Legăturile dintre fenomenele colective

Statistica studiază *legăturile cauzale* dintre fenomenele colective. Legile naturale îmbracă întotdeauna forma enunţării unei legături fie între manifestările a două sau mai multor fenomene, fie între prezenţa şi absenţa anumitor caractere Spunem, de exemplu, că electricităţile de sens contrar se atrag, iar cele de acelaşi sens se resping, că temperatura descreşte odată cu creşterea altitudinii, că infracţionalitatea creşte dacă scade nivelul de trai Acestea sunt legi cunoscute. Prima este o lege rigidă, se verifică, prin urmare, aplicată oricărui caz. Celelalte sunt legi *stochastice*..

Forma sub care am enunţat legea stochastică de mai sus constă numai în existenţa legăturii dintre fenomene şi prin aceasta suntem privaţi de cunoaşterea unui element destul de interesant în această direcţie: *gradul de asociaţie între cele două fenomene sau caractere*.

Prezentul capitol este consacrat mijloacelor de care dispune statistica matematică pentru a măsura acestui nou element.

Este clar că o definiţie perfect circumscrisă a gradului de legătură între două variabile nu poate fi atinsă. În procesul de cercetare a legăturilor cauzale existente între fenomenele naturale avem de-a face cu contopirea acţiunilor unei multitudini de factori (cauze), dintre care unii esenţiali, alţii neesenţiali, unii pot fi determinaţi, alţii nu. Tocmai de aceea, în cercetarea legăturii reciproce dintre două fenomene apar dificultăţi mari deoarece pot să existe cauze necunoscute. În astfel de situaţii este util să determinăm gradul de corelare şi apoi să analizăm separat unele din aceste cauze. Cu alte cuvinte, trebuie să definim relaţii posibile între diferiţi factori, evenimente, atribute sau caracteristici care ar putea avea o influenţă, cel puţin parţială asupra datelor experimentale. În acest mod este posibilă stabilirea unui tablou al condiţiilor în care se desfăşoară un anumit fenomen natural, fapt care duce la modelarea matematică a sa. Din această cauză, diferite puncte de vedere au putut sta alături pe această chestiune, fără ca să se decidă care este cel mai adecvat. Şi natural, fiecare din ele a condus la un mijloc diferit de măsură a acestei legături.

În multe procese naturale, pe lîngă complexa întrepătrundere cu alte fenomene (procese), acestea mai sunt supuse unor evoluţii care la prima vedere pot fi considerate probabiliste (aleatoare). De aceea, pentru cunoaşterea modului de evoluţie probabilă în viitor - prognoza unui anumit fenomen - trebuie să ne bazăm pe *cunoaşterea evoluţiei trecute*, precum şi pe situaţia prezentă. Experienţa unui mare număr de observaţii (probe) au dus la concluzia că între diferitele mărimi variabile pot exista următoarele tipuri de relaţii:

a) ***Relaţia de dependenţă,*** Y depinde de X sau invers. O modificare a unei variabile duce la o modificare a celei de-a doua. În cazul unei astfel de relaţii s-ar putea aminti existenţa relaţiei cauzale în care o variabilă este cauza, iar cealaltă este efectul, cauza fiind un fenomen sau un complex de fenomene care provoacă, generează sau determină un alt fenomen - efectul. Operaţia logică prin care efectul este dedus din cauză se numeşte ***inferenţă cauzală***.

b) ***Relaţia de interdependenţă***, Y depinde de X şi X depinde de Y. În acest caz modificarea unei variabile provoacă modificarea celei de-a doua variabile, iar modificarea acesteia din urmă are influenţă asupra primei variabile.

c) ***Relaţia de tranziţie***, X se transformă parţial sau total în Y şi invers.

d***) Corelaţia statistică*** sau ***covarianţa***. Pentru X există întotdeauna Y şi invers. Este o relaţie reciprocă dintre două variabile, dintre care una în mod logic apelează la alta şi pe baza analizei datelor experimentale se poate pune în evidenţă o asociere între ele.

e) ***Relaţia stochastică***; dacă se realizează X atunci cu o anumită probabilitate se realizează şi Y, sau invers.

Datorită caracterului complex al fenomenelor şi datorită multitudinii de factori esenţiali şi întâmplători care intervin, aceste legături se manifestă *sub formă* ***de tendinţă;*** ele pot fi identificate în condiţiile acţiunii legii numerelor mari, în colectivităţile de volum ridicat.

O problemă importantă în cadrul analizei seriilor interdependente o reprezintă ***identificarea legăturilor*** *cu adevărat* ***semnificative şi stabile.***

În cazul în care s-a stabilit că între două variabile există un raport de dependenţă, se ridică o altă problemă, mult mai dificilă: găsirea unei ***măsuri a acestei legături***, prin intermediul unui indicator sintetic de corelaţie.

Există însă şi situaţii înşelătoare, în care variaţiile (creşterea sau scăderea) celor două fenomene aparent interdependente sunt similare, ***dar nu există nici o legătură logică între ele***. De exemplu, în ultimii ani în România cresc simultan atât rata sărăciei, cât şi înzestrarea populaţiei cu telefoane mobile. Alteori explicaţia *tendinţei de asociere* a celor două variabile este dată de ***existenţa unei cauze comune.*** De exemplu, creşterea simultană şi semnificativă a vânzărilor la pulovere de lână şi la medicamente antigripale are aceeaşi cauză: venirea iernii. Analiza calitativă a datelor statistice rezolvă probleme de acest tip.

SERIILE INTERDEPENDENTE

TIPURI DE LEGĂTURI

Analiza seriilor interdependente urmăreşte

1. ***verificarea existenţei*** şi
2. ***măsurarea intensităţii legăturilor dintre fenomene.***

Cu cât fenomenele sunt mai complexe, cu atât numărul factorilor care le influenţează este mai mare, ceea ce face legăturile cauzale dificil de evidenţiat. Analiza dependenţelor se complică atunci când factorii de influenţă intercondiţionează determinând apariţia de cauzalităţi în lanţ.

Pentru a putea determina intensitatea relaţiilor de corelaţie, este necesar să se analizeze, în primul rând, conţinutul şi forma acestor relaţii.

Vom nota cu:

- *y* fenomenul a cărui variaţie este influenţată (*variabilă dependentă sau rezultativă*) şi cu
- *x* factorul de influenţă (*variabila independentă sau factorială*).

După **natura legăturilor de interdependenţă** deosebim două categorii de relaţii:

a) funcţionale (deterministe) şi

b) ***stochastice (statistice).***

a**) *Funcţionale: y = f (x).***

Fenomenul cauză x determină în mod univoc fenomenul efect *y*, astfel încât *unei valori a variabilei x îi corespunde o valoare unică a variabilei y.*.

b**) *Stochastice: y = f (x_1 , x_2 ,..., x_n).***

Fenomenul efect este influenţat de o multitudine de factori, esenţiali şi întâmplători. *Unei valori a fenomenului cauză x_i îi pot corespunde mai multe valori diferite ale fenomenului efect y, în funcţie de acţiunea combinată a celorlalţi factori de influenţă.* De exemplu, infracţionalitatea unei persoane depinde de: profilul psihologic,sex, gradul de educaţie, gradul de sărăcie, situaţia familială, starea de sănătate etc.

Fiecare caracteristică x_i determină numai o parte a variaţiei fenomenului *y*, restul variaţiei fiind explicat prin alte caracteristici, care din punct de vedere al legăturii x_i -*y* sunt întâmplătoare.

Variaţia fenomenului *y* poate fi analizată în funcţie de unul sau mai mulţi factori de influenţă (x_1, x_2,... , x_n), dar întotdeauna va rămâne *o variaţie neexplicată*, determinată de factorii neînregistraţi.

În cazul în care se identifică şi se analizează factorii de influenţă esenţiali, componenta aleatoare, care sintetizează acţiunea factorilor întâmplători, va avea o pondere redusă şi nu va influenţa semnificativ veridicitatea rezultatelor.

Legăturile stochastice sunt specifice fenomenelor din societate şi economie.

Extrema diversitate a legăturilor stochastice impune sistematizarea lor după mai multe criterii.

După **numărul caracteristicilor** analizate, legăturile stochastice pot fi simple sau multiple.

a) *Legături simple* - *se alege o singură caracteristică determinantă pentru variaţia fenomenului y,* toate celelalte caracteristici care îl influenţează, fie că sunt esenţiale sau întâmplătoare, sunt considerate cu acţiune constantă.
De exemplu, analiza legăturii dintre recolta totală şi suprafaţa cultivată.

b) *Legături multiple* - *se analizează variaţia fenomenului y în funcţie de mai multe caracteristici esenţiale x_1, x_2,..., x_n.* Rămâne şi în acest caz o componentă aleatoare, care sintetizează acţiunea, presupusă constantă, a celorlalţi factori de influenţă.
De exemplu, analiza variaţiei salariului într-o colectivitate în funcţie de productivitate, vechime şi calificare.

După **natura caracteristicii** pot exista legături stochastice de asociere sau de corelaţie.

a) ***Asocierea statistică*** se referă la raporturile de interdependenţă dintre caracteristicile calitative, sau dintre o caracteristică numerică şi una calitativă. De exemplu, legătura dintre locul de muncă şi studii, calificare şi productivitate, , între zona geografică şi clasa de fertilitate a terenurilor agricole etc.
Analiza statistică a raporturilor de asociere este posibilă doar dacă se găseşte o modalitate de exprimare numerică a variantelor. De exemplu, clasele de calitate ale produselor pot fi codificate şi ierarhizate: 0 – produs inferior, 1 - produs mediu, 2 - produs superior.

b) ***Corelaţia statistică*** exprimă raporturile de cauzalitate dintre două sau mai multe caracteristici exprimate cantitativ. De exemplu, analiza cifrei de afaceri în funcţie de numărul salariaţilor şi valoarea capitalului fix ,analiza gradului de poluare în funcţie de producţia de substanţe chimice .

După **sensul relaţiei de cauzalitate**, legăturile stochastice pot fi directe sau inverse.

a) ***Legături directe*** există atunci când modificarea într-un anumit sens (creştere sau scădere) a fenomenului cauză *x* determină modificarea în acelaşi sens a fenomenului efect *y*.
De exemplu, legătura dintre numărul salariaţilor şi volumul producţiei, dintre mărimea creditului şi masa dobânzii, dintre costul unitar şi costul total etc.

b) ***Legăturile inverse*** există atunci când modificarea într-un anumit sens a lui x determină o modificare în sens contrar a lui *y*. De exemplu, legătura dintre profitul unitar şi costul unitar de producţie, dintre impozitul pe profit şi profitul net, dintre mărimea dividentelor şi profitul reinvestit etc.

După **forma matematică** a legăturilor, acestea pot fi liniare sau neliniare.

a) ***Liniare****:* legătura se realizează după ecuaţia dreptei.

b) ***Neliniare****:* exponenţiale, hiperbolice, parabolice, logaritmice.

Forma legăturii este, de regulă, vizibilă pe grafic. Atunci când legătura grafică nu este clară, se poate continua analiza pe variantele sugerate de grafic folosind metode analitice şi folosind anumite criterii pentru a alege varianta cea mai bună.

După **momentul** producerii lor deosebim legături sincrone şi asincrone.

a) ***Sincrone*** - modificarea lui x determină modificarea imediată a lui y. De exemplu, creşterea veniturilor populaţiei determină mărirea imediată a cererii de consum, creşterea producţiei se obţine concomitent cu creşterea cheltuielilor etc.

b) ***Asincrone*** - fenomenul x determină variaţia fenomenului efect y după o perioadă de timp. De exemplu, legătura dintre investiţii şi creşterea producţiei sau legătura dintre rata dobânzii şi volumul masei monetare.

METODE ELEMENTARE DE VERIFICARE A EXISTENŢEI LEGĂTURILOR

Studierea legăturilor dintre fenomenele economice presupune parcurgerea mai multor etape:

1. ***depistarea factorilor*** care influenţează variaţia fenomenului analizat şi ***ierarhizarea*** acestora;
2. ***alegerea factorului sau factorilor esenţiali*** a căror influenţă asupra fenomenului dependent urmează să fie analizată;
3. culegerea şi sistematizarea datelor referitoare la variabilele studiate; verificarea gradului de cuprindere a datelor înregistrate (dacă datele provin dintr-un sondaj interpretarea rezultatelor se va face în sens probabilistic);
4. ***verificarea existenţei şi formei legăturii*** dintre caracteristicile corelate în vederea alegerii corecte a procedeelor statistico -matematice de măsurare a dependenţei statistice;
   Verificarea existenţei legăturilor se poate face cu ajutorul unor metode simple:
   - metoda seriilor paralele interdependente;
   - metoda grupărilor;
   - metoda tabelului de corelaţie;
   - metoda grafică;
   - analiza dispersională.

   Metodele elementare evidenţiază ***direcţia legăturii***, iar unele dintre ele pot indica şi ***forma*** acesteia. Aplicarea acestor metode trebuie completată cu o analiză calitativă a fenomenelor, bazată pe conţinutul lor, pe legătura logică dintre ele
5. ***măsurarea intensităţii legăturii*** cu ajutorul indicatorilor de corelaţie selectaţi în funcţie de forma de legătură şi de natura informaţiilor de care dispunem;
6. ***testarea semnificaţiei indicatorilor de corelaţie*** calculaţi când datele au provenit dintr-un sondaj.

Corelaţia statistică

Eficienţa aplicării metodei corelaţiei depinde de punerea (enunţarea) corectă a problemei în studiu precum şi de aplicarea corectă a statisticii matematice.

Caracterul complex al dependenţei statistice pune pe primul plan problema identificării existenţei legăturilor. Calculul indicatorilor de corelaţie este admis cu condiţia stabilirii anticipate a unei legături cauzale reale între fenomenele cercetate. Statistica nu poate să rezolve o astfel de problemă fără ajutorul ştiinţei din domeniul căreia face parte fenomenul studiat. Cu alte cuvinte, specialistul din domeniul respectiv trebuie să cunoască temeinic noţiunile analizei statistice implicate pentru a da o interpretare corectă a rezultatelor. Pentru a asigura deducţii suficient de

întemeiate, este necesar includerea în cercetare, dacă este posibil, a tuturor factorilor cu acţiune esenţială.

La fenomenele simple, unde cauzele acţionează separat, relaţia dintre fenomenul-efect şi fenomenul-cauză se reprezintă sub forma:

$y=f(x)$

unde x reprezintă cauza, iar y efectul.

La fenomenele complexe, dependenţa se exprimă sub forma generală:

$y = f(x_1 , x_2 ,..., x_n)$

Fenomenul y este generat de acţiunea comună a factorilor $x_1 ,x_2 ,.., x_n$ (cauze), din care luăm însă în calcul numai o parte.

Să admitem că am luat în calcul factorul x_1. Întrebarea care se pune este următoarea: în ce condiţii indicatorii corelaţiei obţinuţi exprimă măsura reală a influenţei variabilei x_1 asupra variabilei y? Numai cu condiţia ca factorul x_1 să fie hotărâtor în determinarea lui y, ceilalţi fiind nesemnificativi. În cazul în care fenomenul este sub acţiunea unui complex de factori esenţiali şi aceasta este situaţia obişnuită, pentru a exprima influenţa şi gradul de intensitate a legăturilor în raport cu un singur factor trebuie să eliminăm influenţa celorlaţi.

Să considerăm o colectivitate statistică caracterizată prin mărimile X şi Y. Efectuând o serie de determinări experimentale (sau observaţii) asupra acestei colectivităţi, putem întocmi tabela datelor respective:

X | $x_1, x_2 , ..., x_n$

Y | $y_1 , y_2 , ..., y_n$

Repartiţia empirică a celor două variabile se poate afişa grafic, într-un sistem de axe XOY, unde vom reprezenta punctele de coordonate x_i şi y_i. Un ansamblu de astfel de puncte se numeşte ***câmp de corelaţie***, ***tabel de corelaţie*** sau ***nor statistic***.

În Excel, acest lucru se poate face utilizând ***diagrama XY prin puncte(XY-scatter)***

Analiza vizuală a organizării şi formei norului de puncte obţinut poate oferi indicii importante asupra relaţiei dintre variabile.

Datele de sondaj vor susţine ipoteza asocierii între variabile dacă forma norului de puncte se apropie de o curbă funcţională.

Dacă punctele $M_i(x_i ,y_i)$ sunt distribuite de-a lungul unei fâşii, care în general, urmează o curbă determinată, spunem că între mărimile respective există o dependenţă funcţională. Dacă punctele $M_i(x_i ,y_i)$ nu arată o dependenţă funcţională strictă, dar există o tendinţă ca valorile lui Y să depindă de cele ale lui X , deşi nu în mod riguros, între mărimile X şi Y există o corelaţie Aceasta

poate să fie liniară (fig. 1) sau neliniară (fig. 2). În cazul când între X şi Y nu există nici un fel de dependenţă, câmpul de distribuţie se va prezenta asemănător cu acela arătat în fig. 3, 4. Cele două caracteristici sunt independente. Într-un caz particular, dependenţa corelaţionară se poate transforma într-o dependenţă funcţională, dar cu un anumit grad de certitudine. Apare problema de stabili cantitativ (numeric) în ce măsură dependenţa corelaţională se apropie sau se depărtează de dependenţa funcţională.

În foarte multe cazuri, din observarea fenomenelor naturale sau a proceselor sociale, fără a cunoaşte natura exactă a acestora şi nici cauzele prin care este pusă în evidenţă o anumită caracteristică, se pot trage concluzii foarte importante prin examinarea corelaţiei dintre aceste trăsături şi alte evenimente. În acest mod se poate aprecia existenţa unei relaţii statistice între două sau mai multe variabile, adică, în astfel de cazuri se vorbeşte despre corelaţii dintre mărimile care indică o dependenţă reciprocă.

Corelaţia este o metodă statistică de determinare a relaţiei dintre două variabile existente.

Regresia este o metodă statistică utilizată pentru a descrie natura relaţiei dintre variabile daca este pozitivă ,liniară sau neliniară

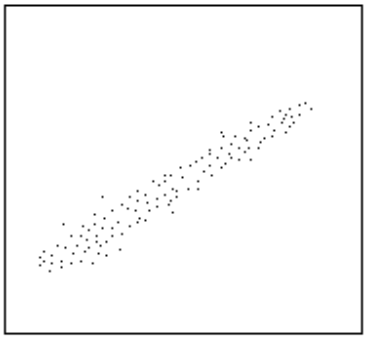


Figura 1 *Distribuţie liniară*

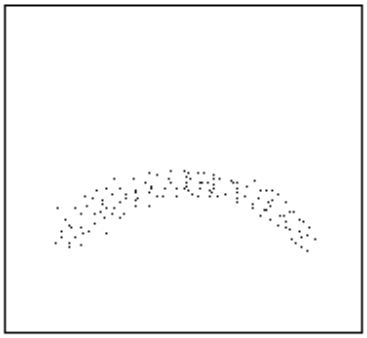


Figura 2 *Distribuţie neliniară*

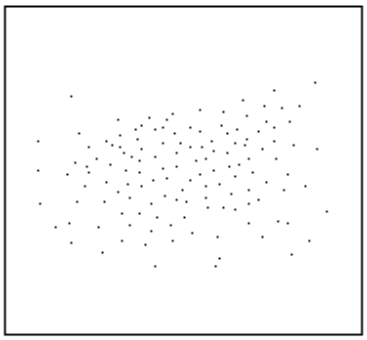


Figura 3 *Distribuţie aleatoare*

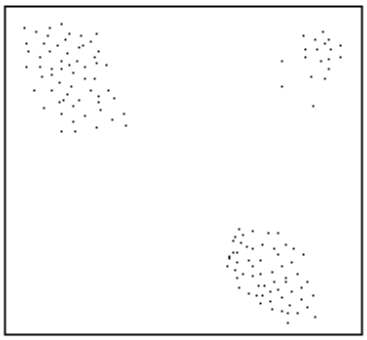


Figura 4 *Distribuţie grupată*

Fiind date două variabile X şi Y, se pune problema dacă între cele două variabile, respectiv între fenomenele descrise de acestea, există o anumită dependenţă numită şi ***corelaţie.*** O primă concluzie se poate obţine reprezentând grafic într-un sistem de coordonate XY, cele două şiruri de date observate pentru cele două variabile. Dacă punctele graficului se împrăştie pe toată suprafaţa fără a urma o anumită regulă, atunci vom spune că cele două variabile nu sunt corelate. Dacă în schimb punctele descriu o anumită curbă, numita şi ***curbă de regresie***, atunci vom spune că există corelaţie şi ea este cu atât mai intensă, cu cât domeniul pe care se întind punctele este mai îngust. Mai mult, dacă punctele se aşează pe o curbă care poate fi aproximată de o curbă clasică (dreaptă, parabolă, exponenţială, etc.) atunci vom spune că legătura dintre cele două variabile este una liniară sau parabolică sau exponenţială, etc. şi vom folosi ecuaţia acelei curbe clasice pentru prognoză.

Exemplu:

Un exemplu de studiu a legăturii specifice dintre :

X= *producţia totală de substanţe chimice* din Europa înregistrată între anii 1996 până în 2007

$X=(x_i \ ,i=1..12)$ şi

Y=*gradul de poluare a aerului* prin măsurarea cantităţii de gaze emise $Y=(y_i \ ,i=1..12)$

În Excel :

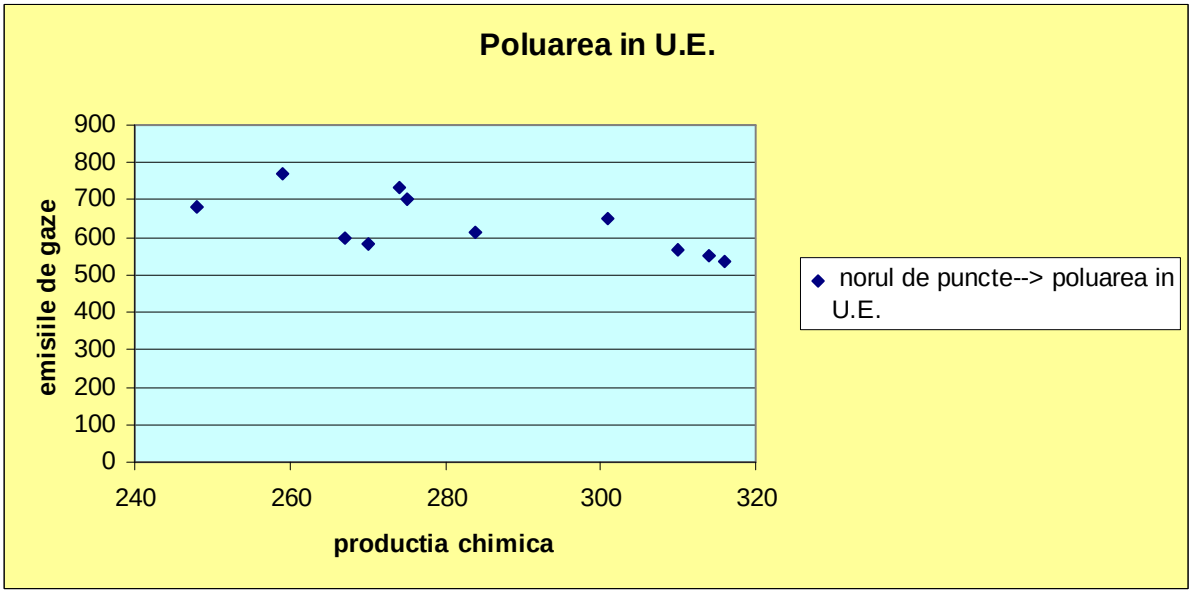
x = (259, 274, 275, 248, 301, 284, 267, 270, 310, 314, 316 ,317) şi

y = (769.75, 732.58, 702.34, 682.04 ,648.38, 614.57, 600.43, 583.63, 565.42, 552.28, 533.93, 521.32)

Datele au fost obţinute de pe site-ul www.eurostat.eu

.

anii	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
productia chimica	**259**	**274**	**275**	**248**	**301**	**284**	**267**	**270**	**310**	**314**	**316**	**317**
volumul emisiilor de gaze	**769.75**	**732.58**	**702.34**	**682.04**	**648.38**	**614.57**	**600.43**	**583.63**	**565.42**	**552.28**	**533.93**	**521.32**



Metoda celor mai mici pătrate.

Dependenţa funcţională a unei variabile *y* faţă de altă variabilă *x* poate fi studiată empiric, pe cale experimentală, efectuîndu-se o serie de măsurători asupra variabilei *y* pentru diferite valori ale lui *x*. Rezultatele se pot prezenta sub formă de tabel sau grafic.

Problema care se pune în acest caz este de a găsi reprezentarea analitică a dependenţei funcţionale căutate, adică de a alege o formulă care să descrie rezultatele experimentului.

Formula se alege dintr-o mulţime de formule de tip determinat, de exemplu.

$y = ax + b$, $y = ax^2 + bx + c$, $y = ae^{bx} + c$, $y = a + h \sin(\omega t + \varphi)$

Cu alte cuvinte, problema constă în a determina parametrii *a*, *b*, *c*, ai formulei, în timp ce tipul formulei este cunoscut dinainte ca urmare a unor considerente teoretice sau după forma prezentării grafice a materialului empiric.

Să notăm, la modul general când avem *n* parametrii, dependenţa funcţională prin:

$y = f(x; a_0 ,a_1 ,...,a_n)$

Parametrii $a_0, a_1,..., a_n$ nu se pot determina exact pe baza valorilor empirice $y_1, y_2,...,y_n$ ale funcţiei, deoarece acestea din urmă conţin erori aleatoare. Este vorba de obţinerea unei estimaţii "suficient de bune".

Formularea problemei

Dacă toate măsurătorile valorilor funcţiei sunt y_1, y_2,...,y_n atunci estimaţiile parametrilor a_0, a_1,..., a_n se determină din condiţia ca suma pătratelor abaterilor valorilor măsurate y_k de la cele calculate $f(x_k; a_0, a_1,..., a_n)$, adică expresia

$$S = \sum_{k=1}^{n} [y_k - f(x_k; a_0, a_1, ..., a_n)]^2$$

să ia valoarea minimă.

Aflarea valorilor parametrilor a_0, a_1,..., a_n, care conduc la cea mai mică valoare a funcţiei

$$s = s(a_0, a_1, ..., a_n)$$

revine la rezolvarea sistemului de ecuaţii

$$\frac{\partial S}{\partial a_0} = 0\,,\ \frac{\partial S}{\partial a_1} = 0\,,\ldots,\ \frac{\partial S}{\partial a_n} = 0$$

Dacă formula empirică depinde liniar de parametrii necunoscuţi atunci sistemul de mai sus va fi de asemenea liniar.

Dreapta de regresie.

În cazul cel mai simplu se studiază numai două variabile X, Y şi se doreşte găsirea dependenţei:

$$Y = aX + b$$

în ipoteza că X este cauza şi Y este efectul.

În urma celor n probe se cunosc datele (x_i, y_i), i=1,..., n şi trebuie să determinăm coeficienţii a şi b astfel încât suma

$$S(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

să fie minimă. Se obţine

$$a = \frac{\bar{c}_{xy}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} \cdot \frac{\bar{c}_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_y}{\sigma_x} \cdot r_{xy} \qquad\qquad b = \bar{y} - a \cdot \bar{x}$$

unde σ_x^2 este dispersia variabilei x, iar σ_y^2 este dispersia variabilei y.

Mărimea

$$\bar{c}_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$$

se numeşte ***covarianţa variabilelor X şi Y***. sau notaţia echivalentă :

$$\operatorname{cov}(x,y) = M(x \cdot y) - M(x) \cdot M(y)$$

$$M(x \cdot y) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N}$$

Pentru a analiza dacă între variabilele X şi Y există o legătură liniară, se calculează ***coeficientul de corelaţie liniară***, dat de formulele echivalente:

$$r_{xy} = \bar{c}_{xy} / \sigma_x \sigma_y$$

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Funcţia Excel corespunzătoare este CORREL.

Sintaxa funcţiei :

CORREL (X,Y)= coeficientul de corelaţie

-parametrii de intrare reprezintă doi vectori de aceeaşi dimensiune care conţin valorile celor două variabile pentru care dorim să calculăm coeficientul de corelaţie.

Interpretare:

Valorile coeficientului de corelaţie sunt în intervalul $[-1,1]$.

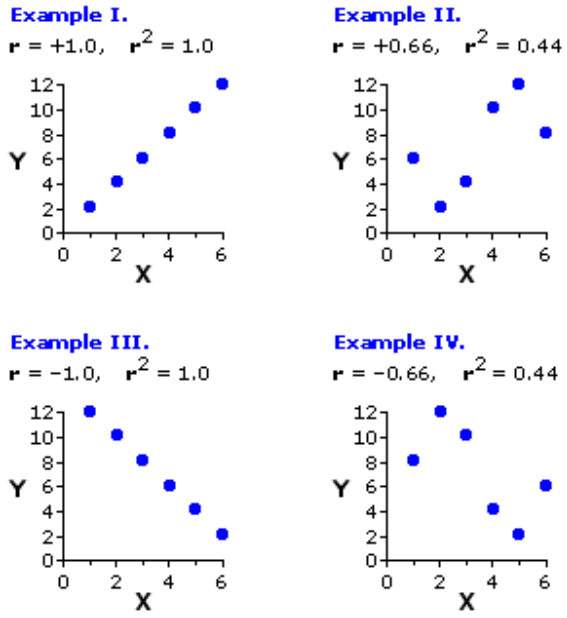
Dacă **r = 0** între cele două variabile **nu există corelaţie**.

Dacă **r = 1**, corelaţia între cele două variabile este **maximă şi directă**.

Dacă **r = -1**, corelaţia între cele două variabile este **maximă şi inversă**.

Cu cât avem o valoare mai apropiată de 1 sau -1 cu atât corelaţia e mai puternică (directă pentru valori pozitive şi inversă pentru valori negative), cu cât avem o valoare apropiată de 0 corelaţia este mai slabă.

În ambele cazuri (r=-1, sau 1) , sintagma "*tendinţa de a fi asociat*" este un alt mod de a spune că variabilitatea lui X tinde să fie asociată cu variabilitate în Y, şi vice-versa, sau, pe scurt, că *X şi Y au tendinţa de a varia împreună.*

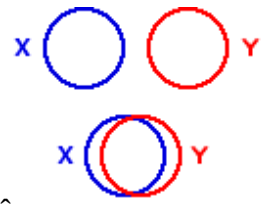


Coeficient de determinare

r ^{2} = ***coeficient de determinare***

Dacă reprezentăm variaţia lui X ,respectiv Y prin suprafaţa a două cercuri (suprafaţa cercului reprezinta 100% din variaţia fie a lui X sau Y). În caz de corelaţie zero nu există nici o tendinţă pentru X şi Y pentru a co-varia; şi, prin urmare, după cum este ilustrat ,avem două cercuri separate în partea de sus;

corelaţia nu este zero,atunci cele două cercuri se suprapun(o parte din variabilitatea lui X este explicată de variabilitatea lui Y). şi anume **r** ^{2} (care reprezintă procentul din variaţia lui y determinată de variaţia lui x)



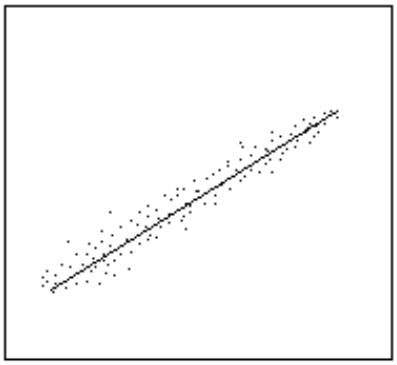
În acest caz cele două cercuri se suprapun (**r** = - 0.86 **r** ^{2} =0.74) Zona de suprapunere reprezintă 74% din variaţia lui Y (Yeste cuplat în variabilitate cuX, şi invers), precum şi faptul că 26% din variaţia lui Y este fără nici o legătură cu variabilitatea lui X, la fel ca din variaţia de X este fără nici o legătură cu variabilitate în Y.(zonele de ne-suprapunere reprezintă 26%) Aceasta zonă de ne-suprapunere sau porţiune de dezacord, între X şiY se numeşte **variaţie reziduală** = 1- **r** ^{2} .

Dar faptul că relaţiile de cauzalitate între variabile pot produce corelaţii nu implică faptul că o relaţie de cauzalitate se află în spatele fiecărui exemplu de corespondenţă.

În final se va obţine ecuaţia de regresie:

$$Y - \bar{y} = \frac{\sigma_y}{\sigma_x} r_{xy} (X - \bar{x})$$

Această dependenţă reprezintă o dreaptă numită dreaptă de regresie a variabilei Y în raport cu variabila X.



Dreapta de regresie

Observaţie. Se poate vorbi şi de dependenţa variabilei X în funcţie de Y. Urmând un calcul asemănător se ajunge la dreapta de regresie a variabilei X în raport cu Y:

$$X - \bar{x} = \frac{\sigma_x}{\sigma_y} r_{xy} (Y - \bar{y})$$

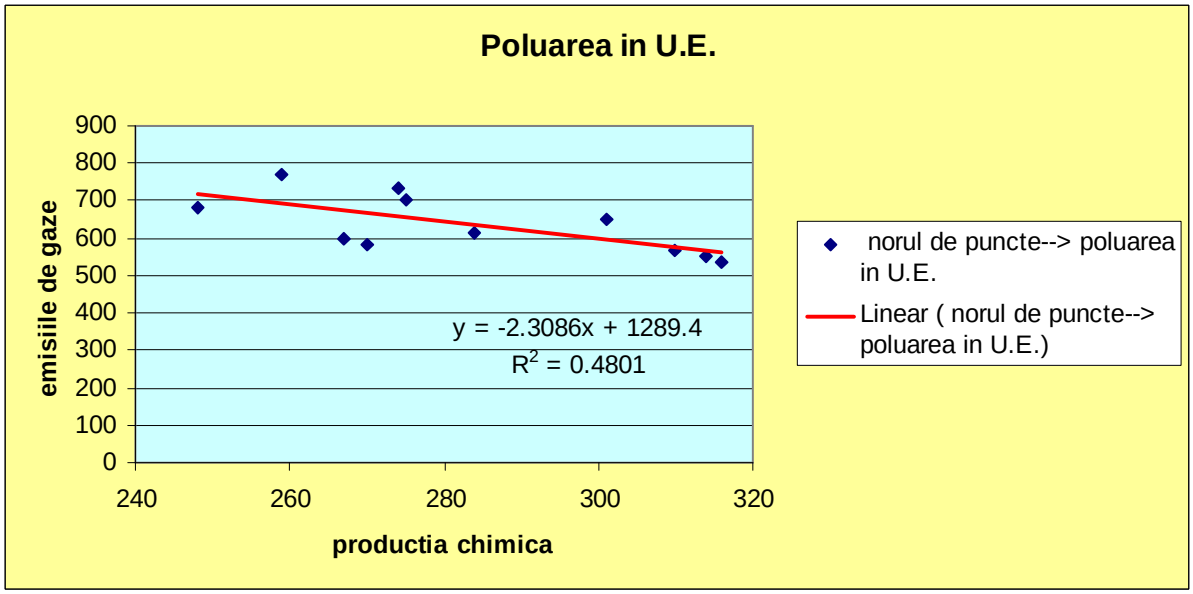
Se observă că cele două drepte de regresie coincid dacă şi numai dacă $r_{xy}^2 = 1$.

Observaţii.

1. Trebuie să facem observaţia că, indiferent de gradul de împrăştiere al punctelor, întotdeauna se poate găsi o dreaptă de regresie, dar în cazul unei dispersii mari aceasta devine inutilă. De aceea un studiu preliminar a distribuţiei punctelor în plan sau spaţiu se impune cu necesitate.

2. Coeficientul de corelaţie este o mărime foarte importantă în cadrul regresiei liniare. El măsoară gradul de dependenţă liniară între cauză şi efect şi are o valoare cuprinsă între –1 şi 1. Apropierea de 1 implică o dependenţă liniară puternică între mărimi, iar apropierea de zero indică o lipsă a corelatiei. Valorile negative semnifică o corelaţie inversă.

EXEMPLE


Poluarea in U.E.
emisiile de gaze
900
800
700
600
500
400
300
200
100
0
240
260
280
300
320
productia chimica
y = -2.3086x + 1289.4
R2 = 0.4801
norul de puncte--> poluarea in U.E.
Linear (norul de puncte--> poluarea in U.E.)