normal' (vezi mai jos ***distribuția Gauss***), și din acest punct de vedere, este esențială caracterizarea depărtării de la normalitate (v. *Statistici pentru caracterizarea depărtării de normalitate a variabilelor cantitative*).

Simbol şi măsură	Referă	Expresie	Mărimi ce 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		$G_2 = B_2 - 3\cdot\frac{(n_Y-1)^2}{(n_Y-2)(n_Y-3)}$	

Tabelul 20. Statistici pentru caracterizarea depărtării de normalitate a variabilelor cantitative

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

Tabelul 21. Medii şi varianțe ale mediei şi varianței observabilei Y ce rezultă din distribuția de eşantionare din populația cu caracteristica X

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

Tabelul 22. Valori aproximative pentru mediile şi varianțele mediei şi varianței observabilei Y în ipotezele teoremei limită centrale

Un alt caz foarte important de distribuție este ***distribuția de eşantionare***. Extragerea repetată de eşantioane (de volum dat) dintr-o populație face ca valorile obținute să urmeze o distribuție, numită distribuția de eşantionare. Tabelul 21 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 [18], aceştia pot fi aproximați cu ajutorul estimatorilor acestora (Tabelul 19). Formulele de calcul aproximativ ale mediei şi varianței pentru medie şi varianță sunt redate în Tabelul 22.

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 19

estimatorii parametrilor statistici ai populației din măsurătorile (statisticile) efectuate asupra eşantionului.

Legi de distribuție şi statistici ale acestora

Tabelele 23-41 dau expresiile unor mărimi statistice (valabile pentru populație) în timp ce expresiile pentru estimatori se pot obține din Tabelul 19.

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 23. Mărimi statistice ale distribuției discrete uniforme

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, $1 \le k$
Media; varianța	p; p(1-p)	
Asimetria; excesul de boltire	0; $(6p^2-6p+1)/(p(1-p))$	

Tabelul 24. Mărimi statistice ale distribuției discrete Bernoulli

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 25. Mărimi statistice ale distribuției discrete binomiale

Mărime statistică	Expresie de calcul
Suport	$k = 0, 1, ...; \lambda \ge 0$
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 26. Mărimi statistice ale distribuției discrete Poisson

Mărime statistică	Expresie de calcul
Suport; Minim; Maxim	$x \in [a, b]$; 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 27. Mărimi statistice ale distribuției continue uniforme

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; Funcția de repartiție	$\frac{1}{\gamma\pi\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)}$; $\frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}$
Mediana şi moda	x_0

Tabelul 28. Mărimi statistice ale distribuției continue Cauchy-Lorentz

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 29. Mărimi statistice ale distribuției continue Student t

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

Tabelul 30. Mărimi statistice ale distribuției continue Fisher-Snedecor F

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 31. Mărimi statistice ale distribuției continue χ^2

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

Tabelul 32. Mărimi statistice ale distribuției continue exponențiale

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\lambda, k \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate; funcția de repartiție	$k x^{k-1} e^{-(x/\lambda)^k} / \lambda^k$; $1 - e^{-(x/\lambda)^k}$
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 33. Mărimi statistice ale distribuției continue Weibull

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	$\frac{1 + \operatorname{erf}\left((\ln(x)-\mu)/\left(\sigma\sqrt{2}\right)\right)}{2}$; $\operatorname{erf}(z) = 2\int_0^z e^{-t^2} dt \Big/ \sqrt{\pi}$
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 34. Mărimi statistice ale distribuției continue Log-normale

Mărime statistică	Expresie de calcul	
Suport	$b \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$	
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 35. Mărimi statistice ale distribuției continue Laplace (dublu exponențială)

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

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 37. Mărimi statistice ale distribuției continue Gamma

Mărime statistică	Expresie de calcul
Suport	$\beta \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$
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 38. Mărimi statistice ale distribuției continue Gumbel (log-Weibull)

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 39. Mărimi statistice ale distribuției continue Beta

Mărime statistică	Expresie de calcul
Suport	$\sigma \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$
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 40. 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

Tabelul 41. Alte mărimi statistice ale distribuției continue Gauss (normale)

Agrementul între observație și model: statistici

Fie $Y = (Y_1, ..., Y_n)$ și $X = (X_1, ..., X_n)$ serie de observații pereche și obiectivul să fie găsirea unei funcții $f(x, a_1, ..., a_m)$, pentru care $Y = f(X)$ este cea mai bună soluție posibilă a aproximare $\hat{Y} \sim Y$. Atingerea acestui obiectiv presupune găsirea expresiei funcției f și a valorilor parametrilor $a_1, ..., a_m$. Sub ipoteza de acord între observație și model, expresia funcției f se presupune a fi cunoscută (sau cel puțin ar trebui, atunci când se desfășoară o căutare după un anumit set de expresii alternative).

Astfel, a rămas obținerea valorilor parametrilor $a_1, ..., a_m$. Pentru a avea o soluție unică pentru valorile parametrilor $a_1, ..., a_m$, cel puțin se cere ca $m \leq n$ să fie asigurată. O serie de alternative sunt disponibile pentru aproximarea $Y \sim f(X)$, iar cele considerate cele mai importante sunt exemplificate în continuare.

Minimizarea erorii de acord (minimizarea dezacordului)

În această ipoteză o serie de alternative sunt disponibile (ecuația de mai jos, cu diferite alegeri pentru p și q):

$$S(p,q) = \sum_{i=1}^{n} |Y_i - f(X_i)|^p / f^q(X_i) = \min., \; q = 0, 1, {}^p/_2, p$$

În cazul în care seria Y reprezintă frecvența (nenulă) a observațiilor distincte X, atunci f(x) ar trebui să fie o funcție pozitivă de asemenea, și modul numărătorului nu mai este necesar.

Distribuția Gauss [19] a valorilor termenilor sumei sunt traduse în p = 2, și distribuția Laplace [20] apare când p = 1 (vezi și [21]). Funcția densitate de probabilitate (PDF) a unor reprezentanți ai familiei care conține distribuțiile Gauss și Laplace standard ($\mu = 0$ și $\sigma = 1$), sunt exemplificate în fig. 11. Minimizarea erorii de acord pentru valori p diferite oferă soluții diferite pentru parametrii, și cum se poate observa în figura 1, sunt asociate cu diferite forme de eroare.

$$GL(0;0.5)=\sqrt{15/2}\cong 2.739$$
$$GL(0;1.0)=\sqrt{1/2}\cong 0.707$$
$$GL(0;2.0)=1/\sqrt{2\pi}\cong 0.399$$
$$GL(0;3.0)=\ldots\cong 0.342$$
$$GL(0;4.0)=\Gamma^2(3/4)\cdot 2^{1/4}\cdot\pi^{-3/2}\cong 0.321$$
$$GL(x;p)=\frac{p}{2}\frac{\Gamma^{1/2}(3/p)}{\Gamma^{3/2}(1/p)}\exp\left(-\frac{|x|^p}{\left(\frac{\Gamma(1/p)}{\Gamma(3/p)}\right)^{p/2}}\right)$$

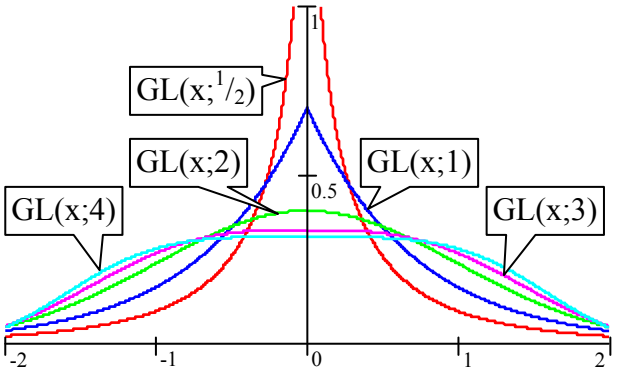


Fig.11. Distribuția Gauss-Laplace

Două cazuri particulare sunt de obicei utilizate pentru a estima parametrii necunoscuți ai unei distribuție atunci când p = 2 (vezi [22] şi [23]). Abordarea cea mai generală de obținere a parametrilor de distribuție este de a ghici valorile sau să aplice o procedură care reduce în fiecare etapă cantitatea dată de ecuația (1) până când cantitatea redusă este mult mai mică decât cea rămasă.

Utilizarea momentelor

Sub ipoteza că Y~f(X) ar trebui să fie chiar mai precisă (a doua ipoteză fiind caracterul aleatoriu al erorii, cu o medie de zero) şi aproximarea $\Sigma X_i^k Y_i \sim \Sigma X_i^k \cdot f(x_i)$ este folosită pentru k ≥ 0. Metoda momentelor dă greutate maximă la primele momente, astfel, o soluție $a_1, \ldots, a_m$ a Y~f(X) poate proveni direct din ecuația $\Sigma X_i^k Y_i = \Sigma X_i^k \cdot f(x_i)$. Modul cel mai convenabil pentru cazul general este repetarea căutării parametrilor $a_1, \ldots, a_m$ începând de la anumite valori inițiale):

$$\sum_{i=1}^{n} X_i^k Y_i \sim \sum_{i=1}^{n} X_i^k f(X_i),\ k = 0, 1, \ldots$$

Utilizarea momentelor centrale

Se întăreşte aproximarea $\Sigma X_i Y_i \sim \Sigma X_i \cdot f(X_i)$ şi $\Sigma (Y_i - \overline{Y})^k \sim \Sigma (X_i - \overline{X})^k \cdot f(X_i)$ pentru k≥2. Metoda momentelor dă greutate maximă primelor momente centrale, deci o soluție $a_1, \ldots, a_m$ a Y~f(X) poate proveni din ecuația:

$$\sum_{i=1}^{n} X_i Y_i \sim \sum_{i=1}^{n} X_i f(X_i) \text{ and } \sum_{i=1}^{n} (Y_i - \overline{Y})^k \sim \sum_{i=1}^{n} (X_i - \overline{X})^k f(X_i),\ k = 2, 3 \ldots$$

Folosind statisticile referitoare la populație

O uşoară modificare a metodei anterioare poate beneficia de disponibilitatea expresiei pentru media, deviația standard, asimetria şi excesul asimetriei populației în funcție de parametrii distribuției pentru un număr mare de distribuții bine cunoscute.

Estimare pe baza şansei maxime

Principiul şansei maxime este că o estimare rezonabilă pentru un parametru este cea care maximizează probabilitatea (P în ecuația de mai jos) de observare a datelor experimentale [24], iar setul mai probabil al parametrilor $a_1, \ldots, a_m$ va face P maxim (şi implicit MLE este maxim).

$$MLE = \log_2(P) = \sum_{i=1}^{n} \log_2(f(X_i))$$

Statistica Benford

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

Un şir de numere urmează distribuția Benford dacă probabilitatea de distribuție a unei cifre (d_i)

a numerelor ($d=d_0d_1\ldots$) reprezentate în baza de numerație b (uzual baza 10) urmează legea (Benford):

$$p(d_0) = \log_b\left(1+\frac{1}{d_0}\right),\ d_0 = 1..(b\text{-}1);$$

$$p(d_1) = \sum_{k=1}^{b-1}\log_b\left(1+\frac{1}{k\cdot b + d_1}\right),\ d_1 = 0..(b\text{-}1)$$

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

…

Fig.12. Statistica Benford

Ipoteza acestei legi de distribuție este că valorile măsurătorilor rezultate din observație sunt frecvent distribuite logaritmic şi astfel logaritmul setului de măsurători este distribuit uniform. Legea de distribuție este numită după fizicianul Frank BENFORD care a formulat-o intuitiv în 1938 [25], dar demonstrația acesteia a fost dată mult mai târziu [26].

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, prețurile acțiunilor, numerele populației, 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 numerație) are loc independent de baza de numerație în care se exprimă numerele, chiar dacă proporțiile de reprezentare se schimbă. De aici, ***acest rezultat poate fi folosit pentru a verifica datele în suspiciunea de alterare (mistificare) a acestora prin compararea frecvențelor teoretice cu cele observate pentru prima cifră a acestora***.

Statistica Jarque-Bera

Testul Jarque-Bera [27,28] calculează şi atribuie probabilitatea statistică ca valorile unui eşantion ce provine din populație normal distribuită să îşi abată simultan asimetria şi excesul de boltire de la valorile teoretice corespunzătoare distribuției normale.

Statistica Jarque-Bera se calculează cu relația:

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

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

Statistica JB are o distribuție 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 [29] poate fi folosit pentru verificarea ipotezei că un eşantion de date urmează o anumită lege de distribuție (redat în continuare), precum şi pentru compararea legilor de distribuție ale populațiilor din care provin două eşantioane [30].

Statistica Kolmogorov-Smirnov verifică dacă observațiile independente $X=(X_i)_{1\le i\le n}$ provin dintr-o populație ce urmează legea de distribuție dată de funcția cumulativă de probabilitate $F_t(x)$ prin calcularea maximumului diferenței absolute între $F_t(x)$ şi funcția cumulativă de probabilitate observată $F_o(x)$ în toate punctele observației:

$D = \max_{1\le i\le n}\left\|F_t(X_i) - F_o(X_i)\right\|$	(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 [31] 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 [32, 33]. Formulele de calcul şi interpretarea testului pentru compararea de eşantioane se găsesc la adresa [34].

În cazul comparației unei legi de distribuție discrete cunoscute cu legea de distribuție observată în eşantion varianța statisticii A^2 se calculează cu formula ([35], 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 [36]. 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

Experimentul variaţelor ce conduce la distribuţia χ^2

Distribuţia χ^2 a fost descoperită de Karl PEARSON [37] în urma încercării de a explica varianţa observată a numerelor care provin din distribuţia normală.

Astfel, dacă se consideră distribuţia normală standard N(0,1) şi variabila întâmplătoare X ce urmează această distribuţie (Figura 13), probabilitatea (dp) de a extrage valorile -x şi x din N(0,1) sunt ambele egale şi egale cu diferenţiala funcţiei de densitate de probabilitate a distribuţiei normale ($PDF_{N(0,1)}$).

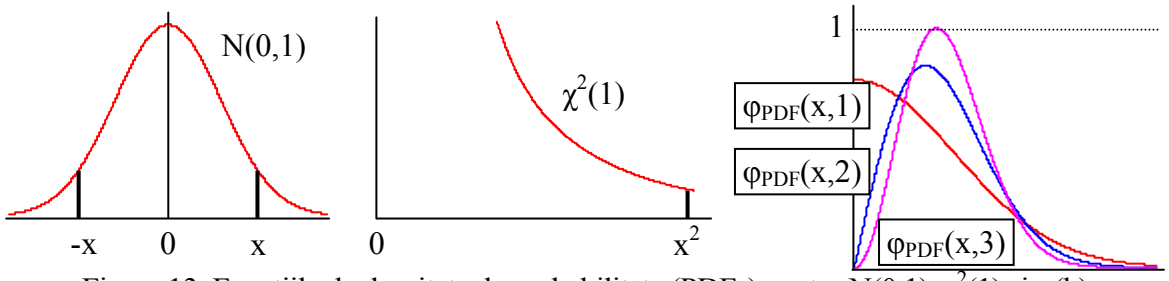


Figura 13. Funcţiile de densitate de probabilitate (PDFs) pentru N(0,1), $\chi^2(1)$ şi φ(k)

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

Distribuţia normală standard are media 0; astfel, pentru a exprima probabilitatea de observaţie pentru deviaţia x^2 trebuie adunate două probabilităţi (pentru -x şi x) date de relaţia (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 distribuţia Chi Square cu un grad de libertate (extragerea lui X din distribuţia normală). Dacă sunt extrase mai multe valori (k valori) din distribuţia normală atunci se obţine distribuţia Chi Square cu k grade de libertate, şi demonstraţia că ecuaţia (4) este adevărată poate fi găsită în [38].

Calea directă de la distribuţia normală la distribuţia χ^2 nu este reversibilă (Figura 13); astfel, definind variabila φ ca în relaţia (5) - ce reprezintă o expresie modificată a coeficientului de asociere definit de LIEBETRAU [39]:

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

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

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

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 distribuţiei 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 13).

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
- testează asocierea între două variabile cu valori grupate pe categorii;
- 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;
- 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);
- 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
- testează dacă mai multe populații sunt similare (sau omogene sau egale) în anumite caracteristici (acele caracteristici care sunt incluse în testare);
- 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);
- uzual caracteristica supusă observației este un moment central (ex. valoare medie, varianță);

÷ Testul Chi Square pentru verificarea asocierii în distribuție
- testează dacă un model teoretic poate fi asociat observațiilor;
- 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 frecvente 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 [37] 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, [40]), şi cu numărul parametrilor necunoscuți ai distribuției teoretice estimați din observații din măsuri ale tendinței centrale ([41]).

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):

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

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 şi ipoteză, testarea independenței şi a omogenității, definirea cadrului de aplicare al acesteia este dintre cele mai complexe [42].

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

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

÷ calcularea prin rotunjire a numărului de clase de frecvență din entropia Hartley [43] a observației vs. expectație: $\log_2(2N)$, unde N este numărul de observații (EasyFit [44] folosește această procedură);

÷ calcularea numărului de clase de frecvență odată cu lărgimea clasei folosind histograma ca estimator al densității [45] şi alegerea pe baza acesteia a criteriului optimal pentru lărgimea clasei (Dataplot [46] generează automat clasele de frecvență folosind această regulă: lărgimea clasei de frecvență este $0.3 \cdot s$ unde s este deviația standard a eşantionului; limitele inferioară şi superioară sunt date de medie $\pm 6 \cdot s$ şi clasele de frecvență observată 0 marginale sunt omise;

O a doua problemă este lărgimea claselor de frecvență; şi 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.

Măsuri ale tendinței centrale

Fie X un şir de n valori $X_1, X_2, \ldots, X_n$. Pentru calculul valorii medii, următorii indicatori sunt cei mai folosiți:

÷ Media Aritmetică, AM(X), dată de:

$$AM(X) = \frac{\sum_{i=1}^{n} X_i}{n}$$

÷ Media Geometrică, GM(X), obținută din expresia (de notat că pentru n par, expresia pentru GM poate fi nedeterminată, când produsul ΠX_i este negativ):

$$GM(X) = \sqrt[n]{\prod_{i=1}^{n} X_i} \underset{X_i>0}{=} \exp(AM(\ln(X))$$

÷ Media Armonică, HM(X), dată de:

$$HM(X) = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} = 1/AM(1/X)$$

÷ Media lui Euler, EM(X), calculată ca:

$$EM(X) = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} = \sqrt{AM(X^2)}$$

÷ Valoarea mediană, m(X), este numărul (π a.î. X_π este şir ordonat):

$$m(X)=\begin{cases}\left(X_{\pi(\frac{n}{2})}+X_{\pi(\frac{n}{2}+1)}\right)\Big/2, & \text{pt. } n \text{ par}\\ X_{\pi(\frac{n+1}{2})}, & \text{pt. } n \text{ impar}\end{cases}$$

÷ Valorile la modă sunt numerele date de:

$$\breve{X}=\left\{X_i \mid f_i=\sup\{f_j, 1\le j\le n\},\ f_j \text{ frecventa aparitiei lui } X_j\right\}$$

Binary and multinomial variables analysis: binomial confidence intervals

The origins of the mathematical study of natural phenomena are found in the fundamental work [47] of Isaac Newton [1643-1727].

Even if first studies about binomial expressions were made by Euclid [48], the mathematical basis of the binomial distribution study was put by Jacob Bernoulli [1654-1705], of which studies of especially significance for the theory of probabilities [49] was published 8 years later after his death by his nephew, Nicolaus Bernoulli. In *Doctrinam de Permutationibus & Combinationibus* section of this fundamental work he demonstrates the Newton binomial series expansion. Later, Abraham De Moivre [1667-1754] put the basis of approximated calculus for, using the normal distribution for binomial distribution approximation [50]. Later, Johann Carl Friedrich Gauss [1777-1855] with work [51] put the basis of mathematical statistics. Abraham Wald [1902-1950, born in Cluj] do his contributions on binomial confidence intervals approximation, elaborated and published the confidence interval that carry his name now [52].

Nowadays, the most prolific researcher on confidence intervals domain is Allan Agresti, which it was named the Statistician of the Year for 2003 by American Statistical Association, and at the prize ceremony (October 14, 2003) it spoken about Binomial Confidence Intervals [53, 54, 55, 56].

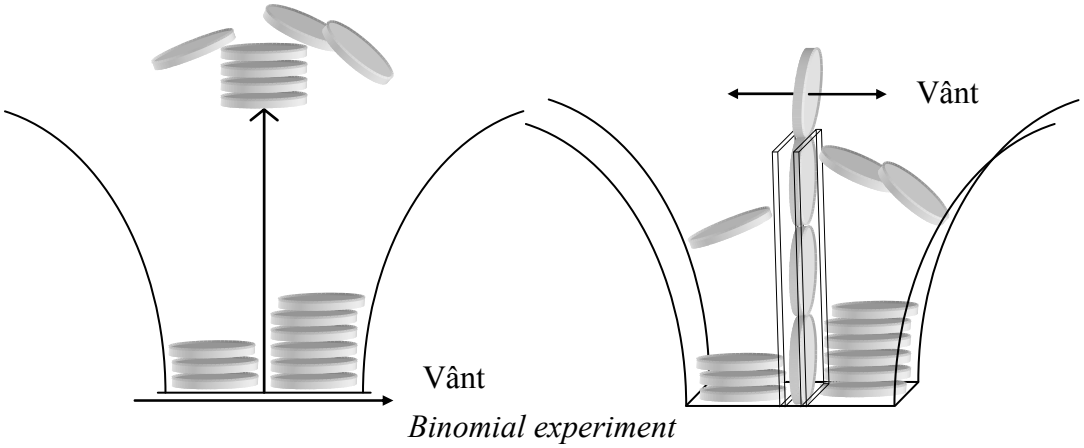
Binary observations always generate binomial distribution. A broad range of experiments are subject of binomials. Thus, law of binomial distribution are proved at heterometric bands of tetrameric enzyme in [57], the stoichiometry of the donor and acceptor chromophores implied in enzymatic ligand/receptor interactions in [58], translocation and exfoliation of type I restriction endonucleases in [59], biotinidase activity on neonatal thyroid hormone stimulator in [60], the parasite induced mortality at fish in [61], the occupancy/activity for proteins at multiple nonspecific sites containing replication in [62].

Finally, let us to give a reference to a very good essay about the frame of binomial distribution model applied to the natural phenomena [63].

(binomial design and model)

Let us give now a close look at the binomial experiment (see *Binomial experiment*). In a binary experiment, two values may come from observation, and may be encoded as 0 and 1 in the informational space (let's say if the coin goes to the left, then is recorded 0 or it goes to the right, and then is recorded 1. In a binomial experiment, we count the number of 0's and 1's from a series of n repetitions of binary experiments (let's say from n fallen coins in our *Binomial experiment*).

If from the total time *t* of running the experiment, wind has t_1 time the direction from left to right and t_2 the direction from right to left, the expectance is to found about $100 \cdot t_1/(t_1+t_2)$% coins in the right basket and $100 \cdot t_2/(t_1+t_2)$% coins in the left basket, but the truth is that we will see a such proportion only if we will spend enough time counting the coins. Going further, we may want to estimate the wind direction ratio from counting the coins, but then what level of confidence we have for our outcome?



Binomial experiment

The previous experiment is a construction meant to estimate a ratio between two real variables (times of wind having a certain direction) in which the estimation is given from a repeated binary experiment. As longer time we will spend on counting fallen coins as better accuracy will have on estimating times ratio. This *experiment* proofs that is no difference in the quality of the measurement encoded by binary values than the measurement encoded by the real values (e.g. any real number can be encoded as a succession of 0's and 1's as in our experiment. More than that, counting the coins can be even more accurate than any other instrument if the time goes to infinity.

Binomial distribution

Coins fallen	In the right basket (probability)	In the left basket (probability)	Explanation from observation of the right basket
1	0 (p)	1 (p)	0 = 0
	1 (1-p)	0 (1-p)	1 = 1
2	0 (p^2)	2 (p^2)	0 = 0 + 0
	1 (p·(1-p)·2)	1 (p·(1-p)·2)	1 = 0 + 1 = 1 + 0
	2 ($(1-p)^2$)	0 ($(1-p)^2$)	2 = 1 + 1
…	…	…	…
n	0 (p^n)	n (p^n)	0 = 0 + 0 + … + 0
	1 (p^{n-1}·(1-p)·(n))	n-1 (p^{n-1}·(1-p)·(n))	1 = 1 + 0 + … + 0 = … = 0 + … + 0 + 1
	…	…	…
	k	n-k	?
	…	…	…
	n-1 (p·$(1-p)^{n-1}$·(n))	1 (p·$(1-p)^{n-1}$·(n))	n-1 = 0 + 1 + … + 1 = … = 1 + … + 1 + 0
	n ($(1-p)^n$)	0 ($(1-p)^n$)	n = 1 + 1 + … + 1

?: The probability of falling k from n comes from the Newton's binomial expansion (and may be verified by induction):

$$\left(p+(1-p)\right)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Let's go back to the binomial experiment. The probability to see a certain number of coins in the right basket are related with wind speed ratio (or true proportion in the population) and the total number of fallen coins (see *Binomial distribution*).

According to [64] in general if p is a population parameter, T_1 a sufficient statistic in estimating p, and T_2 is any other statistic, the sampling distribution of simultaneous values of T_1 and T_2 must be such that for any given value of T_1, the distribution of T_2 does not involve p (if $f(p,T_1,T_2)dT_1dT_2$ - probability that T_1 and T_2 to fall in ranges dT_1 and dT_2 - it separates as follows: $f(p,T_1,T_2) = f_1(p,T_1)\cdot f_2(T_1,T_2)$.

The probability of drawing in order any particular sample $x_1, x_2, \ldots, x_n$ (of 0's and 1's) is:

$$p^{x_1}(1-p)^{1-x_1}\cdot p^{x_2}(1-p)^{1-x_2}\cdot\ldots\cdot p^{x_n}(1-p)^{1-x_n} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$$

This quantity can be divided in two factors:

$$p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} = \frac{n!}{(\Sigma x_i)!(n-\Sigma x_i)!} p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \cdot \frac{(\Sigma x_i)!(n-\Sigma x_i)!}{n!}$$

of which the first represent the probability that the actual total Σx_i should have been scored, and the second the probability, given this total, that the partition of it among the n observations should be actually observed. In the latter factor, p, the parameter sought, does not appear. Now when the mean Σx_i is known, any further information which the sample has to give depends on the observed partition ($\Sigma x_i/(n-\Sigma x_i)$), but the probability of any particular distribution is wholly independent of the value of p.

The parameter p can be estimated from the mean of the observed sample by using the Maximum Likelihood Estimation (MLE) method [65]:

$$\mathrm{MLE} = \ln\left(\frac{n!}{(\Sigma x_i)!(n-\Sigma x_i)!}p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}\right) = \max.\ \text{then}\ \frac{\partial \mathrm{MLE}}{\partial p} = \frac{\partial \mathrm{MLE}}{\partial n} = 0$$

The parameters *n* and *p* are independent, and then from $\partial MLE/\partial p = 0$ it results:

$$\frac{\partial}{\partial p}\ln\left(\frac{n!}{(\Sigma x_i)!(n-\Sigma x_i)!}\right) + \frac{\partial}{\partial p}\ln\left(p^{\Sigma x_i}\right) + \frac{\partial}{\partial p}\ln\left((1-p)^{n-\Sigma x_i}\right) = 0$$

This previous equation gives a relationship between estimators under assumption of the maximum likelihood:

$$\frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p} = 0,\ \text{from which}\ \hat{p}\cdot\hat{n} = \Sigma x_i$$

Solution of the second derivative ($\partial MLE/\partial n = 0$) has no analytical close form and can be computed only numerically [66]. Thus, if we assume that we draw always *n* objects, then $T_1 = \Sigma x_i$ is a sufficient statistics.

More, if we obtain an estimate of $p, \hat{p}$, from $\hat{p}\cdot n = \Sigma x_i$, then (as long is the solution of MLE) then:

$$\mathrm{var}(\hat{p}) = \left(-\left(\frac{\partial^2 \mathrm{MLE}}{\partial p^2}\right)\right)^{-1} = \left(\frac{\Sigma x_i}{p^2} + \frac{n-\Sigma x_i}{(1-p)^2}\right)^{-1} =_{\hat{p}\cdot n=\Sigma x_i} \left(\frac{\hat{p}\cdot n}{p^2} + \frac{n-\hat{p}\cdot n}{(1-p)^2}\right)^{-1} =_{p=\hat{p}} \frac{p(1-p)}{n}$$

Confidence for a binomial proportion is relatively simple to be expressed as long as we enumerate entire probability space of the binomial distribution.

A variable (Y) confined to the whole domain of values (from 0 to n) is binomial distributed if the probability of its taking any particular value *j* is:

$$P_B(Y=j) = \frac{n!}{j!(n-j)!}p^j(1-p)^{n-j}$$

When we already conducted a previous experiment in which we seen *i* objects on the right from *m* objects in total, then we can use *i*/*m* as an estimate for *p* and the probability of *Y* taking *j* values on the right become:

$$P_B(Y=j,n,i/m) = \frac{n!}{j!(n-j)!}\left(\frac{i}{m}\right)^j\left(1-\frac{i}{m}\right)^{n-j}$$

(binomial confidence intervals)

In order to collect the probabilities given by the previous relation and to express a 95% (or other threshold) confidence interval we must solve a combinatorial problem since $P_B(j)$ function is not monotone nor continuous (see *$P_B(j,n,7/10)$ binomial probability table*).

$P_B(j,n,7/10)$ binomial probability

j	0	1	2	3	4	5	6	7	8	9	10
$P_B(j)$	$5.9\cdot10^{-6}$	$1.4\cdot10^{-4}$	$1.4\cdot10^{-3}$	$9.0\cdot10^{-3}$	$3.7\cdot10^{-2}$	$1.0\cdot10^{-1}$	$2.0\cdot10^{-1}$	$2.6\cdot10^{-1}$	$2.3\cdot10^{-1}$	$1.2\cdot10^{-1}$	$2.8\cdot10^{-2}$
OC	1	2	3	4	6	7	9	11	10	8	5
$\Sigma_{i\le OC}$	$5.9\cdot10^{-6}$	$1.4\cdot10^{-4}$	$1.6\cdot10^{-3}$	$1.1\cdot10^{-2}$	$7.6\cdot10^{-2}$	$1.8\cdot10^{-1}$	$5\cdot10^{-1}$	$1.0\cdot10^{0}$	$7.3\cdot10^{-1}$	$3.0\cdot10^{-1}$	$3.9\cdot10^{-2}$
j	0	1	2	3	4	5	6	7	8	9	10
OC: Order by chance (from unlikely to likely); $^{95\%}$CI(j,10,7/10) = [4,9]; Err([4,9]) = 3.9% (<5%)											

It is a fact that we cannot express a mathematical formula indicating how many numbers from left and how many from right (see *$P_B(j,n,7/10)$ binomial probability table*) should be excluded in general when we express a confidence interval from a binomial experiment. For this reason usually expressed confidence intervals uses approximating formulas. Following table (see *Binomial and Binomial-like confidence intervals****) lists the most known of it.

*Binomial and Binomial-like confidence intervals****

Group	Name	Method	Acronym*	Refs
Normality	Wald	Classic	Wald_N	[52],[67], [68], [69]

Group	Name	Method	Acronym*	Refs
approximation		Continuity corrected	Wald_C	[70]
	Agresti-Coull	Classic	A_C_N	[53]
		Continuity corrected	A_C_C	[70]
		Continuity corrected	A_C_D	-**
	Wilson	Classic	Wilson_N	[71]
		Continuity corrected	Wilson_C	[68]
Harmonic approximation	ArcSine	Classic	ArcS_N	[72]
		Continuity corrected	ArcS_C	[70]
		Continuity corrected	ArcS_D	[70]
		Continuity corrected	ArcS_E	-
Log-normality approximation	Logit	Classic	Logit_N	[73]
		Continuity corrected	Logit_C	[74]
Binomial approximation	Bayes (Fisher)	Classic	BetaC11	[75]
	Clopper-Pearson	Classic	BetaC01	[76], [77]
	Jeffreys	Classic	BetaCJ0	[78]
	BetaC00	Continuity corrected	BetaC00	-
	BetaC10	Continuity corrected	BetaC10	-
	BetaCJ1	Continuity corrected	BetaCJ1	-
	BetaCJ2	Continuity corrected	BetaCJ2	-
	BetaCJA	Continuity corrected	BetaCJA	-
Obtained from optimization	Blyth-Still-Casella	Probabilistic optimization	B_S_C	[79], [80]
	OptiBin	Numerical optimization	OptiBin	[81]
	ComB	Combinatorial enumeration	ComB	[82], [83]

* According to http://l.academicdirect.org/Statistics/confidence_intervals/;
** New corrections
** English translation of Table 1 p. 42 from [84]

(contingency of binomials)

A more complex problem occurs when expressions containing proportions requires confidence intervals. This is the common case when ratios or differences of proportions express a relative or absolute measure of comparison between two populations, or two treatments on populations. This sort of expressions is often used in medical and biological studies [84].

The general shape of a study involving two binomial proportions is given by a 2X2 contingency table (see *2X2 contingency*):

2X2 contingency

Experiment 1\2	Left	Right
Up	a	b
Down	c	d

Depending on the type of the experiment, a list of six binomial expressions involving two binomial variables may comprise the currently reported in the literature [84] (see *Expressions involving two binomials*) - where the place of X and Y in the contingency table may vary (for example one arrangement is when X = a, m = a+b, Y = c, n = c+d and other one is when X = a, m = a+c, Y = b, n = b+d).

Expressions involving two binomials

Function	f3	f4	f5	f6	f7	f8
Expression	$\frac{X(n-Y)}{Y(m-X)}$	$\frac{Y}{n}-\frac{X}{m}$	$\left\|\frac{Y}{n}-\frac{X}{m}\right\|$	$1/\left\|\frac{Y}{n}-\frac{X}{m}\right\|$	$\frac{Xn}{Ym}$	$\left\|1-\frac{Xn}{Ym}\right\|$

(multinomial distribution)

If the observations are split into the informational space in more than two categories then knowing as a fact that we observe *n* objects which falls in *k* classes with the probabilities p_1, p_2, …, p_k ($1 = p_1 + p_2 + \ldots + p_k$) then the probability $P_M(x_1, \ldots, x_n; n; p_1, \ldots, p_k)$ to observe a certain distribution of the objects ($n = x_1 + x_2 + \ldots + x_k$) in the classes (1, …, k) is:

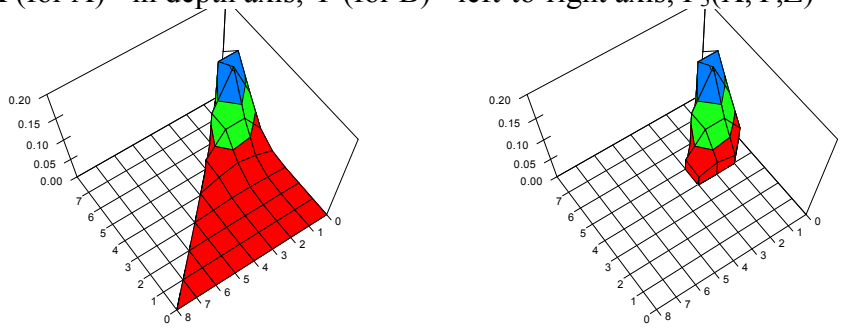
$$P_M(x_i; n; p_i) = n! \prod_{j=1}^{k} p_j^{x_j} \Big/ x_j!$$

We have many multinomial distributions when we discuss about the distribution of living organisms in a certain area, for example distribution of bacteria species belonging on a class of bacteria in a Petri plate after treatment with an antibiotic. Also for these cases we want to express with a certain level of confidence, the effect of the treatment, for example.

Our knowledge about expressing confidence for expressions containing binomial variables can be extent to cover trinomial (see *Trinomial(6A,1B,1C) - distribution and 95% coverage confidence interval*) or multinomial variables as well [85].

Trinomial(6A,1B,1C) - distribution and 95% coverage confidence interval

Depicted: X (for A) - in depth axis, Y (for B) - left-to-right axis, $P_3(X,Y,Z)$ - vertical axis



Ordinal variables analysis: ranks statistic

In not at all few cases the measurement or observation has the purpose to arrange a number of individuals (such as molecules) in order according to some quality which they possess to a varying degree (such as a molecular structure descriptor or compound chromatographic elution time). The arrangement in which every member of the whole has a rank are called ranking.

Ranked material can arise in many different ways [86]:

÷ Purely arrangement of objects considered only by reference to their position in space or time;

÷ According to some quality which we cannot measure on an objective scale. "If A scratches B when the two are rubbed together" gives a scale of "hardness" then this is such kind of scale and, by transitivity, we may only rank the hardness;

÷ According to some measurable or countable quality. We may rank molecules according to melting point or to the number of molecular formula isomers. It may not always be necessary to carry out the actual measurements in such cases, as for instance, if we arrange the molecules by the number of atoms standing in place of the number of isomers;

÷ According to some quality which we believe to be measurable but cannot measure for practical or theoretical reasons; the electronegativity we cannot measure, but we can conduct practical or theoretical experiments to see near to are located the binding electrons.

We can always rank a series of quantitative measurements according to their position on the scale. We may replace then the values with their ranks. Disadvantage of ranks use in the place of the values is because (and when) the information regarding the closeness between individuals is omitted. Compensating accuracy lose, the advantage are found in generality, because with ranks we gain the independence of ranks on stretching the scale of measurement - ranking is invariant under stretching the scale of measurement.

(order statistics)

If X (X_1, …, X_n) is a random sample coming from a continuous distribution f_X (with F_X=CDF(f_X)) then U = F_X(X) comes from the standard uniform distribution ($U_1 = F_X(X_1)$, …, $U_n = F_X(X_n)$). We will note through $U_{(k)}$ the k-th value in the (ascending) ordered series of U.

By splitting the [0,1] interval in three parts, and counting the number of cumulative probabilities $U_{(k)}$ ($1 \le k \le n$) falling in these three intervals we actually count the number of objects in a multinomial distribution with three categories (see *Probability of order statistic falling in an interval*).

If we have two random samples coming from (two) continuous distribution functions, and $U_{(i)}=F_X(X_{(i)})$ and $V_{(j)}=F_Y(Y_{(j)})$ the associated order statistics, then the joint probability density function of the two order statistics $U_{(i)}<U_{(j)}$ are constructed from a multinomial distribution with five categories (see *Probability of joint order statistics falling in an interval*).

When distribution functions are known (we known or we derive what distribution function has X and Y) the above obtained relations may allow us to obtain the PDF of the order statistics (u=F_X(x), v=F_Y(y), du=f_X(x)dx, dv=f_Y(y)dy) - see *Probability distribution function of order statistics*.

Probability of order statistic falling in an interval

Interval	[0,u]	(u,u+du]	(u+du,1]
Probability	u	du	1-u-du
Objects falling in	k-1	1	n-k

$$P_3(U_{(k)};n;u)=\frac{n!}{(k-1)!\cdot 1!\cdot(n-k)!}u^{k-1}\cdot du\cdot(1-u-du)^{n-k}\cong\frac{n!}{(k-1)!\cdot(n-k)!}u^{k-1}(1-u)^{n-k}du$$

Probability of joint order statistics falling in an interval

Interval	[0,u]	(u,u+du]	(u+du,v]	(v,v+dv]	(v+dv,1]
Probability	u	du	v-u	dv	1-v-dv
Objects falling in	i-1	1	j-i-1	1	n-j

$$P_5(U_{(i)} < U_{(j)}; n; u) = \frac{n!}{(i-1)!\cdot 1!\cdot (j-i-1)!\cdot 1!\cdot (n-j)!} u^{i-1}\cdot du\cdot (v-u)^{j-i-1}\cdot dv\cdot (1-v-dv)^{n-j}$$

Probability distribution function of order statistics

Expression	Meaning
$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\cdot(n-k)!}(F_X(x))^{k-1}(1-F_X(x))^{n-k} f_X(x)$	Probability to see the k-th value in the ordered (ranked) list of observations
$f_{X_{(i)},X_{(j)}}(x,y) = n!\frac{(F_X(x))^{i-1}}{(i-1)!}\frac{(F_X(y)-F_X(x))^{j-i-1}}{(j-i-1)!}\frac{(1-F_X(y))^{n-j}}{(n-j)!}dxdy$	Probability to see the i-th and j-th values in the ordered (ranked) list of observations

(Fisher-Yates correlation coefficient) If $(X_i,Y_i)_{1\le i\le n}$ are already sorted by X values and U_i=rank(X_i) and V_i=rank(Y_i) then (and if) U = (1, 2, …, n) - meaning that are no tied X values - the Fisher-Yates correlation coefficient are defined by ($\xi(i|n)$ is normal order statistic - the expected value of the i-th largest standardized deviate in a sample of size n from a normal population):

$$r_{FY} = \frac{\sum_{1\le i\le n}\xi(i\,|\,n)\xi(v_i\,|\,n)}{\sum_{1\le i\le n}\xi^2(i\,|\,n)}$$

(generalized correlation coefficient)

For *n* pairs of values (x_i,y_i) we may construct two matrices $(a_{i,j})$ and $(b_{i,j})$ for defining scores and possessing two properties (see *Scores for correlation coefficients*) - $a_{i,i} = b_{i,i} = 0$, $a_{i,j} = a_{j,i}$, $b_{i,j} = b_{j,i}$.

Following table gives the generalized correlation coefficient [86] and three particular cases of it (See *Pearson, Spearman and Kendall correlation coefficients*).

Scores for correlation coefficients

$(a_{i,j})$					$(b_{i,j})$				
x_i vs. y_j	y_1	y_2	…	y_n	y_j vs. x_i	x_1	x_2	…	x_n
x_1	0	$a_{1,2}$	…	$a_{1,n}$	y_1	0	$b_{1,2}$	…	$b_{1,n}$
x_2	$-a_{1,2}$	0	…	$a_{2,n}$	y_2	$-b_{1,2}$	0	…	$b_{2,n}$
…	…	…	0	…	…	…	…	0	…
x_n	$-a_{1,n}$	$-a_{2,n}$	…	0	y_n	$-b_{1,n}$	$-b_{2,n}$	…	0

Pearson, Spearman and Kendall correlation coefficients

Generalized formula	Pearson [87]	Spearman [88]	Kendall [89]
$Q = \frac{\Sigma a_{i,j}b_{i,j}}{\sqrt{\Sigma a_{i,j}^2}\sqrt{\Sigma b_{i,j}^2}}$	$a_{i,j} = x_j - x_i$ $b_{i,j} = y_j - y_i$	$a_{i,j} = \text{rank}(x_j) - \text{rank}(x_i)$ $b_{i,j} = \text{rank}(y_j) - \text{rank}(y_i)$	$a_{i,j} = \text{sign}(x_j - x_i)$ $b_{i,j} = \text{sign}(y_j - y_i)$

A correlation coefficient should (and these three it) follow a list of rules:

÷ If the agreement is perfect (if $x_i=y_i$, $1\le i\le n$) then Q should be 1 and if the disagreement is perfect (if $x_i=y_{n-i}$, $1\le i\le n$) then Q should be -1; indeed, $(\Sigma a_{i,j}b_{i,j})^2 \le (\Sigma a_{i,j})^2(\Sigma b_{i,j})^2$ - being Cauchy- Bunyakovsky-Schwarz inequality [90, 91, 92];

÷ Increasing values from -1 to 1 should reflect increasing agreement between (x_i) and (y_i); this

should be made by construction of $(a_{i,j})$ and $(b_{i,j})$ matrices.

(Jackknife method) Suppose we are interested in estimating some parameter θ using $\hat{\theta} = f(x_1,...,x_n)$ where (x_1, …, x_n is a sample of *n* independent observations with cumulative distribution function F(θ,X). Assuming that $\hat{\theta}$ is a good estimate of θ, then following steps (see *Jackknife algorithm*) gives us so called jackknife estimate of first order (then the bias is of O(1/n) order) [93]:

Jackknife algorithm

- For each i = 1..n
  - Remove x_i and compute $\hat{\theta}_{-i} = f(x_1,...,x_{i-1},x_{i+1},...,x_n)$ from reminders;
  - Compute pseudo value $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$;
- EndFor

In order to remove high order bias we must proceed further to remove all pairs of size 2, 3, and so on [94].

For groups up to size k, the bias is removed to order $O(1/n^k)$ and the formula for the k-th order jackknife is obtained ($\hat{\bar{\theta}}_{(j)}$ the mean of the estimates based on removing groups of size j). Based on computed values, Jackknife estimates for the parameter and its variance are obtained (see *Jackknife estimates*).

Jackknife estimates

Estimates using	θ	$Var_{est}(\theta)$
first order jackknife	$J_n^1(\theta) = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$	$SE^2(J_n^1(\theta)) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\theta}_i - J_n^1(\theta))^2$
second order jackknife	$J_n^2(\theta) = \frac{1}{C_2^n}\sum_{1\le i<j\le n}\hat{\theta}_{i,j}$	$SE^2(J_n^2(\theta)) = \frac{1}{C_2^n(C_2^n-1)}\sum_{1\le i<j\le n}(\hat{\theta}_{i,j} - J_n^2(\theta))^2$
k-th order jackknife	$J_n^k(\theta) = \frac{1}{C_k^n}\sum_{1\le i_1<...<i_k\le n}\hat{\theta}_{i_1,...,i_k}$	$SE^2(J_n^k(\theta)) = \frac{1}{C_k^n(C_k^n-1)}\sum_{1\le i_1<...<i_k\le n}(\hat{\theta}_{i_1,...,i_k} - J_n^k(\theta))^2$
combined till k-th order	$\tilde{J}_n^k(\theta) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(n-j)^k J_n^j(\theta)$	probability from p_1, p_2, …, p_k using [95]: "combining independent tests of significance"

(Bootstrap method) Suppose we are interested in estimating some parameter θ using $\hat{\theta} = f(x_1,...,x_n)$ where (x_1, …, x_n is a sample of *n* independent observations with cumulative distribution function F(θ,X). Assuming that $\hat{\theta}$ is a good estimate of θ, then a series of steps gives (- see *Bootstrap algorithm*) us the bootstrap estimates [96].

Bootstrap algorithm

- Construct empirical probability distribution $\hat{F}$ giving weight of 1/n to each x_i (this is bootstrap population):

$$\hat{F}\begin{matrix}\text{Value}\\ \text{Probability}\end{matrix}\begin{pmatrix}x_1 & ... & x_n\\ 1/n & 1/n & 1/n\end{pmatrix}$$

- For each i = 1..N (50≤N≤200):
  - Extract a sample (bootstrap sample) of same size n (as the initial sample) from the bootstrap population using discrete uniform distribution (Random(1..n));
  - Calculate the estimate $\hat{\theta}_{(i)}$ of θ for the sample;
- EndFor

Following table gives the bootstrap estimates (see *Bootstrap estimates*).

Bootstrap estimates

θ	$Var_{est}(\theta)$
$B_n(\theta) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{(i)}$	$SE^2(B_n(\theta)) = \frac{1}{N(N-1)}\sum_{i=1}^{N}(B_n(\theta)-\hat{\theta}_{(i)})^2$

(tied values and fractional ranking)

Coming back to our example relating number of atoms with the number of isomers, following table gives the number of isomers for first X alkanes (see Ranking the isomers of alkanes):

Ranking the isomers of alkanes

MF	CH_4	C_2H_6	C_3H_8	C_4H_{10}	C_5H_{12}	C_6H_{14}	C_7H_{16}	C_8H_{18}	C_9H_{20}	$C_{10}H_{22}$	$C_{11}H_{24}$	$C_{12}H_{26}$	$C_{13}H_{28}$	$C_{14}H_{30}$
NC	1	2	3	4	5	6	7	8	9	10	11	12	13	14
IS	1	1	1	2	3	5	9	18	35	75	159	355	802	1858
r_{NC}	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_{IS}	2	2	2	4	5	6	7	8	9	10	11	12	13	14
MF: Molecular formula; NC: Number of C atoms; IS: Number of structural isomers														

A lot of particular cases do not provide necessary information to distinguish between different objects, such as in our example of ranking alkanes isomers. Even more, we wish to measure the degree of correspondence between these two things and we may involve ranks since the relationship is clear to not be linear.

When ranking are conducted using a given property only (as number of structural isomers), then we are not able to distinguish between the objects (as between methane, ethane and propane) and we should assign the same rank. We can do this keeping constant the sum of ranks by averaging all equal rank values (as in: Ranking the isomers of alkanes). This sort of ranking may provide fractional ranks.

(Spearman ρ)

For given set of *n* objects with two measured characteristics *X* and *Y* ranked as $(q_i)_{1\le i\le n}$ and $(r_i)_{1\le i\le n}$ using fractional ranking following formula defines Spearman ρ correlation coefficient [88], where $diff_i = q_i - r_i$ (see Isomers of alkanes - Spearman ρ correlation measures):

$$\rho = 1 - \frac{6\sum_{i=1}^{n} diff_i^2}{(n^3 - n)}$$

Isomers of alkanes - Spearman correlation measures

MF	CH_4	C_2H_6	C_3H_8	C_4H_{10}	C_5H_{12}	C_6H_{14}	C_7H_{16}	C_8H_{18}	C_9H_{20}	$C_{10}H_{22}$	$C_{11}H_{24}$	$C_{12}H_{26}$	$C_{13}H_{28}$	$C_{14}H_{30}$
NC	1	2	3	4	5	6	7	8	9	10	11	12	13	14
IS	1	1	1	2	3	5	9	18	35	75	159	355	802	1858
r_{NC}	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_{IS}	2	2	2	4	5	6	7	8	9	10	11	12	13	14
diff	-1	0	1	0	0	0	0	0	0	0	0	0	0	0
$diff^2$	1	0	1	0	0	0	0	0	0	0	0	0	0	0
$\Sigma_i diff_i^2 = 2$; $T = 0$; $U = 3^3-3 = 24$; $\rho = 2718/2730$; $\rho_a = 2706/2730$; $\rho_b = \sqrt{2706}/\sqrt{2730}$														
MF: Molecular formula; NC: Number of C atoms; IS: Number of structural isomers														

When are sets of ties in the rankings then we must count the total number of their pairs for *X* ($T = \Sigma_i(t_i^3 - t_i)$) and for *Y* ($U = \Sigma_j(u_j^3 - u_j)$). Two formulas are derived in the presence of ties T and U [86]:

$$\rho_a = \frac{(n^3 - n) - 6\sum_{i=1}^{n} diff_i^2 - \frac{T+U}{2}}{(n^3 - n)}, \quad \rho_b = \frac{(n^3 - n) - 6\sum_{i=1}^{n} diff_i^2 - \frac{T+U}{2}}{\sqrt{(n^3 - n) - T}\sqrt{(n^3 - n) - U}}$$

The distribution of ρ_a is approximately normal. Two approximate statistics are available:

÷ $\mathrm{var}(\rho_a)=\frac{1.060}{n-3}$, $z_{\rho a}=\sqrt{\frac{n-3}{1.06}}z_r$, $z_r=\frac{1}{2}\ln\frac{1+r}{1-r}$ as gives in [97];

÷ $t=r\sqrt{\frac{n-2}{1-r^2}}$ as given in [98];

(Kendall τ)

For given set of *n* objects with two measured characteristics *X* and *Y* ranked as $(q_i)_{1\le i\le n}$ and $(r_i)_{1\le i\le n}$ the number C of concordant and D of discordant pairs using fractional ranking must be computed (see *XY concordance contingency*).

XY concordance contingency

C=cnn+cpp		(Y_j-Y_i)		
D=dpn+dnp		<0	=0	>0
(X_j-X_i)	<0	cnn	tny	dnp
	=0	txn	txy	txp
	>0	dpn	tpy	cpp

Then following formula defines Kendall τ_a and τ_b correlation coefficients [99], where τ_b is the correction for ties:

$$\tau_a=\frac{2(C-D)}{(n^2-n)},\ \tau_b=\frac{2(C-D)}{\sqrt{(n^2-n)-\Sigma_i t_i(t_i-1)}\sqrt{(n^2-n)-\Sigma_j u_j(u_j-1)}},\ \tau_c=\frac{2(C-D)}{n^2\frac{\min(\mathrm{rows},\mathrm{cols})-1}{\min(\mathrm{rows},\mathrm{cols})}}$$

The sampling distribution of τ_a has an expected value of zero. The distribution of τ_a cannot be expressed in a close form. May be calculated exactly for small samples and it is common to use an approximation to the normal distribution, for larger samples, with zero mean and variance given in the next formula and $z_{\tau a}$ statistic approximates the distribution of τ_a [86]:

$$\mathrm{var}(\tau_a)\cong\frac{2(2n+5)}{9n(n-1)},\ z_{\tau a}=\frac{\tau_a}{\sqrt{\mathrm{var}(\tau_a)}}=\frac{2(C-D)}{n(n-1)}\frac{\sqrt{9n(n-1)}}{\sqrt{2(2n+5)}}=\frac{3(C-D)}{\sqrt{n(n-1)(2n+5)/2}}$$

$$\text{or [97]: } \mathrm{var}(\tau_a)\cong\frac{0.437}{n-4}$$

The following statistic $z_{\tau b}$ provides an approximation for the τ_b distribution:

$$z_{\tau b}=\frac{C-D}{s_{CD}},\ {s_{CD}}^2=\frac{v_0-v_t-v_u}{18}+v_1+v_2$$

$$v_0=n(n-1)(2n+5),\ v_t=\Sigma_i t_i(t_i-1)(2t_i+5),\ v_u=\Sigma_j u_j(u_j-1)(2u_j+5)$$

$$v_1=\frac{\left(\Sigma_i t_i(t_i-1)\right)\left(\Sigma_j u_j(u_j-1)\right)}{2n(n-1)},\ v_2=\frac{\left(\Sigma_i t_i(t_i-1)(t_i-2)\right)\left(\Sigma_j u_j(u_j-1)(u_j-2)\right)}{9n(n-1)(n-2)}$$

When we assess the correlation between the properties of more than 10 objects in the expression of z_τ we should decrease the absolute value of 2(C-D) with one unit since the frequency of it is only in paired points. For example [86] when *n = 9* and *2(C-D) = 20* the exact probability is 4.4%, the not corrected one is 3.7% and the corrected one is 4.8%, and thus with continuity corrected expression we obtain a better approximation when $n \ge 10$.

(***Goodman and Kruskal's γ***) By using same notation as for Kendall τ, gamma correlation coefficient is given by [100]:

$$\gamma=\frac{C-D}{C+D}$$

The γ coefficient approximately follows a Student t distribution with (n-1) degrees of freedom [101]:

$$t_\gamma\cong\gamma\sqrt{\frac{C+D}{n(1-\gamma^2)}}$$

(Hoeffding D) A more general measure of dependence was developed [102] (where $D_1 = \Sigma_i(Q_i-1)(Q_i-2)$, $D_2 = \Sigma_i(R_i-1)(R_i-2)(S_i-1)(S_i-2)$, $D_3 = \Sigma_i(R_i-2)(S_i-2)(Q_i-1)$, $R_i = \text{Rank}(X_i)$, $S_i = \text{Rank}(Y_i)$, $Q_i = 1 + |\{(x_j, y_j) \text{ s.th. } x_j<x_i, y_j<y_i\}|$):

$$D = 30\frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

It was slightly changed in definition [103] to point out the calculation behind (see *Partition on 0^2 and Hoeffding D*):

Partition on 0^2 and Hoeffding D

$X_i \backslash Y_i$	$y_j \le Y_i$	$y_j > Y_i$
$x_j \le X_i$	a_i	b_i
$x_j > X_i$	c_i	d_i
$B_n = \Sigma_i(a_i d_i - b_i c_i)^2/n^4$		

(other statistics) There are many statistics developed to measure in certain manners the agreement between the ranks in data series. A brief list of other methods is given in the next table (see *Other ordinal correlation measures*).

Other ordinal correlation measures

Expression	Ref
$d = (C-D)/(C+D+T_Y)$	Sommer's d [104]
$\tau_K = 4 \cdot cnn \cdot (n(n-1))^{-1} - 1$	Kendall's τ_K [105]
$\rho_F = 1 - \frac{4}{n^2}\sum_{1\le i,j\le n}\lvert \text{rank}(x_i) - \text{rank}(y_i)\rvert$	Spearman's footrule ρ_F [105]

(ordered contingencies)

Let us take a general two-way table representing the contingency of two variables $(X_i)_{1\le i\le R}$ and $(Y_j)_{1\le j\le C}$ (see *R\C ordered contingency*).

R\C ordered contingency

R\C	Y_1	…	Y_j	…	Y_C
X_1	$n_{1,1}$	…	$n_{1,j}$	…	$n_{1,C}$
…	…	…	…	…	…
X_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,C}$
…	…	…	…	…	…
X_R	$n_{R,1}$	…	$n_{R,j}$	…	$n_{R,C}$
$X_1 < \ldots < X_i < \ldots < X_R$ $Y_1 < \ldots < Y_j < \ldots < Y_C$					

In order to give computing formula for rank correlation measures (see *Computing rank correlation measures for multinomial contingencies*) a series of notations should be introduced (see *Notations for computing rank correlation measures for ordered contingencies*).

Notations for computing rank correlation measures for ordered contingencies

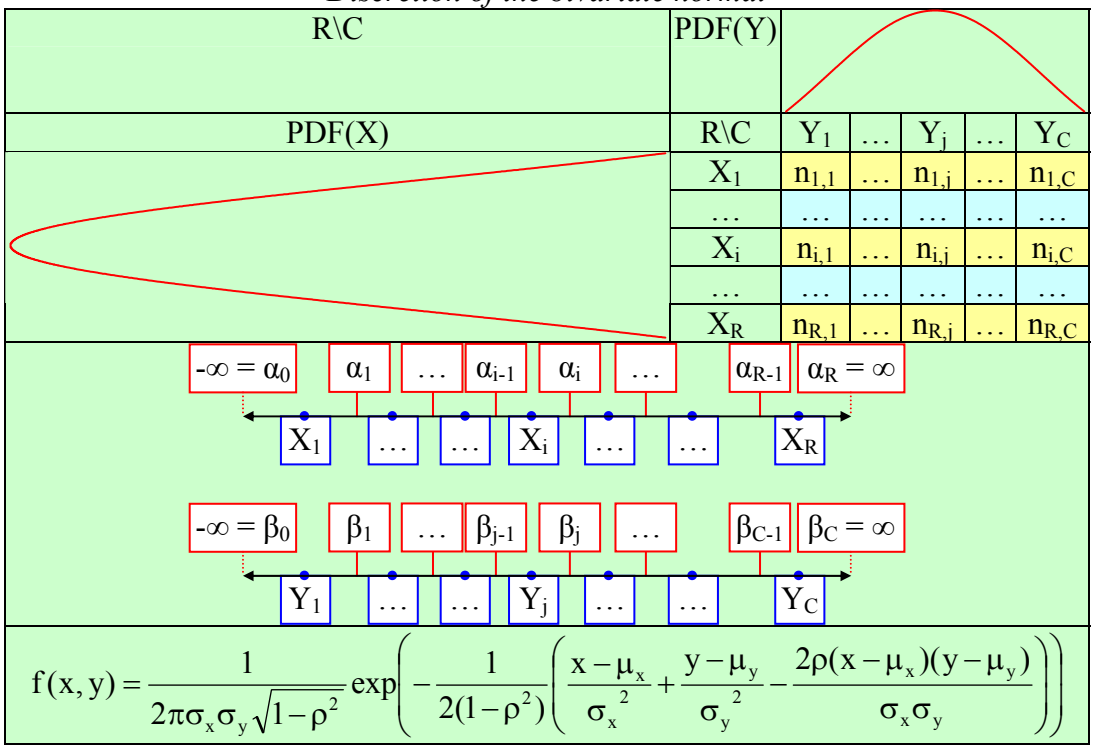
Label	Formula	Meaning
$n_{i,\cdot}$; $n_{\cdot,j}$; n	$\sum_j n_{i,j}$; $\sum_i n_{i,j}$; $\sum_i\sum_j n_{i,j}$	row, column, and overall totals
R_i; C_j	i or $R1_i$; j or $C1_j$	score for row *i* and for column *j*
$\overline{R}$; $\overline{C}$	$\sum_i n_{i,\cdot}R_i \big/ n$; $\sum_j n_{\cdot,j}C_j \big/ n$	average row and column score
$R1_i$; $C1_j$	$\sum_{k<i} n_{k,\cdot} + \frac{n_{i,\cdot}+1}{2}$; $\sum_{l<j} n_{\cdot,l} + \frac{n_{\cdot,j}+1}{2}$	rank score of row *i* and column *j*
$C_{i,j}$; $D_{i,j}$	$\sum_{k>i}\sum_{l>j} n_{k,l} + \sum_{k<i}\sum_{l<j} n_{k,l}$; $\sum_{k>i}\sum_{l<j} n_{k,l} + \sum_{k<i}\sum_{l>j} n_{k,l}$	concordances and discordances
C; D	$\sum_i\sum_j n_{i,j}C_{i,j}$; $\sum_i\sum_j n_{i,j}D_{i,j}$	twice number of concordances twice number of discordances

Computing rank correlation measures for ordered contingencies

Measure	Its expression
γ	$(C-D)/(C+D)$
Asymptotic variance	$\frac{16}{(C+D)^4}\sum_i\sum_j n_{i,j}(D\cdot C_{i,j}-C\cdot D_{i,j})^2$
Variance to be different from 0	$\frac{4}{(C+D)^2}\left(\sum_i\sum_j n_{i,j}(C_{i,j}-D_{i,j})^2-\frac{(C-D)^2}{n}\right)$
τ_b	$\frac{C-D}{w}$, $w=\sqrt{w_r w_c}$, $w_r=n^2-\Sigma_i n_{i,\cdot}{}^2$, $w_c=n^2-\Sigma_j n_{\cdot,j}{}^2$ $d_{i,j}=C_{i,j}-D_{i,j}$, $v_{i,j}=n_{i,\cdot}w_c-n_{\cdot,j}w_r$
Asymptotic variance	$\frac{1}{w^4}\left(\sum_i\sum_j n_{i,j}(2wd_{i,j}+t_b\cdot v_{i,j})^2-n^3 t_b{}^2(w_r+w_c)^2\right)$
Variance to be different from 0	$\frac{4}{w_r w_c}\left(\sum_i\sum_j n_{i,j}(C_{i,j}-D_{i,j})^2-\frac{(C-D)^2}{n}\right)$
$d_{C\|R}$	$\frac{C-D}{w_r}$, $w_r=n^2-\Sigma_i n_{i,\cdot}{}^2$, $d_{i,j}=C_{i,j}-D_{i,j}$
Asymptotic variance	$\frac{4}{w_r{}^4}\left(\sum_i\sum_j n_{i,j}(w_r d_{i,j}-(C-D)(n-n_{i,\cdot}))^2\right)$
Variance to be different from 0	$\frac{4}{w_r{}^2}\left(\sum_i\sum_j n_{i,j}(C_{i,j}-D_{i,j})^2-\frac{(C-D)^2}{n}\right)$
r	ss_{rc}/w, $w=\sqrt{ss_r ss_c}$ $ss_r=\sum_i\sum_j n_{i,j}(R_i-\overline{R})^2$, $ss_c=\sum_i\sum_j n_{i,j}(C_j-\overline{C})^2$ $ss_{rc}=\sum_i\sum_j n_{i,j}(R_i-\overline{R})(C_j-\overline{C})$, $b_{i,j}=(R_i-\overline{R})^2 ss_c+(C_j-\overline{C})^2 ss_r$
Asymptotic variance	$\frac{1}{w^4}\left(\sum_i\sum_j n_{i,j}\left(w(R_i-\overline{R})(C_j-\overline{C})-\frac{b_{i,j}ss_{rc}}{2w}\right)^2\right)$
Variance to be different from 0	$\frac{1}{w^2}\left(\sum_i\sum_j n_{i,j}(R_i-\overline{R})(C_j-\overline{C})-\frac{ss_{rc}{}^2}{n}\right)$
ρ_S	$\frac{v}{w}$, $v=\sum_i\sum_j n_{i,j}R0_i C0_j$, $w=\frac{1}{12}\sqrt{FG}$, $F=n^3-\Sigma_i n_{i,\cdot}{}^3$, $G=n^3-\Sigma_j n_{\cdot,j}{}^3$, $R0_i=R1_i-n/2$, $C0_j=C1_j-n/2$ $\overline{z}=\sum_i\sum_j n_{i,j}z_{i,j}$, $z_{i,j}=wv_{i,j}-vw_{i,j}$, $w_{i,j}=\frac{-n}{96w}(Fn_{\cdot,j}{}^2+Gn_{i,\cdot}{}^2)$ $v_{i,j}=n\left(R0_i C0_j+\frac{1}{2}\sum_l n_{i,l}C0_l\frac{1}{2}\sum_k n_{k,j}R0_k+\sum_l\sum_{k>i} n_{k,l}C0_l+\sum_k\sum_{l>j} n_{k,l}R0_k\right)$
Asymptotic variance	$\frac{1}{n^2 w^4}\sum_i\sum_j n_{i,j}(z_{i,j}-\overline{z})^2$
Variance to be different from 0	$\frac{1}{n^2 w^2}\sum_i\sum_j n_{i,j}(v_{i,j}-\overline{v})^2$, $\overline{v}=\sum_i\sum_j n_{i,j}v_{i,j}/n$

(polychoric correlation) Assuming for two variables that comes from normal distribution, and the contingency counts the number of observations for a given set of discrete values of X (X_1, …, X_R) and Y (Y_1, …, Y_C) then we may estimate the correlation between the population of X and Y by using the bivariate normal distribution (see *Discretion of the bivariate normal*).

Discretion of the bivariate normal



The values of μ_X and σ_X may come from MLE on $(X_i)_{1\le i\le R}$ values under assumption of normality and the values of μ_Y and σ_Y may come from MLE on $(Y_j)_{1\le j\le C}$ values under assumption of normality and are assumed known with known data series.

The joint probability of X_i and Y_j falling in $(\alpha_{i-1},\alpha_i]X(\beta_{j-1},\beta_j]$ is estimated by the observed count, from which the likelihood of all $(X_i)_{1\le i\le R}$ and $(Y_j)_{1\le j\le R}$ falling in the designed intervals $(\alpha_{i-1},\alpha_i]X(\beta_{j-1},\beta_j]$ can be derived. Maximum likelihood is obtained when derivative of log of likelihood is null (MLE method, [65]), giving us gives a system of equations from which is possible to estimate (numerically) of unknown "thresholds" $(\alpha_i)_{1\le i\le R-1}$ and $(\beta_j)_{1\le j\le C-1}$ as well as correlation coefficient ρ.

$$n_{i,j} \sim \int_{\alpha_{i-1}}^{\alpha_i}\int_{\beta_{j-1}}^{\beta_j} f(x,y)dxdy\,,\ P(\alpha;\beta;\rho) = \prod_{i=1}^{R}\prod_{j=1}^{C}\int_{\alpha_{i-1}}^{\alpha_i}\int_{\beta_{j-1}}^{\beta_j} f(x,y)dxdy$$

$$\ln(P(\alpha;\beta;\rho)) = \sum_{i=1}^{R}\sum_{j=1}^{C}\int_{\alpha_{i-1}}^{\alpha_i}\int_{\beta_{j-1}}^{\beta_j} f(x,y)dxdy\,,\ \frac{\partial\left(\ln(P(\alpha;\beta;\rho))\right)}{\partial\alpha} = \frac{\partial\left(\ln(P(\alpha;\beta;\rho))\right)}{\partial\beta} = \frac{\partial\left(\ln(P(\alpha;\beta;\rho))\right)}{\partial\rho} = 0$$

Linear regression analysis: linear models

The simplest association model is linear association. The model assumes that exist a relationship between two paired characteristics expressed by a straight line. We may start expressing this association by using the implicit equation of a straight line: ax + by + c = 0.

If a = 0 then the equation of the line reduces to by + c = 0. If further c ≠ 0 gives a relationship which defines the mean of the Y associated characteristic but no relationship with X. Similarly if b = 0 then the equation of the line reduces to ax + c = 0 and if further c ≠ 0 gives a relationship which defines the mean of the X associated characteristic but no relationship with Y. The remained case, if c = 0 defines a degenerated linear model in which is no intercept between the characteristics X and Y.

Which expression of the linear equation we should use is a matter of experimental error treatment. Going further, if a linear model defines the relationship between the X and Y characteristics, then if we take samples $(x_i, y_i)_{1\le i\le n}$ of these two (X and Y) characteristics then we should still see the relationship in terms of a experimental error.

We may make errors measuring both characteristics, and thus if we express as $(\hat{x}_i, \hat{y}_i)$ the expected (or true) values each corresponding to a pair of measurements (x_i, y_i) then:

÷ The linear model is given by: $a\hat{x}_i + b\hat{y}_i + c = 0$, i = 1..n;

÷ The experimental errors are given by: $\varepsilon_i = x_i - \hat{x}_i$ and $\eta_i = y_i - \hat{y}_i$;

In order to be able to find the linear model (e.g. it's coefficients a, b and c) we must make further assumptions on experimental errors. Thus, at least we should know the expected value of the experimental error. If $\hat{x}_i$ is the expected value of x_i then the expected value of ε_i, namely $\hat{\varepsilon}_i$ is 0: $\hat{\varepsilon}_i = 0$, but this fact happens only if we not made systematically errors on measurement of the characteristic X. Similarly, if $\hat{y}_i$ is the expected value of y_i then the expected value of η_i, namely $\hat{\eta}_i$ is 0: $\hat{\eta}_i = 0$, but this fact happens only if we not made systematically errors on measurement of the characteristic Y. More than that, under presence of systematically errors on measurement of X or Y is no way to obtain the relationship between X and Y excepting the situation when we know the (expected) value of the systematic error (but then we can systematically extract it from the values to obtain new series of data which do not posses systematically errors). Thus, first assumptions were made, namely:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)}{n} = 0 \text{ (no systematically error on X);} \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} = 0 \text{ (idem for Y)}$$

Even more strong assumptions are generally accepted and are required in order to solve (or obtain) the relationship, namely:

$$\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)}{n} \cong 0 \text{ and } \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} \cong 0$$

Please note that these two relationships above do not imply that the absolute or other sort of modulus based differences are null to. In general:

$$\frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|^{\alpha}}{n} \ge 0 \text{ and } \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|^{\beta}}{n} \ge 0$$

We should know more about the distribution of the error. A common assumption is to expect that an error ε_i (or η_i) to occur in equal probability as an error $-\varepsilon_i$ (or $-\eta_i$), and then the distribution of the experimental error is symmetrical.

A choice is possible here coming from the experiment: to give different weight to the errors (and then we have a weighted regression). Usually the weight are function of the observable and/or expectance ($v_i = f(x_i, \hat{x}_i)$, $w_i = g(y_i, \hat{y}_i)$). The reason of giving weight to the

errors is to normalize (e.g. the distribution of the errors to become normal or at least to have a known distribution).

Here comes another assumption regarding the experimental error (ε_i and η_i) that we must take in order to obtain estimations of the population parameters: that follows a known distribution.

We may make inferences about the distribution (see for example *Distribution of errors when binary responses are recorded*).

Distribution of errors when binary responses are recorded

x_i	$p_i=p(x_i)$	$\hat{x}_i$	$q_i=p(\hat{x}_i)$	$x_i-\hat{x}_i$	$\lvert x_i-\hat{x}_i\rvert^{\alpha}$ ($\alpha>0$)	p_iq_i
0	p	0	q	0	0	pq
0	p	1	(1-q)	-1	1	p(1-q)
1	(1-p)	0	q	1	1	(1-p)q
1	(1-p)	1	(1-q)	0	0	(1-p)(1-q)

$x_i-\hat{x}_i$	$p(x_i-\hat{x}_i)$	$n=n_{-1}+n_0+n_1$ (observation errors)
-1	$p_{-1}=p(1-q)$	$P_3(n_{-1},n_0,n_1;n;\{p(1-q),(1-p)q,2pq\})=$ $\frac{n!(p(1-q))^{n_{-1}}((1-p)q)^{n_1}(1-p-q+2pq)^{n_0}}{n_{-1}!\cdot n_1!\cdot n_0!}=$ $\frac{n!\cdot 2^{n_0}\cdot p^{n-n_1}q^{n-n_{-1}}(1-p)^{n_1}(1-q)^{n_{-1}}(1-p-q+2pq)^{n_0}}{n_{-1}!\cdot n_0!\cdot n_1!}$
0	$p_0=1-p-q+2pq$	
1	$p_1=q(1-p)$	
$\lvert x_i-\hat{x}_i\rvert^{\alpha}$	$p(\lvert x_i-\hat{x}_i\rvert^{\alpha})$	$n=n_0+n_1$ (observation errors)
0	$p_0=1-p-q+2pq$	$P_2(n_0,n_1;n;\{1-p-q-2pq,p+q-2pq\})=$ $\frac{n!(1-p-q+2pq)^{n_0}(p+q-2pq)^{n_1}}{n_0!\cdot n_1!}=$ $\text{Binomial}(1-p-q+2pq;n)$
1	$p_1=p+q-2pq$	

The common continuous symmetrical distribution assumed to be followed by the experimental error is the Gauss-Laplace distribution [106]:

$$GL(x;\mu,\sigma,p)=\frac{p}{2\sigma}\frac{\Gamma^{1/2}(3/p)}{\Gamma^{3/2}(1/p)}\exp\left(-\left|\frac{x-\mu}{\sigma}\right|^{p}\Big/\left(\frac{\Gamma(1/p)}{\Gamma(3/p)}\right)^{p/2}\right)$$

When p = 1 the Laplace distribution occurs [107] (often referred in physical sciences) and when p = 2 the Gauss distribution occurs [108] (often referred in biological sciences).

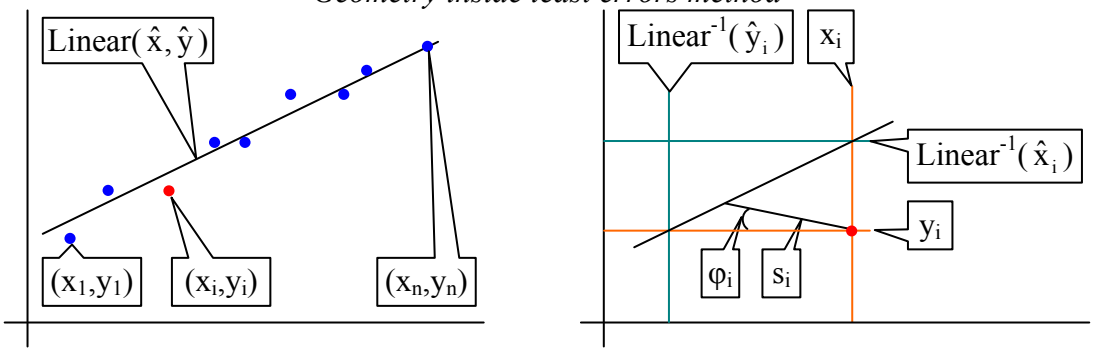
The relationship of Gauss-Laplace distribution with the distribution of the experimental error is immediate (see *Gauss-Laplace and least errors method*).

Gauss-Laplace and least errors method

Errors distribution assumption	Regression model	Least errors method
$\left(\lvert x_i-\hat{x}_i\rvert^{\alpha}\right)_{1\le i\le n}\sim GL(\varepsilon;0,\sigma_X,\alpha)$	$x\sim\hat{x}=b\cdot y+c$	$\sum_{i=1}^{n}\varepsilon_i^{\alpha}=\sum_{i=1}^{n}\lvert x_i-\hat{x}_i\rvert^{\alpha}=\min.$
$\left(\lvert y_i-\hat{y}_i\rvert^{\beta}\right)_{1\le i\le n}\sim GL(\eta;0,\sigma_Y,\beta)$	$y\sim\hat{y}=a\cdot x+c$	$\sum_{i=1}^{n}\eta_i^{\beta}=\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert^{\beta}=\min.$
$\left(\left\lvert\frac{x_i-\hat{x}_i}{\hat{x}_i}\right\rvert^{\gamma}\right)_{1\le i\le n}\sim GL(\varepsilon;0,\sigma_X,\gamma)$ $\left(\left\lvert\frac{y_i-\hat{y}_i}{\hat{y}_i}\right\rvert^{\gamma}\right)_{1\le i\le n}\sim GL(\eta;0,\sqrt{1-\sigma_X^{\,2}},\gamma)$	only when c = 1: $1\sim\hat{1}=a\cdot x+b\cdot y$	$\sum_{i=1}^{n}0_i^{\gamma}=\sum_{i=1}^{n}\lvert ax_i+by_i-1\rvert^{\gamma}=\min.$

The geometrical interpretation of the least errors method is depicted below (see *Geometry inside least errors method*). Subject to minimization is $S(\gamma,\varphi) = \Sigma_i s_i^{\gamma}$. Different values of φ_i and γ give different assumptions on errors, different linear models and different methods of calculation [109] (see *Calculations in linear regression - least errors method*).

Geometry inside least errors method



Calculations in linear regression - least errors method

φ_i	γ	Regression model	$S \to \min.$	Method	Further assumption
0	2	$x \sim \hat{x} = b \cdot y + c$	$\Sigma_i(x_i - by_i - c)^2$	analytical from $\frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$	measurement of $(y_i)_{1 \le i \le n}$ without errors
	1		$\Sigma_i\lvert x_i - by_i - c\rvert$	numerical from $\frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$	
$\pi/2$	2	$y \sim \hat{y} = a \cdot x + c$	$\Sigma_i(y_i - ax_i - c)^2$	analytical from $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial c} = 0$	measurement of $(x_i)_{1 \le i \le n}$ without errors
	1		$\Sigma_i\lvert y_i - ax_i - c\rvert$	numerical from $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial c} = 0$	
s.th. $s_i \perp$ on Linear($\hat{x}, \hat{y}$)	2	$x \sim \hat{x} = b \cdot y + c$	see [109]	analytical from $\frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$	$\varepsilon_i=\varepsilon(x_i)$; $\eta_i=\eta(y_i)$ $\varepsilon_i/\eta_i=\Sigma x_i^2/\Sigma y_i^2$
	1			numerical from $\frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$	$\varepsilon_i=\varepsilon(x_i)$; $\eta_i=\eta(y_i)$ $\varepsilon_i/\eta_i=\Sigma\lvert x_i\rvert/\Sigma\lvert y_i\rvert$
	2	$y \sim \hat{y} = a \cdot x + c$		analytical from $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial c} = 0$	$\varepsilon_i=\varepsilon(x_i)$; $\eta_i=\eta(y_i)$ $\varepsilon_i/\eta_i=\Sigma x_i^2/\Sigma y_i^2$
	1			numerical from $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial c} = 0$	$\varepsilon_i=\varepsilon(x_i)$; $\eta_i=\eta(y_i)$ $\varepsilon_i/\eta_i=\Sigma\lvert x_i\rvert/\Sigma\lvert y_i\rvert$
s.th. s_i is median	2	$x \sim \hat{x} = b \cdot y + c$	…	…	…
	1		…	…	…
	2	$y \sim \hat{y} = a \cdot x + c$	…	…	…
	1		…	…	…
	2	$1 \sim a \cdot x + b \cdot y$	…	…	…
	1		…	…	…

The multiple regression models are generalizations of simple linear regression. We may conduct the same rationalizations (as in Gauss-Laplace and least errors method or in Calculations

in linear regression - least errors method) in order to derive representatives of the family of multiple regression models. We will take only the simplest and in the same time the most common case:

$$y \sim \hat{y} = a_0 \cdot 1 + a_1 \cdot x_1 + \ldots + a_m \cdot x_m \text{ for } (y_i, x_{1,i}, \ldots, x_{m,i})_{1 \le i \le n} \text{ and at least } m<n$$

Deriving of $\Sigma_i(a_0+a_1x_{1,i}+\ldots+a_mx_{m,i}-y_i)^2=\min.$ conduct to the following system of equations:

$$S_A = \begin{matrix} 0 & 1 & \ldots & m \end{matrix} \begin{pmatrix} \Sigma_i 1 \cdot 1 & \Sigma_i x_{1,i} \cdot 1 & \ldots & \Sigma_i x_{m,i} \cdot 1 \\ \Sigma_i 1 \cdot x_{1,i} & \Sigma_i x_{1,i} \cdot x_{1,i} & \ldots & \Sigma_i x_{n,i} \cdot x_{1,i} \\ \ldots & \ldots & \ldots & \ldots \\ \Sigma_i 1 \cdot x_{n,i} & \Sigma_i x_{1,i} \cdot x_{n,i} & \ldots & \Sigma_i x_{n,i} \cdot x_{m,i} \end{pmatrix} \begin{matrix} 0 \\ 1 \\ \ldots \\ m \end{matrix}, \quad S_B = \begin{matrix} 0 \\ 1 \\ \ldots \\ m \end{matrix} \begin{pmatrix} \Sigma_i y_i \cdot 1 \\ \Sigma_i y_i \cdot x_{1,i} \\ \ldots \\ \Sigma_i y_i \cdot x_{m,i} \end{pmatrix}, \quad S_A \times A = S_B$$

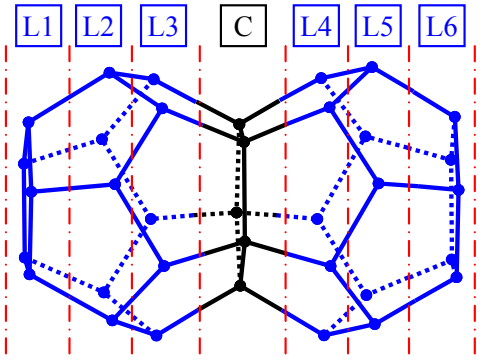
from which the solution is immediate: $S_A^{-1} \times S_A \times A = S_A^{-1} \times S_B \rightarrow A = S_A^{-1} \times S_B$.

The Gauss-Jordan numeric method [110] exploits the feature of elementary row operations (multiplication/division and addition/subtraction keep the matrix equality unchanged) to obtain simultaneous with the coefficients their variance (see *Gauss-Jordan procedure for linear multiple regression*).

Aplicații

Problemă: Nivele de reprezentare

Se doreşte să se realizeze un experiment în care să fie colectate (în spațiul informațional) observații cu privire la structura de mai jos, în care pe fiecare din cele 6 nivele pot fi prezenți atomi de carbon, azot şi bor.



Câte informații distincte se vor regăsi, maxim, în spațiul informațional, ştiind că structura este rigidă şi fixată într-un ansamblu mai mare (cazul a) şi respectiv mobilă (cazul b).

Soluție

Straturi: 6 (L1, L2, L3, L4, L5, L6); Nivele: 3 (B, N, C); Număr total de combinații: 729; Structuri distincte: 378

Studiu de caz: Proprietăți din Spartan '10

MolVol: Molecular volume $Å^3$; SurfA: Surface area $Å^2$; Ovality: Ovality dimensionless 1.234; HOMO: Highest Occupied Molecular Orbital Energy eV; LUMO: Lowest Unoccupied Molecular Orbital Energy eV; Estimated polarizability: $*10^{-30}$ m^3; Electronegativity= -HOMO+LUMO/2; Hardness= -HOMO-LUMO/2

Property	Distinct	Grp.	Zero		Property	Distinct	Grp.	Zero
DipoleT_0	376	352			Lumo-Homo_0	353	375	
DipoleT_1	377	351			Lumo-Homo_1	331	397	
DipoleT_2	377	351			Lumo-Homo_2	372	356	
EnergyHF_0	131	510	87		MolVol_0	175	512	41
EnergyHF_1	139	507	82		MolVol_1	167	530	31
EnergyHF_2	346	382			MolVol_2	372	356	
HOMO_0	337	389	2		Ovality_0	46	516	166
HOMO_1	320	408			Ovality_1	51	590	87
HOMO_2	367	361			Ovality_2	56	645	27
LUMO_0	318	410			Polariz_0	349	379	
LUMO_1	300	428			Polariz_1	316	412	
LUMO_2	367	361			Polariz_2	371	357	
Lumo+Homo_0	355	373			SurfA_0	221	477	30
Lumo+Homo_1	324	404			SurfA_1	202	520	6
Lumo+Homo_2	356	370	2		SurfA_2	368	360	

Problemă: Precizie şi exactitate

Să se analizeze sub aspectul preciziei următoarele măsurători ale vitezei vântului în raport cu afirmația că în acea zi nu s-au înregistrat deplasări de aer.

Moment	Viteza	Direcția	Durata
9^{30}	$2.8ms^{-1}$	NV	1 min.
10^{30}	$1.0ms^{-1}$	E	2 min.
11^{30}	$1.4ms^{-1}$	NE	1 min.
12^{30}	$3ms^{-1}$	S	1 min.

Ce observație lipsește pentru ca să se susțină afirmația?

Soluție

Se alcătuieşte un tabel de forma celui de mai jos:

v	t	dir	dx	dy	sdx	sdy
2.8	60	NV	-118.8	118.8	-118.8	118.8
1	120	E	120.0	0.0	1.2	118.8
1.4	60	NE	59.4	59.4	60.6	178.2
3	60	S	0.0	-180.0	60.6	-1.8

Problemă: Separare maximă

În cromatografie se folosesc o serie de măsuri pentru a caracteriza separarea compuşilor dintr-un amestec:

÷ $RF(i,e) = l(i,e)/l(e)$, în care *i* este un compus separat, *e* faza mobilă, *l(i,e)* distanța de migrare a lui *i* în *e*, şi *l(e)* distanța de migrare a lui *e* şi ce conduce la *mulțimea factorilor de retardare* (RF)

÷ $RFO(i,e) = l(\pi(i),e)/l(e)$, unde *π(i)* este o permutare ce transformă lista într-o listă ordonată (crescător sau descrescător) şi ce conduce la *şirul ordonat al factorilor de retardare* (RFO)

÷ $nc(e) = \text{Count}(\{i \mid l(\pi(i+1),e) - l(\pi(i),e) > (w(\pi(i+1),e) + w(\pi(i),e))/8\})$ ce conduce la *numărul de componenți observați la o distanță (unul de altul) de cel puțin 1σ* (nc)

÷ $mnc = \max\{nc(e) \mid e \in E\}$, în care E mulțimea tuturor fazelor mobile posibile şi ce conduce la *numărul de componenți existenți în amestec* (mnc)

÷ $RSM(i,j,e) = 2 \cdot (l(i,e) - l(j,e))/(w(i,e) + w(j,e))$, unde *w(i,e)* este lărgimea spotului *i* iar *w(j,e)* este lărgimea spotului *j*, şi ce conduce la *matricea rezoluțiilor* (RSM)

÷ $RSO(i,e) = 2 \cdot (l(\pi(i+1),e) - l(\pi(i),e))/(w(\pi(i+1),e) + w(\pi(i),e))$, unde din nou *π* este permutarea ce ordonează (crescător) şirul factorilor de retardare şi ce conduce la *şirul rezoluțiilor spoturilor adiacente* (RSO)

÷ $RSS(e) = \sum_{i=1}^{nc(e)} RSO(i,e)$ ce conduce la *suma rezoluțiilor* (RSS)

÷ $QN_{eff}(e) = \sqrt{N_{eff}(e)} = 4 \cdot l(e) \cdot nc(e) \Big/ \sum_{i=1}^{nc(e)} w(e,i)$ ce conduce la *radical din numărul efectiv de talere* (QN_{eff}); ajustând condițiile experimentale date idealul de separare se obține când $4 \cdot RSS(e) \rightarrow QN_{eff}(e)$

÷ $RSP(e) = 25 \cdot RSS(e)/QN_{eff}(e)$ ce dă *rezoluția împărțită la numărul de talere* (RSP); idealul de separare se obține când $RSP(e) \rightarrow 100$

÷ $RSA(e) = RSS(e)/nc(e)$ ce dă *rezoluția medie a separării* (RSA)

÷ $RRP(e) = \prod_{i=1}^{nc(e)} RSO(i,e) \Big/ \sum_{i=1}^{nc(e)} RSO(i,e)$ care este *produsul rezoluțiilor relative* (RRP)

÷ $RSH(e,p) = \left(\sum_{i=1}^{nc(e)} (RSO(i,e))^p \Big/ nc(e) \right)^{1/p}$ unde p este o valoare reală arbitrară şi ce conduce la *media Hölder a rezoluțiilor* (RSH); când p=2 RSH(e,2) este un mai bun descriptor al separării decât RSA

÷ $RFD(e) = \sqrt{\sum_{i=1}^{nc(e)} (RFO(i+1,e) - RFO(i,e) - 1/mnc)^2} \Big/ \sqrt{nc(e) \cdot (nc(e)+1)}$, unde *1/mnc* este diferența teoretică între 2 factori de retardare şi ce conduce la un indice de separare medie exprimat ca deviație între ideal (1/mnc) şi observat (RFO(i+1,e)-RFO(e)) - *deviația factorilor de retardare* (RFD)

÷ $IEne(e) = mnc^2 - \sum_{i=1}^{mnc} n_i^{\,2}$, unde n_i este numărul de spoturi în al *i*-lea interval echidistant şi ce

conduce la *energia informațională asociată separării* (IEne)

÷ $IEnt(e) = \sum_{i=1}^{mnc} n_i \cdot \log(n_i)$, unde n_i este numărul de spoturi în al *i*-lea interval echidistant şi ce conduce la *entropia informațională asociată separării* (IEnt)

÷ $QF(e) = \min_{1 \le i,j \le nc(e)} RSM(i,j,e) = \min_{1 \le i \le nc(e)} RSO(i,e)$ care conduce la *factorul de calitate a separării* (QF) dat de cea mai defavorabilă separare

Să se calculeze şi apoi să se analizeze proprietățile (operațiile matematice permise) de aceşti parametrii.

Soluție

Imaginând procesul de separare drept un proces în care apariția migrării spoturilor este o problemă de durată, din punct de vedere experimental este de interes ce se întâmplă cu valorile acestor parametrii când se detectează apariția unui nou spot (eventual la începutul sau la sfârşitul şirului de spoturi deja existent).

Fie *e* un eluent (acelaşi). Vom simplifica notațiile făcând referire mereu la acelaşi eluent.

RF. Fie {RF(i)| 1≤i≤nc} mulțimea factorilor de retardare. Apariția unui nou factor face ca la mulțimea factorilor de retardare să se adauge un element. Din acest punct de vedere, mulțimea factorilor de retardare poate fi văzută ca o mulțime de caracteristici. Asocierea între compusul chimic şi factorul de retardare al acestuia e unică (pe calea directă, adică un compus are un singur factor de retardare în condițiile experimentale date; reciproca însă nu este adevărată, putând exista 2 compuşi cu acelaşi factor de retardare) şi poate indica ulterior absența compusului dintr-un amestec prin absența valorii factorului de retardare din mulțimea valorilor măsurate pentru amestec.

RFO. Şirul ordonat al factorilor de retardare nu este o mulțime standard, fiind echipat şi cu o relație de ordine. Din acest punct de vedere, şirul ordonat al factorilor de retardare necesită o operație simplă de adăugare la început sau la sfârşit doar dacă valoarea adăugată este fie mai mică, fie mai mare decât toate valorile deja existente în şir; în caz contrar, adăugarea necesită şi identificarea poziției în care trebuie făcută inserarea.

nc. Numărul de componenți observați îşi modifică valoarea la apariția unui nou spot numai dacă este îndeplinită condiția ca distanța dintre spoturile adiacente să fie cel puțin 1/8 din lărgimea acestora. Din acest punct de vedere, adăugarea unui nou compus în amestec poate avea ca efect unul din următoarele:

÷ nc → nc-1 când noul spot apare între 2 spoturi deja existente, dar apariția acestuia face ca cele 2 spoturi existente inițial împreună cu cel de-al treilea nou spot să nu mai poată fi distinse unul de celălalt;

÷ nc → nc când noul spot apare în vecinătatea altuia dar suficient de departe de oricare alt spot

÷ nc → nc+1 când noul spot apare suficient de departe de oricare alt spot

RSM. Matricea rezoluțiilor suferă o operațiune de adăugare a unei linii şi a unei coloane pentru fiecare nou compus adăugat în amestec. Liniile şi coloanele existente inițial în această matrice rămân cu valori neschimbate.

RSO. Şirul rezoluțiilor ordonate nu se poate obține din ordonarea matricei RSM (de exemplu ordonând o linie prin permutare de coloane). Matricea rezoluțiilor ordonate se obține din diferențele între valorile din şirul ordonat al factorilor de retardare (RFO →RSO). Din acest punct de vedere, rezultatul obținut prin adăugarea unui nou compus în amestec poate produce apariția sau dispariția unui element în şirul rezoluțiilor ordonate sau să se păstreze numărul acestora (vezi nc) necesitând totodată şi schimbarea valorilor adiacente valorii inserate (dacă numărul elementelor rămâne acelaşi sau creşte cu o unitate).

Problemă: Metoda adaosului standard

Se consideră o probă sub formă de soluție în care compuşii dizolvați sunt cunoscuți dar sunt în cantități necunoscute. De asemenea sunt disponibili aceşti compuşi preparați separat sub formă de soluții apoase. Care este modalitatea de determinare a cantităților necunoscute, ştiind că identificarea se realizează cu ajutorul unui detector care produce un semnal proporțional cu concentrația (fie aceasta molară) de compus identificat.

Soluție

Fie x şi y cantitățile în grame de compuşi cunoscuți C_1 şi C_2 dizolvați în 100 ml H_2O (în proba

necunoscută) şi fie a şi b cantităţile în grame de compuşi cunoscuţi dizolvaţi în 100 ml H_2O (în etaloanele preparate din aceşti compuşi). Fie masele molare M_1 (în g, pentru C_1) şi M_2 (în g, pentru C_2). În acest caz avem la dispoziţie 3 soluţii: S_U (soluţia necunoscută), S_1 (din C_1) şi S_2 (din C_2):

Soluţie	S_U	S_1	S_2
Volum	100 ml	100 ml	100 ml
Substanţă dizolvată	x gC_1, y gC_2	a gC_1	b gC_2

Imaginăm o serie de 3 experimente cu aceşti compuşi în care folosim câte 25 ml din soluţia cu cantităţi dizolvate necunoscute (S_U) şi câte un volum (deocamdată neprecizat) din soluţiile etalon. Un calcul simplu dă compoziţia amestecurilor:

Experiment	S_U	S_1	S_2	$V(H_2O)$	$m(C_1)$	$m(C_2)$
1	25 ml	V_{11} ml	V_{12} ml	$25+V_{11}+V_{12}$	$25\cdot x/100+V_{11}\cdot a/100$	$25\cdot y/100+V_{12}\cdot b/100$
2	25 ml	V_{21} ml	V_{22} ml	$25+V_{21}+V_{22}$	$25\cdot x/100+V_{21}\cdot a/100$	$25\cdot y/100+V_{22}\cdot b/100$
3	25 ml	V_{31} ml	V_{32} ml	$25+V_{31}+V_{32}$	$25\cdot x/100+V_{31}\cdot a/100$	$25\cdot y/100+V_{32}\cdot b/100$

Detectorul va înregistra semnale. În acest moment avem 2 soluţii de implementare: când detecţia se realizează după separare, caz în care pentru fiecare compus identificat avem câte un semnal în fiecare experiment (fie R_{11} şi R_{12} semnalele din primul experiment şi asemeni R_{21}, R_{22}, R_{31} şi R_{32}) sau detecţia se realizează direct asupra amestecului, caz în care avem câte un singur semnal din fiecare experiment (fie Z_1 semnalul înregistrat în primul experiment corespunzător detecţiei, Z_2 semnalul înregistrat în al doilea experiment şi Z_3 semnalul înregistrat în al treilea experiment) proporţionale cu concentraţiile molare, deci este necesar să evaluăm aceste concentraţii:

Experiment	$n(C_1)/V(H_2O)$	$n(C_2)/V(H_2O)$
1	$(25\cdot x/100+V_{11}\cdot a/100)/M_1/(25+V_{11}+V_{12})$	$(25\cdot y/100+V_{12}\cdot b/100)/M_2/(25+V_{11}+V_{12})$
2	$(25\cdot x/100+V_{21}\cdot a/100)/M_1/(25+V_{21}+V_{22})$	$(25\cdot y/100+V_{22}\cdot b/100)/M_2/(25+V_{21}+V_{22})$
3	$(25\cdot x/100+V_{31}\cdot a/100)/M_1/(25+V_{31}+V_{32})$	$(25\cdot y/100+V_{32}\cdot b/100)/M_2/(25+V_{31}+V_{32})$
Semnal	Separat pentru C_1	Separat pentru C_2
1	R_{11}	R_{12}
2	R_{21}	R_{22}
3	R_{31}	R_{32}
Semnal	Împreună pentru C_1 şi C_2 în amestec	
1	Z_1	
2	Z_2	
3	Z_3	

Se exprimă raportul între concentraţie şi semnal care pentru un anumit compus este acelaşi de la un experiment la altul. Rezolvăm în continuare pe cele două cazuri:

Cazul semnalului separat

÷ Pentru C_1: $\dfrac{x/4+a\cdot V_{11}/100}{M_1\cdot(25+V_{11}+V_{12})\cdot R_{11}}=\dfrac{x/4+a\cdot V_{21}/100}{M_1\cdot(25+V_{21}+V_{22})\cdot R_{21}}=\dfrac{x/4+a\cdot V_{31}/100}{M_1\cdot(25+V_{31}+V_{32})\cdot R_{31}}$

÷ Pentru C_2: $\dfrac{y/4+b\cdot V_{12}/100}{M_2\cdot(25+V_{11}+V_{12})\cdot R_{12}}=\dfrac{y/4+b\cdot V_{22}/100}{M_2\cdot(25+V_{21}+V_{22})\cdot R_{22}}=\dfrac{y/4+b\cdot V_{32}/100}{M_2\cdot(25+V_{31}+V_{32})\cdot R_{32}}$

Simplificând:

÷ Pentru C_1: $\dfrac{x\cdot 25+a\cdot V_{11}}{(25+V_{11}+V_{12})\cdot R_{11}}=\dfrac{x\cdot 25+a\cdot V_{21}}{(25+V_{21}+V_{22})\cdot R_{21}}=\dfrac{x\cdot 25+a\cdot V_{31}}{(25+V_{31}+V_{32})\cdot R_{31}}$

÷ Pentru C_2: $\dfrac{y\cdot 25+b\cdot V_{12}}{(25+V_{11}+V_{12})\cdot R_{12}}=\dfrac{y\cdot 25+b\cdot V_{22}}{(25+V_{21}+V_{22})\cdot R_{22}}=\dfrac{y\cdot 25+b\cdot V_{32}}{(25+V_{31}+V_{32})\cdot R_{32}}$

Pentru fiecare dintre cei doi compuşi, avem exact 2 soluţii (nu neapărat identice) corespunzătoare celor 2 egalităţi. În fapt, problema poate fi rezolvată şi mai corect minimizând eroarea de observare (temă de casă!).

În continuare, fie aceste două egalităţi următoarele:

$$x\cdot\frac{25}{(25+V_{11}+V_{12})\cdot R_{11}}+\frac{a\cdot V_{11}}{(25+V_{11}+V_{12})\cdot R_{11}}=x\cdot\frac{25}{(25+V_{21}+V_{22})\cdot R_{21}}+\frac{a\cdot V_{21}}{(25+V_{21}+V_{22})\cdot R_{21}}$$

$$x\cdot\frac{25}{(25+V_{11}+V_{12})\cdot R_{11}}+\frac{a\cdot V_{11}}{(25+V_{11}+V_{12})\cdot R_{11}}=x\cdot\frac{25}{(25+V_{31}+V_{32})\cdot R_{31}}+\frac{a\cdot V_{31}}{(25+V_{31}+V_{32})\cdot R_{31}}$$

De unde:

$$x=a\cdot\frac{\frac{V_{21}/R_{21}}{25+V_{21}+V_{22}}-\frac{V_{11}/R_{11}}{25+V_{11}+V_{12}}}{\frac{25/R_{11}}{25+V_{11}+V_{12}}-\frac{25/R_{21}}{25+V_{21}+V_{22}}}\ \text{şi/sau}\ x=a\cdot\frac{\frac{V_{31}/R_{31}}{25+V_{31}+V_{32}}-\frac{V_{11}/R_{11}}{25+V_{11}+V_{12}}}{\frac{25/R_{11}}{25+V_{11}+V_{12}}-\frac{25/R_{31}}{25+V_{31}+V_{32}}}$$

În mod similar rezultă şi expresiile pentru y.

Cazul amestecului

În cazul măsurării directe pe amestec, se suprapune semnalul de la primul compus cu semnalul de la al doilea compus, însă nu în mod necesar cu aceeaşi intensitate:

÷ Experiment 1: $\frac{x/4+a\cdot V_{11}/100}{M_1\cdot(25+V_{11}+V_{12})}\cdot\alpha+\frac{y/4+b\cdot V_{12}/100}{M_2\cdot(25+V_{11}+V_{12})}\cdot\beta=Z_1$

÷ Experiment 2: $\frac{x/4+a\cdot V_{21}/100}{M_1\cdot(25+V_{21}+V_{22})}\cdot\alpha+\frac{y/4+b\cdot V_{22}/100}{M_2\cdot(25+V_{21}+V_{22})}\cdot\beta=Z_2$

÷ Experiment 3: $\frac{x/4+a\cdot V_{31}/100}{M_1\cdot(25+V_{31}+V_{32})}\cdot\alpha+\frac{y/4+b\cdot V_{32}/100}{M_2\cdot(25+V_{31}+V_{32})}\cdot\beta=Z_3$

Ecuaţiile de mai sus formează un sistem de 3 ecuaţii cu 4 necunoscute (x, y, α şi β). Este evident că pentru a rezolva acest sistem este nevoie de încă o ecuaţie, deci de încă un experiment:

Experiment 4: $\frac{x/4+a\cdot V_{41}/100}{M_1\cdot(25+V_{41}+V_{42})}\cdot\alpha+\frac{y/4+b\cdot V_{42}/100}{M_2\cdot(25+V_{41}+V_{42})}\cdot\beta=Z_4$

Sistemul de ecuaţii rezultat se poate scrie prin intermediul valorilor cunoscute astfel (unde c_{11}, ... c_{45} sunt valori cunoscute):

$$x\cdot\alpha\cdot c_{11}+\alpha\cdot c_{12}+y\cdot\beta\cdot c_{13}+\beta\cdot c_{14}=c_{15}$$
$$x\cdot\alpha\cdot c_{21}+\alpha\cdot c_{22}+y\cdot\beta\cdot c_{23}+\beta\cdot c_{24}=c_{25}$$
$$x\cdot\alpha\cdot c_{31}+\alpha\cdot c_{32}+y\cdot\beta\cdot c_{33}+\beta\cdot c_{34}=c_{35}$$
$$x\cdot\alpha\cdot c_{41}+\alpha\cdot c_{42}+y\cdot\beta\cdot c_{43}+\beta\cdot c_{44}=c_{45}$$

Chiar dacă pare complicat, în esenţă este un sistem de ecuaţii uşor de rezolvat. Făcând substituţiile $\gamma=x\cdot\alpha$ şi $\delta=y\cdot\beta$ sistemul devine un sistem liniar şi omogen în necunoscutele α, β, γ şi δ:

$$\gamma\cdot c_{11}+\alpha\cdot c_{12}+\delta\cdot c_{13}+\beta\cdot c_{14}=c_{15}$$
$$\gamma\cdot c_{21}+\alpha\cdot c_{22}+\delta\cdot c_{23}+\beta\cdot c_{24}=c_{25}$$
$$\gamma\cdot c_{31}+\alpha\cdot c_{32}+\delta\cdot c_{33}+\beta\cdot c_{34}=c_{35}$$
$$\gamma\cdot c_{41}+\alpha\cdot c_{42}+\delta\cdot c_{43}+\beta\cdot c_{44}=c_{45}$$

Odată obţinută soluţia acestui sistem (α, β, γ şi δ) valorile necunoscutelor x şi y sunt imediate: $x=\gamma/\alpha$ şi $y=\delta/\beta$.

Problemă: Calculul pH-ului în funcţie de volumul de soluţie adăugată

Să se exprime variaţia pH-ului în titrarea în soluţie apoasă (H_2O) a unui acid slab (HA) cu o bază slabă (BOH).

Soluţie

Modelul reacţiei de titrare porneşte de la scrierea ecuaţiei reacţiei chimice a titrării unui acid slab HA cu o bază slabă BOH, pentru care echilibrul este caracterizat de constanta de solubilitate a sării obţinute:

$$HA + BOH \rightleftharpoons AB + H_2O$$

Dacă reacţia se desfăşoară în apă, trebuie să considerăm influenţa asupra pH-ului din disocierea

moleculelor de apă:

$$HOH \rightleftharpoons HO^- + H^+$$

Procesele de disociere ale acidului, bazei şi sării în apă sunt exprimate de relaţiile:

$$HA \rightleftharpoons H^+ + A^-,\ BOH \rightleftharpoons B^+ + HO^-,\ AB \rightleftharpoons A^- + B^+$$

Titrarea începe cu adăugarea unei mici cantităţi de bază în acid. În acest moment sunt prezente în soluţie speciile: H^+, HO^-, HA şi A^-. Din ecuaţia de disociere a acidului şi apei rezultă:

$$[H^+]\cdot[A^-] = K_a\cdot[HA],\ [H^+]\cdot[HO^-] = K_w$$

unde [·] este exprimă concentraţia molară ($[H^+]$ este concentraţia molară), K_a este constanta de aciditate iar K_w este constanta de disociere a apei la temperatura considerată. Dacă se aplică bilanţul de masă pentru acid şi sare, rezultă că C_a, concentraţia analitică a acidului şi respectiv C_s, concentraţia analitică a sării sunt date de:

$$C_a = [HA] + [H^+] - [HO^-],\ C_s = [A^-] - [H^+] + [HO^-]$$

După substituţiile corespunzătoare în ecuaţiile de mai sus, se obţine o ecuaţie de gradul 3 a pH-ului (ecuaţia Brønsted-Lowry [111], [112]), $[H^+] = x$:

$$x^3 + (K_a+C_s)*x^2 - (K_w+C_x*K_a)*x - K_w*K_a = 0$$

Ecuaţia de mai sus admite o soluţie unică în intervalul (0,1). Ţinând seama că:

$$C_s = C_b*V_x/(V_a+V_x),\ C_x = (C_a*V_a-C_b*V_x)/(V_a+V_x)$$

unde C_b este concentraţia analitică a bazei, V_x este volumul de bază adăugat, C_x concentraţia analitică a acidului după adăugare iar V_a este volumul iniţial de acid. Substituind în ecuaţia Brönsted, aceasta poate fi rezolvată numeric.

La punctul de echivalenţă, se porneşte modelul de la acelaşi punct iniţial şi se consideră toate echilibrele menţionate. La hidroliză mică, $C_s = [A^-] = [B^+]$ aşa că $[H^+] = x$:

$$x = \sqrt{\frac{K_w \cdot K_a \cdot (K_b + C_s)}{K_b \cdot (K_a + C_s)}}$$

După punctul de titrare, prin deduceri similare se obţine că $[H^+] = x$:

$$x^3+(K_w/K_b+C_x)*x^2-(K_w+C_s*K_w/K_b)*x-K_w^2/K_b = 0$$

unde expresiile lui C_x şi C_s sunt:

$$C_x = (C_b*V_x-C_a*V_a)/(V_a+V_x),\ C_s = C_a*V_a/(V_a+V_x)$$

Un program care implementează rezolvarea aceste probleme este disponibil online la adresa de Internet: http://l.academicdirect.org/Education/Training/titration/v1.1/.

Problemă: Modelarea procesului de acţiune enzimatică Michaelis-Menten

- Prima dată a fost observat la invertază [113];
- Ecuaţie: $S + E \leftrightarrow C \rightarrow P + E$, (S - substrat, E - enzimă, C - complex, P - produs (concentraţii: s, e, c, p);
- ipoteză: enzima (E) nu suferă modificări în ceea ce priveşte cantitatea sa totală (şi astfel nci concentraţia) în timp (e + c = constant);
- modelul cinetic permite exprimarea evoluţiei sistemului către echilibru; presupune scrierea ecuaţiilor de viteză de reacţie pentru fiecare reacţie elementară şi aplicarea principiului conservării numărului de atomi;

Soluţie: v. *Algoritm pentru simularea cineticii Michaelis-Menten*

1. Scrierea reacţiilor elementare; scrierea ecuaţiilor de viteză		
(1): $S + E \rightarrow_{k1} C,\ v_{(1)} = k_1\cdot s\cdot e$	(2): $C \rightarrow_{k2} S + E,\ v_{(2)} = k_2\cdot c$	(3): $C \rightarrow_{k3} P + E,\ v_{(3)} = k_3\cdot c$

2. Aplicarea principiului conservării numărului de atomi			
(S): $\dot{s} = v_{(2)} - v_{(1)}$	(E): $\dot{e} = v_{(2)} + v_{(3)} - v_{(1)}$	(C): $\dot{c} = v_{(1)} - v_{(2)} - v_{(3)}$	(P): $\dot{p} = v_{(3)}$

3. Presupuneri şi notaţii				
$s(0) = s_0$	$e(0) = e_0$	$c(0) = 0$	$p(0) = 0$	$e = e_0 - c$

4. Ecuaţii de rezolvat	
$\dot{s} = k_2c - k_1s(e_0 - c)$	$\dot{c} = k_1s(e_0 - c) - (k_2 + k_3)c$

5. Diferite abordări

aproximația "QSSA"	Briggs & Haldane, [114]	$\dot{c} = 0 \Rightarrow$	$c = \frac{e_0 s}{\kappa + s};\ -\dot{s} = \dot{p} = \frac{k_3 e_0 s}{\kappa + s};\ \kappa = \frac{k_2 + k_3}{k_1}$
aproximația "EA"	Henri [115]	$\dot{s} = 0 \Rightarrow$	$c = \frac{e_0 s}{\kappa + s};\ \dot{p} = \frac{k_3 e_0 s}{\kappa + s};\ \kappa = \frac{k_2}{k_1}$
Cazul general	ecuație implicită → nu are soluție analitică!	phases space	$\frac{dy}{dx} = b\frac{x - y - xy}{-x + ay + xy}$
	substituții în ecuația explicită $t = k_1 e_0 \tau$, τ timp inițial	$0 < a < 1$ $b > 0$	$a = \frac{k_2}{k_2 + k_3};\ b = \frac{k_2 + k_3}{k_1 e_0};\ x = \frac{k_1 s}{k_2 + k_3};\ y = \frac{c}{e_0}$ $\dot{x} = -x + ay + xy;\ \dot{y} = b(x - y - xy);$

6. Rezolvare numerică (i=1..n)

$x_0 = 3$ $y_0 = 0$	$x_{i+1}=x_i+\delta(-x_i+ax_i+x_iy_i)$ $y_{i+1}=y_i+b\delta(x_i-y_i-x_iy_i)$	$\delta=10^{-2}$ n=3000	$a \in \{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\}$	$b \in \{\frac{250}{25}, \frac{50}{25}, \frac{10}{25}, \frac{4}{25}\}$

7. Foaie de calcul Excel:

	A	B	C	D	E	F
1	x0=	3		i	xi	yi
2	y0=	0		=0	=B1	=B2
3	δ=	1.0e-2		=D2+1	=E2+$B$3*(-E2+$B$4*E2+E2*F2)	=F2+$B$5*$B$3*(E2-F2-E2*F2)
4	a=	0.2		…	…	…
5	b=	10		…	…	…

Algoritm pentru simularea cineticii Michaelis-Menten

Problemă: Modelarea mecanismului Lindemann - Hinshelwood al complexului activat

Prima oară comunicat de Lindemann [116] şi ulterior analizat de Hinshelwood [117] furnizează un mecanism de acțiune enzimatică. Se modelează mecanismul $R+R \leftrightarrow R^*+R \rightarrow P$.

Soluție: v. *Algoritm pentru simularea mecanismului Lindemann - Hinshelwood*

1. Scrierea reacțiilor elementare; scrierea ecuațiilor de viteză

(1): $R + R \rightarrow R^* + R$, $v_{(1)}=k_1[R]^2$	$R^* + R \rightarrow R + R$, $v_{(2)}=k_2[R][R^*]$	$R^* \rightarrow P$, $v_{(3)}=k_3[R^*]$

2. Aplicarea principiului conservării numărului de atomi (necunoscute [R] = x; [R*] = y; [P] = z):

(R): $\dot{x} = -v_{(1)} + v_{(2)}$	(R*): $\dot{y} = v_{(1)} - v_{(2)} - v_{(3)}$	(P): $\dot{z} = v_{(3)}$

3. Presupuneri şi notații

$r(0) = r_0$	$r^*(0) = 0$	$p(0) = 0$	$k_1 = a$	$k_2 = b$	$k_3 = c$

4. Ecuații de rezolvat

$\dot{x} = -ax^2 + bxy$	$\dot{y} = ax^2 - bxy - cy$	$\dot{z} = cy$

5. Abordare greşită

Căutarea unei soluții analitice este fără de succes.

6. Rezolvare numerică (i=1..n)

$x_0 = 3$ $y_0 = 0$ $z_0 = 0$	$x_{i+1}=x_i+\delta(-ax_i^2+bx_iy_i)$ $y_{i+1}=y_i+\delta(ax_i^2-bx_iy_i-cy_i)$ $z_{i+1}=z_i+\delta cy_i$	$\delta=10^{-2}$ n=3000	$a=10^{-1}$ $b=10^{-2}$ $c=10^{-3}$

7. Foaie de calcul Excel:

	A	B	C	D	E	F	G
1	x0=	1		i	xi	yi	zi
2	y0=	0		=0	=B1	=B2	=B3
3	z0=	0		=D2+1	=E2+(-B$1*E2^2+B$2*E2*F2)*B$4	=F2+(B$1*E2^2-B$2*E2*F2-B$3*F2)*B$4	=G2+B$3*F2*B$4
4	δ=	1e-2		…	…	…	…

5	a=	1e-1		…	…	…	…
6	b=	1e-2		…	…	…	…
7	c=	1e-3		…	…	…	…

Algoritm pentru simularea mecanismului Lindemann - Hinshelwood

Problemă: Simularea cineticii autocatalizei

Autocataliza este un caz special de cinetică chimică în care reacţiile au loc doar în prezenţa a atât reactanţilor cât şi a produşilor de reacţie şi au o ecuaţie generală de reacţie dată de: R → P, cu $v_{(1)} = k\cdot[R]\cdot[P]$.

Soluţie: v. *Algoritm pentru simularea autocatalizei*.

1. Scrierea reacţiilor elementare; scrierea ecuaţiilor de viteză

(1): R → P, $v_{(1)} = k\cdot[R]\cdot[P]$

2. Aplicarea principiului conservării numărului de atomi (necunoscute [R] = x; [R*] = y; [P] = z):

(R): $\dot{r} = -v_{(1)} = -k_1 rp$ | (P): $\dot{p} = v_{(1)} = k_1 rp$

3. Presupuneri şi notaţii

[R] = r	[P] = p	$k_1 = a$	$r+p=r_0+p_0=b$	$r(0)=r_0$	$p(0)=p_0$

4. Ecuaţii de rezolvat

$\dot{p} = a(b-p)p$ | $\dot{r} = -ar(b-r)$

5. Abordări - există soluţie analitică

$$\dot{p} = a(b-p)p \Rightarrow \frac{dp}{p(b-p)} = adt \Rightarrow \frac{1}{b}\ln\frac{p}{b-p} = at + c \Rightarrow \frac{p}{b-p} = e^{b(k_1 t+c)} \Rightarrow$$

$$p = \frac{b}{1+e^{-b(k_1 t+c)}} = \frac{b}{1+e^{-bk_1 t}e^{-bc}}$$

6. Constantele "b" & "c" - din valorile iniţiale ale concentraţiei (la momentul t = 0).

$$\frac{1}{b}\ln\frac{p(0)}{b-p(0)} = a0 + c \Rightarrow \frac{1}{b}\ln\frac{p_0}{r_0} = c;\ bc = \ln\frac{p_0}{r_0};\ -bc = \ln\frac{r_0}{p_0};\ e^{-bc} = \frac{r_0}{p_0}$$

7. Soluţia analitică şi interpretarea ei

$$p = p(t) = p_0\frac{r_0+p_0}{r_0 e^{-(r_0+p_0)k_1 t}+p_0};\ r = r_0 + p_0 - p$$

dacă $p_0 = 0$ atunci $p = 0$ & astfel nu evoluează
dacă $r_0 = 0$ atunci $p = p_0$ & astfel nu evoluează

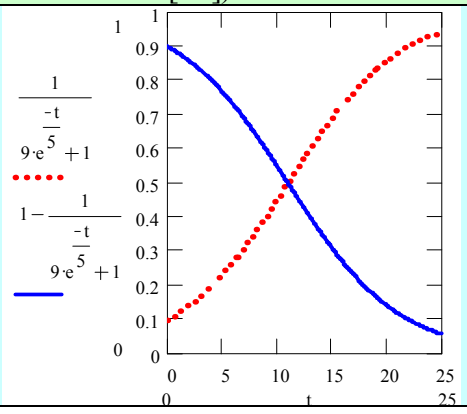
8. Desen în cazul $p_0 \neq 0$ şi $r_0 \neq 0$ (aplicaţie numerică, folosind MathCad [118])

dacă $p_0 = 0.1$; $r_0 = 0.9$; $k_1 = 0.2$ atunci

$$p(t) = \frac{1}{9e^{-0.2t}+1}$$

&

$$r(t) = 1 - \frac{1}{9e^{-0.2t}+1}$$



Algoritm pentru simularea autocatalizei

(Lotka - Volterra mechanism)

Proposed by Lotka for the first time in [119] as a complex mechanism in homogenous phase with damped oscillations. Later [120] a slightly changing of the mechanism was proposed when no damping occurs in the oscillations. Volterra made a statistical analysis of involving these equations [121]. By using same notations ([R]=r, [X]=x, [Y]=y, [P]=p) following algorithm is for the simulation of the Lotka - Volterra mechanism, in which the concentration of the reactant are kept constant by the

experimental design (such as adding substance or having a equilibrium with a nonmiscible phase containing it) - see *Algorithm for simulation of Lotka - Volterra kinetics*.

1. Writing of elementary reactions; writing of reaction rates equations

$R + X \rightarrow 2X$, $v_{(1)} = k_1 rx$	$X + Y \rightarrow 2Y$, $v_{(2)} = k_2 xy$	$Y \rightarrow P$, $v_{(3)} = k_3 y$	$P \rightarrow$, $v_{(4)} = k_4 p$

2. Equations to solve

$\dot{x} = v_{(1)} - v_{(2)} = k_1 rx - k_2 xy$	$\dot{y} = v_{(2)} - v_{(3)} = k_2 xy - k_3 y$	$\dot{p} = v_{(3)} + v_{(4)} = k_3 y + k_4 p$

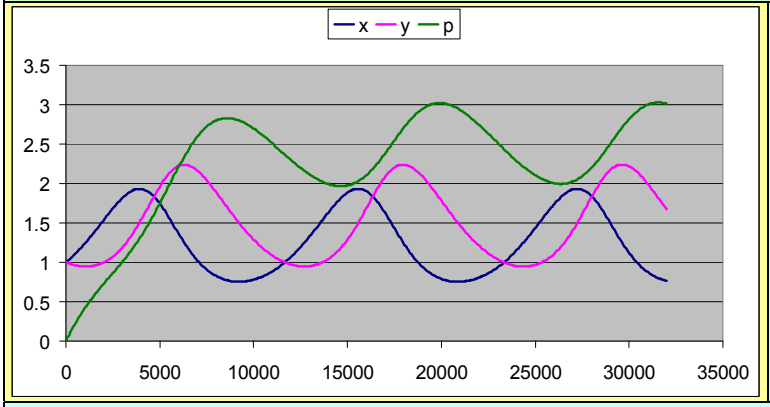
3. Solving numerically (i=1..n)

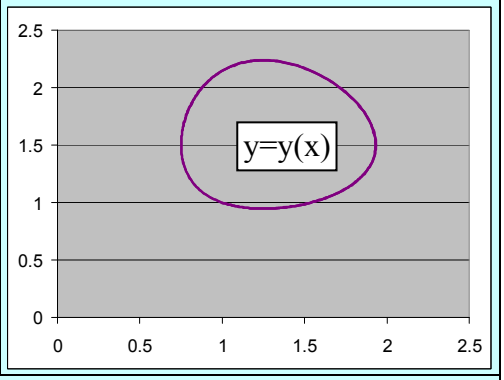
$x_0 = 3$ $y_0 = 0$ $z_0 = 0$	$x_{i+1} = x_i(1+(k_1 r - k_2 y_i)\delta)$ $y_{i+1} = y_i(1+(k_2 x_i - k_3)\delta)$ $p_{i+1} = p_i+(k_3 y_i - k_4 p_i)\delta$	$\delta=10^{-2}$ $n=5\cdot10^5$	$k_1=3$ $k_2=4$ $k_3=5$	$k_4=3$ $r=1$

7. Excel sheet:

	A	B	C	D	E	F	G
1	x0=	1		i	xi	yi	pi
2	y0=	1		=0	=B1	=B2	=B3
3	p0=	0		=D2+1	=D2*(1+(B$5*B$4 - B$6*E2)*B$9)	=E2*(1+(B$6*D2-B$7)*B$9)	= F2+(B$7*E2 -B$8*F2)*B$9
4	r=	2		…	…	…	…
5	k1=	3		…	…	…	…
6	k2=	4		…	…	…	…
7	k3=	5		…	…	…	…
8	k4=	3		…	…	…	…
9	δ=	1e-4		…	…	…	…

8. Plots:





Remarks: the y=y(x) equation is almost impossible to be extracted analitically, but we may extract it aproximately (numerically). Thus, for the simulated mechanism above, the y=y(x) equation is:

$$(x-1.32)^2+0.824\cdot(y-1.57)^2 = 0.35\pm0.05$$

Algorithm for simulation of Lotka - Volterra kinetics

(brusselator mechanism) - autocatalytic

A group of researchers from Bruxelles proposed for the first time a model in which phases space converges to a limit cycle [122]. Many authors modified it and studied systems working under this sort of mechanisms [123, 124, 125]. A simplified variant is given below (as in previous cases the concentration of reactant, *r*, are kept constant, see *Algorithm for simulation of brusselator kinetics*).

The equations of the system (see *Algorithm for simulation of brusselator kinetics*) does not goes all the time to a limit cycle independent of the values of the flow constants and/or initial concentrations. Trying to solve this system is full of surprises: for most of the combinations the equations provide a system going to a equilibrium point; there are values for which the system evolves with damped oscillations to the equilibrium; undamped oscillations have a significant weight between the general solutions of the system - fact proved by the nature system itself in which are numerous living systems basing their existence on such kind of oscillations. Heart periodical pulses are due to

such kind of process. Importance if this type of processes is big. This is the reason for which in 1997, Ilya PRIGOGINE received the Nobel prise in Chemistry for their theoretical studies on dissipative systems (see *Algorithm for simulation of Lotka - Volterra kinetics*).

1. Writing of elementary reactions and of reaction rates equations ([R]=r, [X]=x, [Y]=y, [P]=p):

$R \rightarrow X,\ v_{(1)} = k_1 r$	$X + 2Y \rightarrow 3Y,\ v_{(2)} = k_2 x y^2$	$Y \rightarrow P,\ v_{(3)} = k_3 y$

2. Applying of atoms conservation principle (unknowns [X] = x; [Y] = y):

(X): $\dot{x} = v_{(1)} - v_{(2)} = k_1 r - k_2 x y^2$	(Y): $\dot{y} = \upsilon_{(2)} - \upsilon_{(3)} = k_2 x y^2 - k_3 y$	(P): $\dot{p} = k_3 y$

3. Approach (simplifying r = 1, k_1 = 1 and k_3 = 1):

$\dot{x} = 1 - k_2 x y^2$	$\dot{y} = k_2 x y^2 - y$

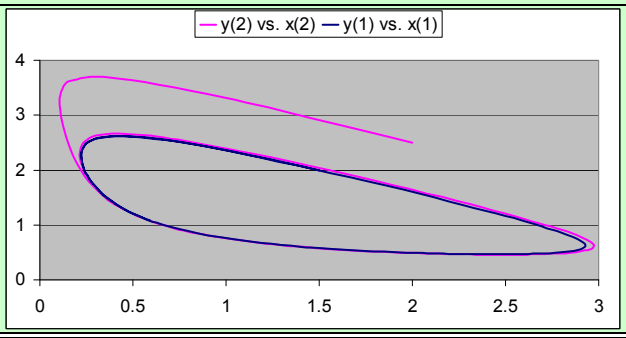
4. Solving numerically (i=1..n):

			Case 1	Case 2
$k_2 = 0.88$	$x_{n+1} = x_n + (1 - k_2 x_n y_n^2)\delta$ $y_{n+1} = y_n + (k_2 x_n y_n^2 - y_n)\delta$	$\delta = 10^{-2}$ n=10000	$x_0 = 1.5$ $y_0 = 2$	$x_0 = 2$ $y_0 = 2.5$

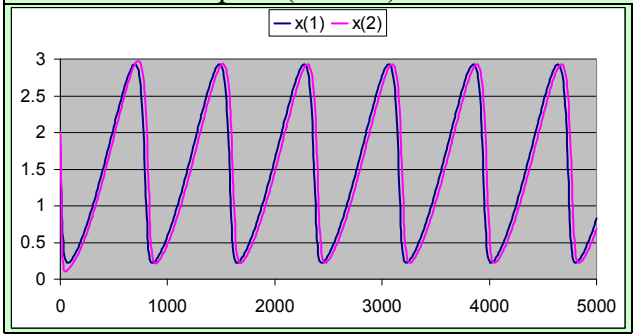
5. Excel sheet:

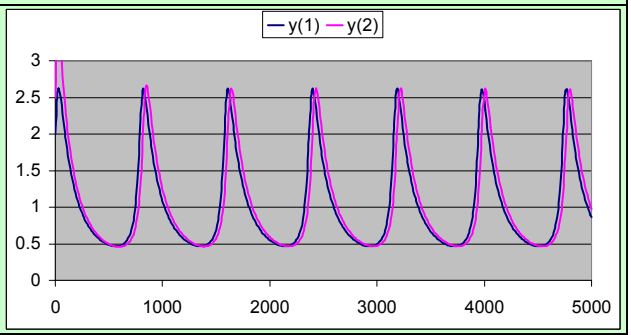
	A	B	C	D	E	F	G	H
1	k2=	0.88			Case 1		Case 2	
2	δ=	1e-2		i	xi	yi	xi	yi
3	Case	1		=0	=B4	=B5	=B7	=B8
4	x0=	1.5		=D2+1	=D3+(1-B$1*D3*E3^2)*B$2	=E3+(B$1*D3*E3^2-E3)*B$2	…	…
5	y0=	2		…	…	…	…	…
6	Case	2		…	…	…	…	…
7	x0=	2		…	…	…	…	…
8	y0=	2.5		…	…	…	…	…

6. Phases plot (n=1000):



7. Intermediaries plots (n=5000):





8. Simulation analysis

Not for any values the attractor appears. For a given k_2 (such is 0.88) exists minimum values of x_0 and y_0 ($x_{0\text{-min}}$, $y_{0\text{-min}}$) from which periodical oscillations occurs and the system tends to the attractor.

Algorithm for simulation of Lotka - Volterra kinetics

There is a fundamental difference between brusselator and Lotka-Volterra mechanisms evolution: if the L-V evolves oscillating around the initial values of the intermediaries, the brusselator converges (in time) to same equation, independent of the initial values of the intermediaries concentrations.

(oregonator mechanism)

Initiated by a group from Oregon (USA), implies 18 elementary stages and 21 different chemical species [126]. A simplified variant is given below (see *Algorithm for simulation of oregonator kinetics*).

1. Writing of elementary reactions and of reaction rates equations ([X]=x, and idem for the rest of):

$A + Y \rightarrow X$ $v_{(1)} = k_1 ay$	$X + Y \rightarrow P$ $v_{(2)} = k_2 xy$	$A + X \rightarrow 2X + Z$ $v_{(3)} = k_3 ax$	$2X \rightarrow Q$ $v_{(4)} = k_4 x^2$	$Z \rightarrow Y$ $v_{(5)} = k_5 z$

2. Applying of atoms conservation principle (unknowns [X] = x; [Y] = y, [Z] = z):

(X): $\dot{x} = k_1 ay - k_2 xy + k_3 ax - 2k_4 x^2$	(Y): $\dot{y} = -k_1 ay - k_2 xy + k_5 z$	(Z): $\dot{z} = k_3 ax - k_5 z$

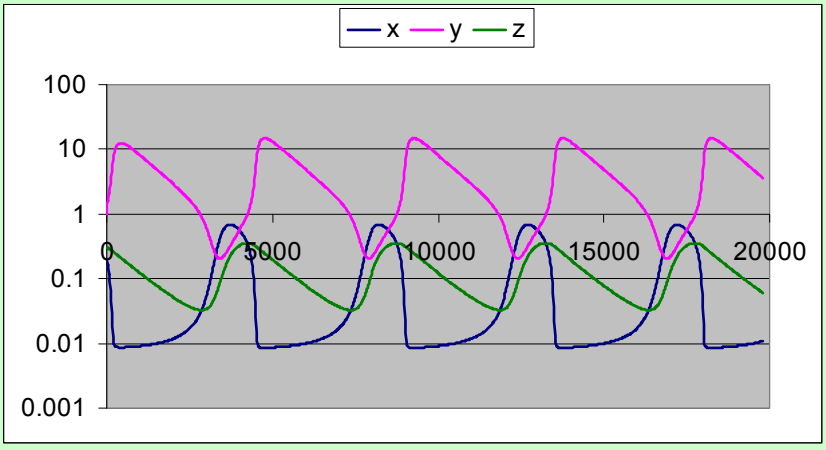
3. Solving numerically (i=1..n, after a long series of substitutions and rescaling):

$x_{n+1} = x_n + (qy_n - x_n y_n + x_n(1-x_n))\delta/\varepsilon$ $y_{n+1} = y_n + (-qy_n - x_n y_n + fz_n)\delta/\eta$ $z_{n+1} = z_n + (x_n - y_n)\delta$	x_0=0.2 y_0=1 z_0=0.3	ε= 8e-3 η=1e-1	q=2e-3 f=1	δ=1e-3 n=19800

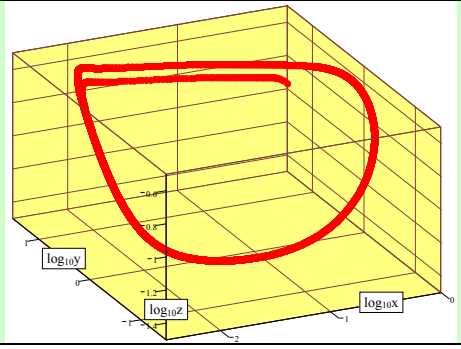
4. Excel sheet:

	A	B	C	D	E	F	F
1	q=	8e-3		i	xi	yi	zi
2	ε=	1e-1		=0	=B6	=B7	=B8
3	η=	2e-3		=D2+1	=IF(D2+(B$1*E2-D2*E2 +D2*(1-D2))*B$5/B$2>0, D2+(B$1*E2-D2*E2+ D2*(1-D2))*B$5/B$2,0)	=IF(E2+(-B$1*E2-D2*E2+ B$4*F2)*B$5/B$3>0, E2+(-B$1*E2-D2*E2+ B$4*F2)*B$5/B$3,0)	=IF(F2+ (D2-F2)*B$5>0, F2+(D2-F2)*B$5,0)
4	f=	1		…	…	…	…
5	δ=	1e-3		…	…	…	…
6	x0=	0.2		…	…	…	…
7	y0=	1		…	…	…	…
8	z0=	0.3		…	…	…	…

5. Intermediaries plots (n=1000):



6. Phases plot (z = z(x,y); 3D plot, MathCad7 [118]):



Algorithm for simulation of oregonator kinetics

Referințe

[1] George BOOLE, 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.
[2] Ronald A. FISHER, 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
[3] Wiki, 2012. List of unsolved problems in physics (http://en.wikipedia.org/w/index.php?oldid=469618085, version of January 5, 2012): http://en.wikipedia.org/wiki/List_of_unsolved_problems_in_physics
[4] Werner HEISENBERG, 1927. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik, Zeitschrift für Physik 43(3-4):172-198, doi: 10.1007/BF01397280
[5] Schrödinger E, 1926. An Undulatory Theory of the Mechanics of Atoms and Molecules. Physical Review 28(6): 1049-1070.
[6] Barry N. TAYLOR, Ambler THOMPSON (Eds.), 2008. The international system of units (SI). NIST Special Publication 330. Gaithersburg: National Institute of Standards and Technology.
[7] Carl WAGNER, 1931. On the theory of the rectifier effect (In German). Physikalische Zeitschrift 32:641-645.
[8] William D. PHILLIPS, Harold J. METCALF, 1982. Laser Deceleration of an Atomic Beam. Physical Review Letters 48(9):596-599.
[9] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, D.E. Knuth, 1996. On the Lambert W function. Advances in Computational Mathematics 5(1): 329-359.
[10] Roberto C. SOTERO, Iturria-Medina YASSER, 2011. From blood oxygenation level dependent (BOLD) signals to brain temperature maps. Bulletin of Mathematical Biology 73(11):2731-2747.
[11] M.A. Jenkins, 1975. Algorithm 493: Zeros of a Real Polynomial [C2]. ACM Transactions on Mathematical Software 1(2):178-189.
[12] http://www.netlib.org/toms/493
[13] William H. PRESS, Saul A. TEUKOLSKY, William T. VETTERLING, Brian P. FLANNERY, 1992. Gauss-Jordan Elimination (§2.1) and Gaussian Elimination with Backsubstitution (§2.2) In: Numerical Recipes in FORTRAN: The Art of Scientific Computing (2nd ed.) Cambridge: Cambridge University Press p. 27-32 (§2.1) and 33-34 (§2.2).
[14] Ronald A. FISHER, 1923. Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties. Journal of Agricultural Science 13:311-320.
[15] Isaac NEWTON, 1711. Analysis through a series of quantities, integrals, and differentials: with a list of lines of the third order (in Latin). London: William Jones.
[16] Robert COTES, 1722. Harmony of measures (in Latin). Cambridge: Robert Smith.
[17] Thomas SIMPSON, 1823. Doctrine and Application of Fluxions. Edinburgh: Bell and Bradfute.
[18] Teorema Limită Centrală

÷ Cronologia contribuțiilor 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.

÷ Enunțul teoremei (fie $(X_n)_{n\geq 1}$ variabile independente și $\exists\delta>0$ 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 $\frac{\sum_{i=1}^{n}(X_n-\mu_1(X_n))}{\sqrt{\sum_{k=1}^{n}\sigma_k^{\,2}}} \xrightarrow[n\to\infty]{} N(0,1)$

[19] Carl F. GAUSS, 1809. Theoria Motus Corporum Coelestium. Perthes et Besser, Hamburg. Translated, 1857, as Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections, trans. C. H. Davis. Little, Brown; Boston. Reprinted, 1963, Dover, New York.
[20] Pierre S. LAPLACE, 1812. Théorie Analytique des Probabilités. Paris: Courcier.
[21] Ronald A. FISHER, 1920. A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. Monthly Notices of the Royal Astronomical Society 80:758-770.
[22] Fisher, R. A. and Mackenzie W. A., 1923. Studies in Crop Variation. II. The manurial response of different potato varieties. Journal of Agricultural Science 13:311-320.

[23] R. A. Fisher, 1924. The Conditions Under Which χ2 Measures the Discrepancy Between Observation and Hypothesis. Journal of the Royal Statistical Society 87:442-450.
[24] R. A. Fisher, 1912. On an Absolute Criterion for Fitting Frequency Curves. Messenger of Mathematics 41:155-160.
[25] BENFORD Frank. 1938. The law of anomalous numbers. Proceedings of the American Philosophical Society 78(4):551-572.
[26] HILL Theodore P. 1995. Base invariance implies Benford's Law. Proceedings of the American Mathematical Society 123(3):887-895.
[27] Carlos M JARQUE, Anil K BERA. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ Lett* 6(3):255-259.
[28] 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.
[29] KOLMOGOROV Andrey. 1941. Confidence Limits for an Unknown Distribution Function. The Annals of Mathematical Statistics 12(4):461-463.
[30] SMIRNOV Nikolay V. 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19(2):279-281.
[31] 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.
[32] Fritz W SCHOLZ, Michael A STEPHENS. 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association 82(399):918-924.
[33] 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.
[34] Lorentz JÄNTSCHI. 2009. http://l.academicdirect.org/Statistics/tests/kAD/, k-sample Anderson-Darling.
[35] 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.
[36] 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
[37] 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.
[38] Ronald A. FISHER, 1935. The Mathematical Distributions Used in the Common Tests of Significance. Econometrica 3:353-365.
[39] LIEBETRAU Albert M. 1983. Measures of association. Newbury Park, CA: Sage Publications. Quantitative Applications in the Social Sciences 32:1-96 (p.13).
[40] 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.
[41] 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.
[42] SNEDECOR George W. and COCHRAN William G. 1989. Statistical Methods, Eighth Edition, Iowa State University Press.
[43] HARTLEY Ralph V L. 1928. Transmission of Information. Bell System Technical Journal 1928:535-563.
[44] Software. 2008. EasyFit v.5. MathWave Technologies. http://www.mathwave.com
[45] SCOTT David. 1992. Multivariate Density Estimation. John Wiley, Chapter 3.
[46] Software. 2005. Dataplot. National Institute for Standards and Technology. http://www.itl.nist.gov/div898/software/dataplot.html
[47] Isaac NEWTON, 1687. Philosophiæ naturalis principia mathematica. London: Juffu Societatis Regiæ, full text: http://www.ntnu.no/ub/spesialsamlingene/ebok/02a019654.html
[48] Julian L. COOLIDGE, 1949. The Story of the Binomial Theorem. The American Mathematical Monthly 56(3):147-157. full text: http://www.jstor.org/stable/2305028
[49] Jacobi BERNOULLI, 1713. Ars Conjectandi. Basel: Thurnisius.
[50] Abraham De MOIVRE, 1738. Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi (presented privately to some friends in 1733), p. 235-243 In: The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play (2nd ed.), London: W. Pearson.
[51] Johann C. F. GAUSS, 1823. Theoria combinations observationum erroribus minnimis obnoxiae. Pars prior, et Pars posterior. Comment. Societ. R. Sci. Gottingensis Recentiores 5:33-90.
[52] Abraham WALD, 1939. Contributions to the Theory of Statistical Estimation and Testing Hypothesis. The Annals of Mathematical Statistics 10(4):299-326.
[53] Allan AGRESTI, Brent A. COULL, 1998. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. American Statistician 52(2):119-126.
[54] Allan AGRESTI, Yongyi MIN, 2001. On small-sample confidence intervals for parameters in discrete distributions. Biometrics 57(3):963-971.

[55] Allan AGRESTI, 2001. Exact inference for categorical data: Recent advances and continuing controversies. Statistics in Medicine 20(17-18):2709-2722.
[56] Allan AGRESTI, 2002. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Statistical Methods in Medical Research 12(1):3-21.
[57] Wolfgang K. ENGEL, 1973. Onset of synthesis of mitochondrial enzymes during mouse development. Synchronous activation of parental alleles at the gene locus for the M form of NADP dependent malate dehydrogenase, Humangenetik 20(2):133-140.
[58] David L. MEADOWS, Jerome S. SCHULTZ, 1991. A molecular model for singlet/singlet energy transfer of monovalent ligand/receptor interactions. Biotechnology and Bioengineering 37(11):1066-1075.
[59] Mark D. SZCZELKUN, 2002. Kinetic models of translocation, head-on collision, and DNA cleavage by type I restriction endonucleases. Biochemistry 41(6):2067-2074.
[60] Tijen TANYALÇIN, François J. M. EYSKENS, Eddy PHILIPS, Marc LEFEVERE, Benal BÜYÜKGEBIZ, 2002. A marked difference between two populations under mass screening of neonatal TSH and biotinidase activity. Accreditation and Quality Assurance 7(11):498-506.
[61] Enrique A. OSSET, Mercedes FERNÁNDEZ, Juan A. RAGA, Aneta KOSTADINOVA, 2005. Mediterranean Diplodus annularis (Teleostei: Sparidae) and its brain parasite: Unforeseen outcome. Parasitology International 54(3):201-206.
[62] Clarke R. CONANT, Marc R. Van GILST, Stephen E. WEITZEL, William A. REES, Peter H. von HIPPEL, 2005. A Quantitative Description of the Binding States and In Vitro Function of Antitermination Protein N of Bacteriophage? Journal of Molecular Biology 348(5):1039-1057.
[63] Matthew A. CARLTON, William D. STANSFIELD, 2005. Making babies by the flip of a coin? American Statistician 59(2):180-182.
[64] Ronald A. FISHER, 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22:700-725.
[65] Ronald A. FISHER, 1912. On an Absolute Criterion for Fitting Frequency Curves. Messenger of Mathematics 41:155-160.
[66] Evan COOCH, Gary WHITE, 2011. Program MARK. A gentle introduction. Internet: E-book (22 Dec. 2011, 927p.).
[67] Bernard ROSNER, 1995. Hypothesis Testing: Categorical Data (p. 345-442) In: Fundamentals of Biostatistics (4th ed.). Belmont: Duxbury Press.
[68] Robert G. NEWCOMBE, 1998. Two-sided confidence intervals for the single proportion; comparison of seven methods. Statistics in Medicine 17(8):857-872.
[69] Ana M. PIRES, 200X. Confidence intervals for a binomial proportion: comparison of methods and software evaluation (web only). URL: http://www.math.ist.utl.pt/~apires/AP_COMPSTAT02.pdf
[70] Lawrence D. BROWN, Tony T. CAI, Anirban DasGUPTA, 2001. Interval estimasion for a binomial proportion. Statistical Science 16(2):101-133.
[71] Edwin B. WILSON, 1927. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 22(158):209-212.
[72] J. R. ANDERSON, Leslie E. BERNSTEIN, Malcolm C. PIKE, 1982. Approximate Confidence Intervals for Probabilities of Survival and Quantiles in Life-Table Analysis. Biometrics 38(2):407-416.
[73] Barnet WOOLF, 1955. On estimating the relation between blood group and disease. Annals of Human Genetics 19:251-253.
[74] John J. GART, 1966. Alternative analyses of contingency tables. Journal of Royal Statistical Society B28:164-179.
[75] Ronald A. FISHER, 1956. Statistical Methods for Scientific Inference. Edinburgh: Oliver and Boyd.
[76] C. J. Clopper, Egon S. PEARSON, 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404-413.
[77] Allan AGRESTI, 2003. Dealing with discretness: making ‘exact’ confidence intervals for proportions, differences of proportions, and odds ratios more exact. Statistical Methods in Medical Research 12:3-21.
[78] Harold JEFFREYS. Theory of Probability (3rd Ed). Clarendon Press, Oxford, 1961. http://lccn.loc.gov/62000074
[79] Colin R. BLYTH, Harold A. STILL, 1983. Binomial confidence intervals. Journal of the American Statistical Association 78(381):108-116.
[80] George CASELLA, 1986. Refining binomial confidence intervals. The Canadian Journal of Statistics 14(2):113-129.
[81] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2008. Optimized Confidence Intervals for Binomial Distributed Samples. International Journal of Pure and Applied Mathematics 47(1):1-8.
[82] Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, 2007. How to Asses Dose-Response Study Outcome: a Statistical Approach. Recent Advances in Synthesys & Chemical Biology VI:P36.
[83] Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, 2010. Exact Probabilities on Confidence Limits for Binomial Samples: Applied to the Difference between Two Proportions, TheScientificWorldJOURNAL 10(5):865-878.
[84] Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI, 2008. Assessment of Confidence Intervals used in Medical Studies (in Romanian), Cluj-Napoca: AcademicPres & AcademicDirect, p. 234 (+p. 10 intro), 2008.

[85] Lorentz JÄNTSCHI, 2011. Sampling distribution in biodiversity analysis - Project proposal PN-II-ID-PCE-2011-3-0103, not funded (I was found "not eligible", see UEFISCDI.ro website).
[86] Maurice G. KENDALL, 1962. Rank Correlation Methods (3rd Ed). London: C. Griffin & Co.
[87] Karl F. R. S. PEARSON, 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(302):157-175. doi: 10.1080/14786440009463897
[88] Charles SPEARMAN, 1904. The proof and measurement of association between two things. Amer. J. Psychol. 15:72-101.
[89] Maurice G. KENDALL, Babington B. SMITH, 1939. The Problem of m Rankings. The Annals of Mathematical Statistics 10(3):275-287.
[90] Augustin-Louis CAUCHY, 1821. Oeuvres 2. III, p. 373
[91] Viktor BUNYAKOVSKY, 1859. Sur quelques inegalités concernant les intégrales aux différences finis (in French). Mem. Acad. Sci. St. Petersbourg 7(1):9.1-9.18.
[92] Hermann A. SCHWARZ, 1888. 15. Einführung der Constante c In: Über ein Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung (in German). Acta Societatis Scientiarum Fennicæ XV:344-345.
[93] Maurice H. QUENOUILLE, 1949. Approximate tests of correlation in time-series. J R Stat Soc B 11(1):68-84.
[94] Maurice H. QUENOUILLE, 1956. Notes on bias in estimation. Biometrika 43(3-4):353-360.
[95] Ronald A. FISHER, 1948. Combining independent tests of significance. The American Statistician 2(5):30.
[96] Bradley EFRON, 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7(1):1-26.
[97] Edgar C. FIELLER, Herman O. HARTLEY, Egon S. PEARSON, 1957. Tests for rank correlation coefficients. I. Biometrika 44:470-481.
[98] Maurice G. KENDALL, Alan STUART, 1973. §31.19 & §31.21, In: The Advanced Theory of Statistics, Volume 2: Inference and Relationship. London: C. Griffin. http://lccn.loc.gov/77360686
[99] Maurice G. KENDALL, 1938. A New Measure of Rank Correlation. Biometrika 30(1-2):81-89.
[100] Leo A. GOODMAN, William H. KRUSKAL, 1954. Measures of Association for Cross Classifications. Journal of the American Statistical Association 49(268):732-764.
[101] Leo A. GOODMAN, William H. KRUSKAL, 1972. Measures of Association for Cross Classifications IV: Simplification of Asymptotic Variances. Journal of the American Statistical Association 67(338):415-421.
[102] Wassily HOEFFDING, 1948. A Non-parametric Test of Independence. Annals of Mathematical Statistics, 19(4):546-557.
[103] Julius R. BLUM, James KIEFER, Murray ROSENBLATT, 1961. Distribution Free Tests of Independence Based on the Sample Distribution Function. Ann Math Statist 32(2):485-498.
[104] Robert H. SOMERS, 1962. A new asymmetric measure of association for ordinal variables. American Sociological Review 27(6):799-811.
[105] Maurice G. KENDALL, 1990. Rank Correlation Methods. Oxford: Oxford University Press.
[106] Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, 2009. Observation vs. Observable: Maximum Likelihood Estimations according to the Assumption of Generalized Gauss and Laplace Distributions. Leonardo Electronic Journal of Practices and Technologies 8(15):81-104.
[107] Pierre-Simon LAPLACE, 1814. Théorie analytique des probabilité (in French). Paris: Courcier.
[108] Carl Friedrich GAUSS. 1809. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Paris: Treuttel & Würtz. London: R.H. Evans.
[109] Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, 2007. The Jungle of Linear Regression Revisited. Leonardo Electronic Journal of Practices and Technologies 6(10):169-187.
[110] William H. PRESS, Saul A. TEUKOLSKY, William T. VETTERLING, Brian P. FLANNERY, 1992. Gauss-Jordan Elimination (§2.1) and Gaussian Elimination with Backsubstitution (§2.2) In: Numerical Recipes in FORTRAN: The Art of Scientific Computing (2nd ed.) Cambridge: Cambridge University Press p. 27-32 (§2.1) and 33-34 (§2.2).
[111] Bronsted J.N., 1923. Some remarks on the concept of acids and bases (In German). Recueil des Travaux Chimiques des Pays-Bas 42(8):718-728, DOI: 10.1002/recl.19230420815
[112] Lowry T.M., 1923. The Uniqueness of Hydrogen. Chemistry and Industry 42(3):43-47. DOI: 10.1002/jctb.5000420302
[113] Leonor MICHAELIS, Maud L. MENTEN, 1913. Die Kinetik der Invertinwirkung. Biochem Z 49:333-369.
[114] George E. BRIGGS, John B.S. HALDANE, 1925. A note on the kinematics of enzyme action. Biochem J 19(2):338-339.
[115] Victor HENRI, 1903. Lois Générales de l'Action des Diastases. Paris: Hermann.
[116] Frederick A. LINDEMANN, Svante ARRHENIUS, Irving LANGMUIR, N. R. DHAR, J. PERRIN, W. C. McC. LEWIS, 1922. Discussion on "the radiation theory of chemical action". Transactions of the Faraday Society, 17:598-606. doi: 10.1039/TF9221700598
[117] Cyril N. HINSHELWOOD, 1926. On the Theory of Unimolecular Reaction. Proc. R. Soc. Lond. A 113:230-233. DOI:10.1098/rspa.1926.0149
[118] MathSoft, 1997. MathCad7 Professional (software). MathSoft. URL: http://www.mathsoft.com

[119] Alfred J. LOTKA, 1909. Contribution to the Theory of Periodic Reactions. The Journal of Physical Chemistry 14(3):271-274. doi: 10.1021/j150111a004
[120] Alfred J. LOTKA, 1909. Undamped oscillations derived from the law of mass action. Journal of the American Chemical Society 42(8):1595-1599. doi: 10.1021/ja01453a010
[121] Vito VOLTERRA, 1926. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem. Acad. Lincei Roma 2:31-113.
[122] Ilya PRIGOGINE, Grégoire NICOLIS, 1967. On Symmetry-Breaking Instabilities in Dissipative Systems. J Chem Phys 46(9):3542-3550.
[123] Richard M. NOYES, 1976. Oscillations in chemical systems. XII. Applicability to closed systems of models with two and three variables. The Journal of Chemical Physics 64(4):1266-1269.
[124] G. B. COOK, Peter GRAY, D. G. KNAPP, Stephen K. SCOTT, 1989. Bimolecular routes to cubic autocatalysis. The Journal of Physical Chemistry 93(7):2749-2755.
[125] Daniel T. GILLESPIE, 1977. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry 81(25):2340-2361.
[126] Richard J. FIELD, Richard M. NOYES, 1974. Oscillations in Chemical Systems IV. Limit cycle behavior in a model of a real chemical reaction, J. Chem. Phys. 60:1877-1884. DOI:10.1063/1.1681288