Partea a III-a

STATISTICĂ DESCRIPTIVĂ BIVARIATĂ

Pentru început să răspundem şi în cazul **statisticii descriptive bivariate**, la aceleaşi întrebări care au fost puse asupra statisticii descriptive univariate.

Ce face ?

Studiază simultan două variabile pentru:

(1a) *clasarea* şi *măsurarea* unei dependenţe argumentate biologic,

ceea ce permite

(1b) **predicţia** comportării unei variabile în raport cu cealaltă variabilă;

(2) **demonstrarea independenţei** a două variabile.

Terminologie:

Dependența între *variabile cantitative* (cantitative sau ordinale) se numeşte **corelație**, iar cea între *variabile calitative* se numeşte **asociere**.

Cum face ?

În mod analog statisticii descriptive univariate, cea bivariată realizează o sinteză grafică, respectiv numerică a datelor.

Sinteza grafică se efectuează, de asemenea, în cei doi paşi cunoscuți, dar prin instrumente specifice statisticii bivariate:

- ♦ gruparea datelor în *tabele statistice cu dublă intrare*, care
  - pentru *variabile cantitative* se numesc **tabele de corelație**, iar
  - pentru *variabile calitative* se numesc **tabele de contingență**;
- ♦ reprezentări grafice
  - pentru *variabile cantitative*:
  - *diagrame de împrăştiere cu puncte* ori *cu areale*
  - *diagrame în batoane în spațiu* (*stereograme*) sau *histograme în spațiu* (*stereohistograme*)
  - pentru *variabile calitative – reprezentare prin areale în dreptunghiuri.*

Sinteza numerică se face în parametri specifici analizei bivariate şi anume parametrii de *corelație*, respectiv *asociere*. În cazul corelațiilor între dimensiuni se obține chiar sinteza datelor în ecuații. Astfel, în cazul unei dependențe argumentate biologic, se va putea obține predicția comportării unei variabile în funcție de *valorile*, respectiv *variantele* celeilalte variabile.

Capitolul 4
TRATAREA SIMULTANĂ A DOUĂ DIMENSIUNI

Deoarece cele două variabile care vor fi tratate simultan în cadrul acestui capitol sunt dimensiuni putem să folosim adjectivul "bidimensional" în locul termenului de "bivariat".

§ 4.1. Sinteza grafică bidimensională

Datele experimentale bidimensionale sunt serii statistice de perechi de valori (x, y).

Exemplul 4.1.

Fie şirul bidimensional:

x: 1 1 1 2 2 2 2 3 3 3
y: 2 4 5 6 6 8 8 5 8 8

Menționăm că acesta poate fi dat şi sub forma →

x	y
1	2
1	4
1	5
2	6
2	6
2	8
2	8
3	5
3	8
3	8

Tabelul cu dublă intrare corespunzător este figurat în continuare.

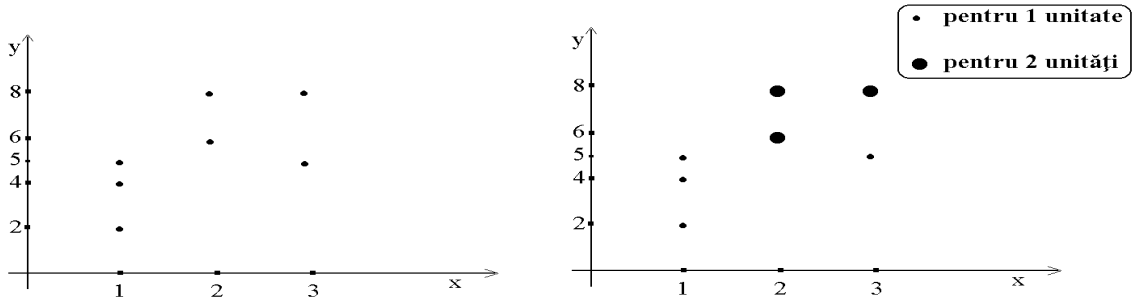
4.1.1. Tabel statistic cu dublă intrare

x =	**1**	**2**	**3**
y :			
8	0	2	2
6	0	2	0
5	1	0	1
4	1	0	0
2	1	0	0

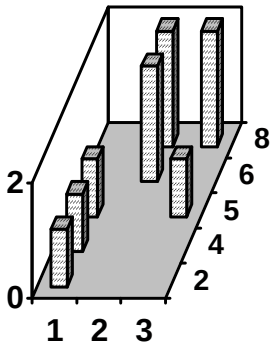
Aşa cum am menționat deja, în cazul variabilelor cantitative şi oridnale acest tabel se numeşte *tabel de corelație* (deoarece foloseşte la studiul corelației).

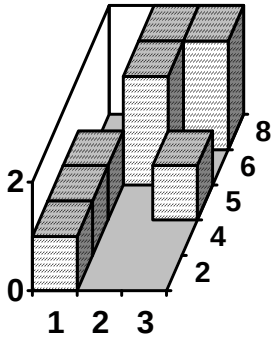
Tabelul poate fi reprezentat grafic în următoarele patru moduri.

4.1.2. Diagramă de împrăştiere cu puncte, respectiv cu areale



4.1.3. Diagramă în batoane în spațiu (stereogramă) şi stereohistogramă



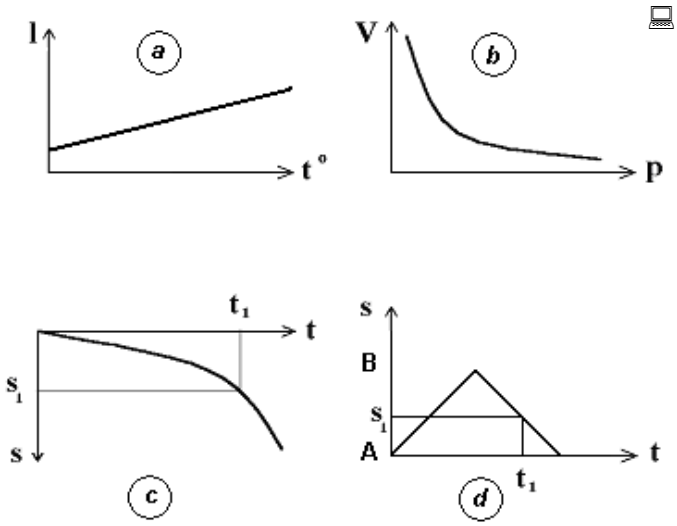


§ 4.2. Dependență funcțională şi dependență statistică

În ştiințele exacte, cum ar fi fizica, astronomia, chimia, apar fenomene descriptibile printr-o **dependență funcțională** de forma $y = f(x)$, adică o relație **univocă** de la x către y: unei valori fixate a lui x îi corespunde o singură valoare a lui y.

4.2.1. Exemple de dependențe funcționale

a. Dilatarea în lungime (l) a unei bare metalice în funcție de temperatură ($t°$), fenomen descris de ecuația
$$l = k \cdot t° + k_0,$$
b. legea Boyle-Mariotte
$$p \cdot V = ct,$$
c. valoarea spațiului parcurs în căderea liberă
$$s = -\frac{1}{2} \cdot g \cdot t^2,$$
d. valoarea spatiului parcurs de un mobil in miscare uniforma de la A la B si inapoi,



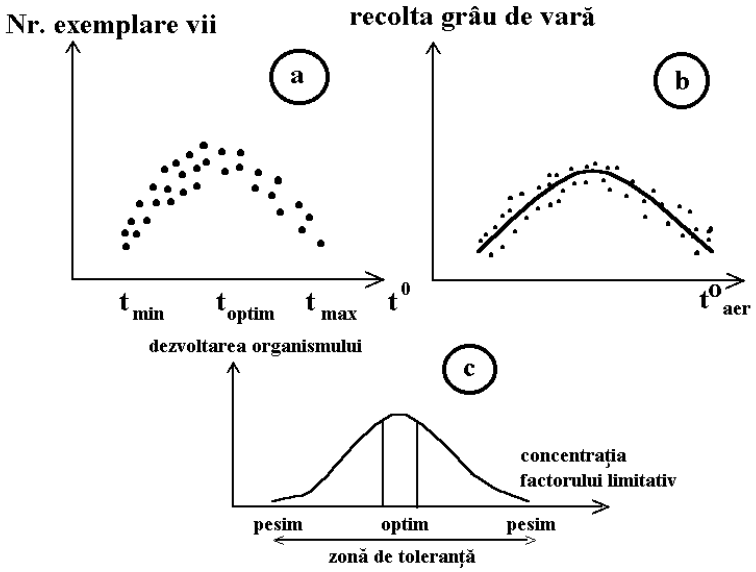
sunt exemple de dependențe funcționale între două variabile cantitative.

În toate aceste situații, cunoscând valoarea unei variabile considerate independentă (şi, de regulă, reprezentată pe axa orizontală), putem determina în mod **univoc** valoarea celeilalte variabile, considerată dependentă de prima variabilă. De exemplu, în cazul căderii libere, la momentul de timp t_1 putem prevedea cu exactitate spațiul parcurs până în acel moment: $s_1 = -\frac{1}{2} \cdot g \cdot t_1^2$ (vezi fig. *c* de mai sus). Prin urmare, dacă figurăm într-un plan perechi de date experimentale pentru asemenea fenomene, punctele se vor plasa pe liniile teoretice respective.

4.2.2. Exemple de dependențe statistice

În ştiințele vieții şi în alte domenii care studiază fenomene complexe, situația de mai sus nu se mai regăseşte decât ca tendință. Acest lucru se întâmplă din cauza variabilității induse de complexitatea fenomenelor respective.

Cateva loturi de câte 100 de indivizi cât mai asemănători din aceeaşi specie, sunt menținute – acelasi interval de timp - la o anumita temperatura constanta. Alte cateva loturi la o alta temperatura constanta si tot asa mai departe, la alte valori constante de temperatura. În final vom consemna pe un grafic perechile de valori (temperatură, număr de indivizi în viață). Se va obține un nor de puncte de forma din figura *a*.



Reprezentarea exprimă ceea ce se poate susține cu argumente biologice. Şi anume faptul că există o anumită temperatură optimă la care rămân în viață cei mai mulți indivizi şi două temperaturi limită, una minimă şi alta maximă, sub care, respectiv peste care nu mai supraviețuieşte nici un individ. Totodată, pe măsură ce temperatura se apropie de valorile limită, efectivele celor care supraviețuiesc scad.

Un alt exemplu pentru această situație este recolta grâului de vară (pe ordonată) în funcție de temperatura aerului (pe abscisă, vezi fig. b de mai sus).

Ambele exemple sunt ilustrări a ceea ce în ecologie este denumit *legea toleranței* [5].

Legea toleranței afirmă că *reacția teoretică a viului la variația unui factor limitativ al mediului urmează o curbă în formă de clopot*[19]*, denumită* **curbă de toleranță**.

În figura *c* de mai sus este desenată o asemenea curbă. Se observă că reacția viului are loc într-o zonă de toleranță (de unde şi denumirile legii şi curbei) mărginită de două zone de **pessim** (una de *minim*, alta de *maxim*, ale intensității factorului respectiv) şi conținând un punct sau o zonă de **optim**.

[19] Putem considera că este o distribuție cvasinormală.

Evident, dependența între cele două variabile nu mai este de tip funcțional, neputându-se prevedea în mod univoc reacția viului la o anumită valoare a factorului de mediu. Cu toate acestea, o prognoză se poate face pe baza tendinței degajate din norul de puncte, tendință care are o formă de dependență funcțională. Aceasta este o **dependență statistică** sau **stocastică** (**stohastică**).

În cazul acestui tip de dependență, dacă urmărim un experiment sau facem o observație şi figurăm perechile de valori corespondente ale celor două variabile într-o *diagramă de împrăştiere*, punctele nu se vor mai plasa în întregime pe o anumită linie (dreaptă sau curbă), ci vor forma un nor în jurul unei anumite linii de dependență.

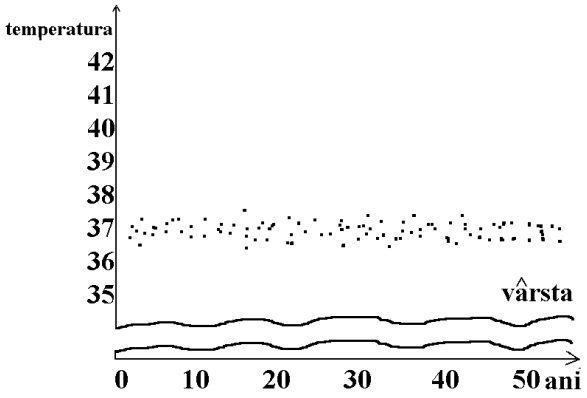
Reformulând sintetic

o **dependență statistică între două variabile** înseamnă un nor de puncte în diagrama de împrăștiere în "spatele" căruia se conturează o *linie de dependență.*

Prin **linie de dependență** înțelegem *orice linie diferită de dreptele paralele la axe.*

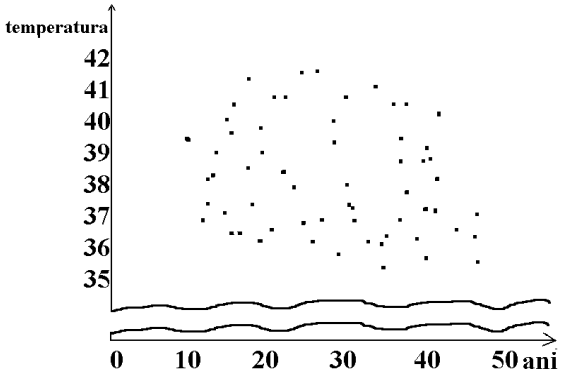
4.2.3. Independență funcțională şi independență statistică totală, respectiv reală

Dreptele paralele cu axele de coordonate exprimă faptul că valoarea unei variabile rămâne constantă indiferent de valorile celeilalte variabile, adică nu depinde de variația celeilalte variabile. De exemplu, dacă reprezentăm pe abscisă vârsta, iar pe ordonată temperatura şi figurăm mai multe puncte corespunzătoare unor indivizi sănătoşi, obținem diagrama alăturată [29].



În spatele acestui nor de puncte se conturează o linie paralelă cu axa orizontală, ceea ce exprimă faptul intuitiv că, dacă am cunoaşte vârsta unui individ, nu aflăm nimic în plus în legătură cu temperatura, față de ceea ce ştiam şi înainte şi anume că aceasta este situată undeva în apropierea temperaturii de $37^{o}C$. Această situație ilustrează, după unii autori [29], **independența totală**. Este vorba, bineînțeles, de **independența statistică totală**, norul de puncte înconjurând o dreaptă paralelă cu una din axe. Dacă punctele se plasează chiar pe o dreaptă paralelă la una din axele de coordonate vorbim despre **independența funcțională (totală)**.

Situația din figura de mai sus este însă mai rar întâlnită în practică deoarece, de regulă, variabilitatea biologică produce, la extremele unei caracteristici, indivizi mai puțini, astfel încât, cel mai adesea, independența a două caracteristici apare în nori de puncte sub formă relativ circulară.



În această figură preluată din [29], punctele reprezintă vârsta, respectiv temperatura unui bolnav. Dacă scările de reprezentare ale celor două variabile se modifică, norul va deveni eliptic cu axele paralele cu axele de reprezentare. Din nor nu se degajă, deci, nici o linie de dependență.

Evident, în condițiile de mai sus, norul este circular (sau eliptic) şi nu pătrat (sau dreptunghic) din cauza rarității vârstelor extreme cu temperaturi extreme şi a abundenței celor cu vârste medii, dar cu temperaturi extreme, aceşti indivizi fiind mai rezistenți. Nici în acest caz, cunoaşterea vârstei nu ne ajută la o predicție a temperaturii. Această situație ar putea fi denumită **independență statistică reală**.

- ✓ Între independența statistică (totală sau reală - vezi ultimele două figuri) şi dependența statistică totală (care este tot una cu dependența funcțională după o linie de dependență în sensul de mai sus), se plasează **dependența statistică** (cum ar fi de exemplu, *legea toleranței*).
- ✓ În cazul unei dependențe statistice este esențial, din punct de vedere intuitiv, ca din norul de puncte să se "străvadă" în mod cât mai puțin echivoc un anumit tip de linie de dependență. În acest caz, putem considera că fenomenul se desfăşoară ca tendință conform liniei, dar variabilitatea biologică induce pentru variabila dependentă, variații în jurul acestei linii astfel încât ordinul lor de mărime pentru o valoare fixată a variabilei independente, este mult mai mic decât variația globală a variabilei dependente [29]. Numai astfel, de fapt, poate fi intuită o linie de dependență şi doar aşa se poate ajunge la o predicție.

  Exemplu:

  În cazul recoltelor de grâu în funcție de temperatură, este evident că la o temperatură apropiată de optim ne putem aştepta la o recoltă bogată tocmai datorită variațiilor mici ale recoltei în jurul unei valori mari, în comparație cu variația globală a recoltei indiferent de temperatură, variație care este cu mult mai mare.
- ✓ Dacă reuşim să evidențiem linia respectivă de dependență şi să identificăm în întregime expresia ei analitică, atunci putem realiza *cel de-al treilea deziderat al statisticii descriptive, şi anume*: **sinteza datelor în ecuații**, pe lângă celelalte două deziderate (*sinteza datelor în grafice* şi *sinteza datelor în numere*).

4.2.4. Postulatul conexiunii între planul fenomenologic şi cel al datelor observate

1° Postulatul

Este extrem de important să percepem diferențiat *planul fenomenologic* de *planul datelor experimentale sau de observație*. Totodată, trebuie reținut că între cele două plane există o conexiune într-un singur sens şi anume:

> "O legătură în plan fenomenologic implică manifestarea unei dependențe (funcționale ori statistice) între variabile în planul datelor, altfel spus a unei corelații între variabile."

Acest **postulat epistemologic**[20] este obligatoriu dacă acceptăm posibilitatea cunoaşterii ştiințifice.

Reciproca acestei implicații şi anume:

> "O corelație între variabile implică existența unei legături în plan fenomenologic",

funcționează **doar ca o posibilitate, nu în mod necesar**.

O dovada este dată de **corelațiile fără sens**. Într-adevăr este posibil să observăm în datele dintr-un experiment, o corelație între intensitatea razelor Lunii şi lungimea unor papuci de gumă. Aceasta este o *corelație fără sens* deoarece este evident pentru oricine că este absurd să vorbim apoi despre "influența razelor de Lună asupra papucilor de gumă".

De aceea**, în cazul unei legături în plan fenomenologic**,

- **biologia** va trebui să vină cu argumente DE SPECIALITATE, LOGICE, DE PRINCIPIU ETC. (în general NESTATISTICE) pentru susținerea acesteia, iar
- **statistica** va ajuta doar la (1) alegerea **formei corelației** şi la (2) măsurarea **intensității acesteia**. Astfel vor fi posibile (3) **predicții** (4) **cu grad de aproximare controlat** prin intensitatea corelației.

Altfel spus, sarcina biologului este să susțină corelația descoperită de biostatistician în planul datelor cu argumente din planul fenomenologic.

✓ Lucrările aplicative despre corelații care se limitează la partea statistică pot fi suspectate că susțin false corelații.

2° Forma echivalentă a postulatului

Formând contrara reciprocei[21] pentru *postulat*, obținem propoziția echivalentă care este, de asemenea, adevărată:

> "Inexistența unei dependențe între variabile (sesizabilă în diagrama de împrăştiere sau prin tabelul de corelație), adică independența (funcțională sau statistică) implică lipsa unei legături în plan fenomenologic."

Ceea ce înseamnă că:

[20] **Epistemologia** este teoria cunoaşterii ştiințifice.

[21] (care scrisă formal este *non b* $\Rightarrow$ *non a* şi care este echivalentă logic cu directa $a \Rightarrow b$)

prin demers statistic exhaustiv putem să dovedim cu certitudine, doar lipsa oricărei legături în plan fenomenologic (reflectată prin datele respective).

- ✓ Pentru diferențierea tranșantă a celor două plane (cel fenomenologic și cel al datelor) precum și a ceea ce ține de epistemologie am introdus termeni diferiți pentru aceeași categorie generală. Astfel în plan fenomenologic vorbim de *legături* ori *lipsa oricăror legături* iar în planul datelor folosim termenii de *dependențe* sau *corelații* ori *independență* sau *lipsa oricărei corelații*, și putem adăuga, între variabile. Pentru a atrage atenția asupra faptului că reflexia epistemologică este "deasupra" celor două plane am utilizat un al treilea termen pentru "legătura" sau "dependența" dintre cele două plane – cel de *conexiune*.
- ✓ O corelație între două variabile *x* și *y* (în planul datelor) poate însemna pentru aspectele $\mathcal{X}$ și $\mathcal{Y}$ din planul fenomenologic că:
  1. $\mathcal{X}$ este cauza lui $\mathcal{Y}$,
  2. $\mathcal{Y}$ este cauza lui $\mathcal{X}$,
  3. ambele sunt efectele unei a treia cauze,
  4. ambele variază concomitent cu un al treilea factor,
     de exemplu, "evoluția paralelă cu vârsta a două caractere biologice, ceea ce de multe ori creează aparența unei legături între ele" [29].
  5. $\mathcal{X}$ și $\mathcal{Y}$ sunt puse în legătură fără sens.

Situația a 4-a se tratează cu ajutorul unui coeficient de corelație special, denumit *coeficient de corelație parțială* [29], iar situația a 5-a reprezintă cazul corelației fără sens.

În general, problema corelației este extrem de complexă deoarece, în realitate, dependențele operează între mai mult de două sau trei variabile, adică este nevoie de statistică multidimensională. De aceea, problema corelației nu poate fi tratată corect fără coordonarea unui biostatistician.

4.2.5. Aplicarea postulatului

1° Alegerea formei corelației, determinarea liniei corespunzătoare și măsurarea intensității corelației de forma respectivă

În situația în care biologul are argumente asupra existenței unei legături în plan fenomenologic sau cel puțin o postulează, statistica îi pune la dispoziție următoarele etape metodologice:

1. alegerea unui anumit tip de corelație teoretică, adică a unei anumite forme de dependență;
2. determinarea parametrilor liniei respective;
3. controlul validității modelului ales.

Cele trei etape sunt denumite, de către diverși autori, mai mult sau mai puțin diferit:

1. *identificare* sau *modelare*;	1. alegerea liniei (dreaptă sau curbă);
2. *ajustare*;	2. ajustarea sau adaptarea unei linii la datele experimentale;
3. *validare* [10];	3. controlul adaptării tipului de linie ales la datele experimentale [29].

1. În legătură cu **prima etapă** (*identificarea* sau *modelarea*) literatura statistică ignoră, de regulă, ceea ce este cel mai important pentru biologi şi anume modurile în care se poate alege o formă de corelaţie. În opinia noastră există următoarele trei moduri sau "filozofii" de alegere a formei corelaţiei:

I. forma este *determinată* de considerente de principiu şi / sau specialitate;
II. forma este *observată* repetându-se pe multe seturi de date similare (proprii sau din literatură);
III. forma este *o aproximaţie convenabilă* pe setul de date respective şi, eventual, alte câteva seturi similare.

În consecinţă sintagma "*alegerea* unei forme de dependenţă" trebuie înlocuită cu "*determinarea, observarea sau aproximarea convenabilă* a unei forme de dependenţă".

✓ Primul mod de alegere este cel mai profund şi mai greu de aplicat. Solicită apelul la anumiţi biomatematicieni sau biofizicieni. Ultimul mod necesită colaborarea cu biostatisticieni. În sfârşit, modul al II-lea este cel mai des întâlnit, căci este practicat de biologi, de regulă, fără asistenţă biometrică, ceea ce conduce deseori la difuzarea în masă a anumitor "imperfecţiuni statistice", ca să ne exprimăm eufemistic.

Exemplele 4.2.5.

I. (a) *Corelaţie după o curbă normală trunchiată* (decupată) *la ambele capete* - pentru modul în care reacţionează efectivele populaţiilor unei anumite specii la gradientul unui anumit factor de mediu. Considerentele ecologice sunt subînţelese în succinta prezentare din 4.2.2.

(b) *Corelaţie după o curbă putere* - pentru corelaţia între înălţime şi greutate la indivizi maturi din aceeaşi specie. Explicaţia de principiu este dată la punctul 4° de la 3.1.2.

II. *O corelaţie liniară* - între înălţimile fiilor la maturitate şi înălţimile taţilor lor.

III. (a) *O corelaţie parabolică de grad* 2 - pentru datele de la I*a*.

(b) *O corelaţie liniară* - pentru norul de la punctul I*b*.

Notă: Curbele amintite aici, în afară de curba normală, sunt prezentate în 4.3.1.

Pentru realizarea primei etape (*identificarea sau modelarea*) statistica pune la dispoziţie *diagrama de împrăştiere*. Altfel spus, orice demers de analiză de corelaţie trebuie să pornească prin construirea diagramei de împrăştiere. Apoi, se va utiliza diagrama respectivă în funcţie de modul de alegere a formei. Adică, dacă forma este dictată de considerente teoretice (primul mod de alegere) ori empirice (al doilea mod) vom verifica vizual cum se "profilează în spatele" norului de puncte forma respectivă. Dacă suntem în situaţia a III-a, specifică noilor cercetări, atunci, cu ajutorul biostatisticianului se va găsi forma care va aproxima cât mai convenabil noul nor de date.

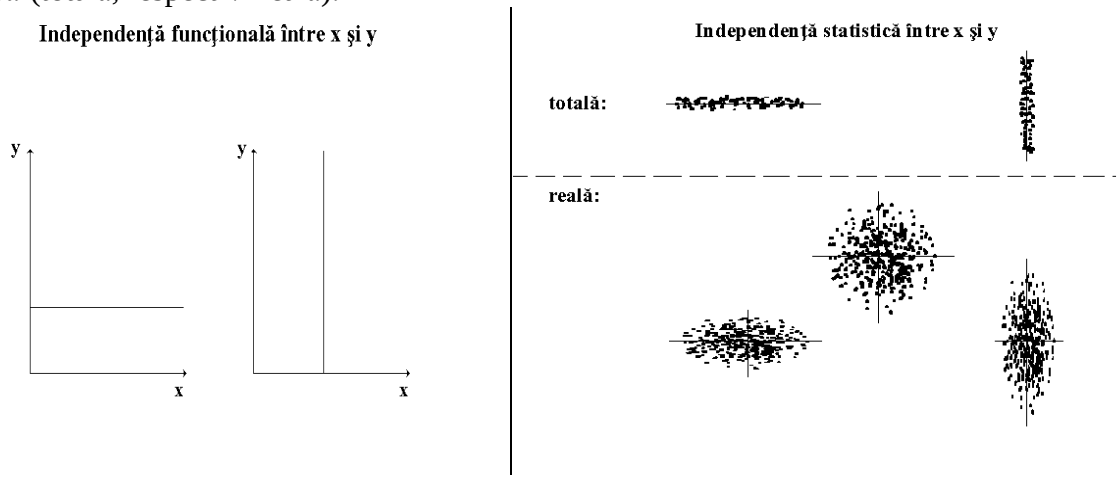
În continuare vom presupune că suntem în situaţia a III-a în care căutăm o formă care să fie o aproximaţie convenabilă a norului respectiv de puncte.

2. Pentru **etapa a doua** (*ajustarea*) statistica are un întreg arsenal de metode de calcul a liniei de forma specificată în etapa anterioară, care se potriveşte la datele experimentale cel mai bine în raport cu un anumit criteriu. Linia respectivă se numeşte *linie* (*dreaptă* sau *curbă*) *de regresie*.

3. În fine, pentru **etapa a treia** (*validarea*) există indicatori specifici fiecărei forme prin care se poate măsura gradul de corelaţie după forma respectivă. Aceştia sunt denumiţi *indici de corelaţie*, cu excepţia dreptei de regresie pentru care se utilizează *coeficientul de corelaţie liniară*.

2° Dovedirea independenței

În situația contrară, în care nu există nici-o legătură în plan fenomenologic, rolul biologului se reduce la zero deoarece dovedirea acestui lucru este rezolvată în întregime de statistică. Astfel, dacă reprezentând populația statistică bidimensională într-o diagramă de împrăştiere obținem una din următoarele forme de nori, putem susține independența (lipsa oricărei legături) în fenomen. În figura din stânga sunt ilustrate cele două cazuri tip de *independență funcțională* (totală), iar în cea din dreapta cele cinci cazuri tip, de *independență statistică* (totală, respectiv reală).



Deci **diagramele de împrăştiere** sunt un instrument foarte puternic în acest caz. Singure, fără nici-un alt instrument şi fără argumente de specialitate, **pot <u>dovedi</u> independența în fenomen**.

✓ Acest mod de utilizare a diagramelor de împrăştiere este foarte rar utilizat în practică din două motive: (a) interesul nostru este focalizat pe legăturile fenomenologice şi nu pe lipsa acestora şi (b) pentru nici-un fenomen nu dispunem de toate datele ca să putem afirma că lucrăm cu populația statistică corespunzătoare. Cu toate acestea, suntem nevoiți, la un moment dat, să considerăm populație statistică un eşantion apreciat ca reprezentativ. Altfel spus, ne asumăm răspunderea să considerăm că datele obținute până în acel moment caracterizează în întregime fenomenul. Aceasta este o limită a metodologiei ştiinței ce nu poate fi depăşită. Prin urmare concluziile noastre astfel obținute sunt perfect valabile din punct de vedere statistic, dar pot fi eronate din cauza acestei limite metodologice.

§ 4.3. Sinteza numerică bidimensională

4.3.1. Forme de corelație

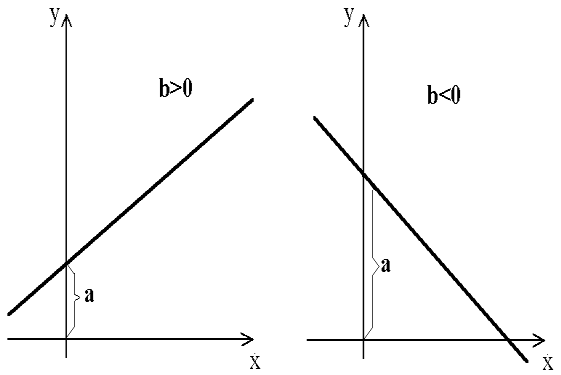
În mod teoretic pot exista o infinitate de forme de corelație. Practic însă apelăm doar la câteva dintre acestea şi anume la cele mai "frumoase", adică *netede* (derivabile) şi cu o *expresie analitică simplă*, căci o **axiomă epistemologică nemărturisită** este aceea că:

"Legile naturii sunt «simple» ".

În statistică în general şi în biostatistică sau ecostatistică în special, pentru exprimarea corelației sunt utilizate frecvent mai multe forme de dependență. Pentru nivelul de începător este suficient să amintim însă numai următoarele funcții $f(x)$. Acestea sunt prezentate în continuare prin ecuațiile graficelor lor, $Y = f(x)$, precum şi prin graficele corespunzătoare:

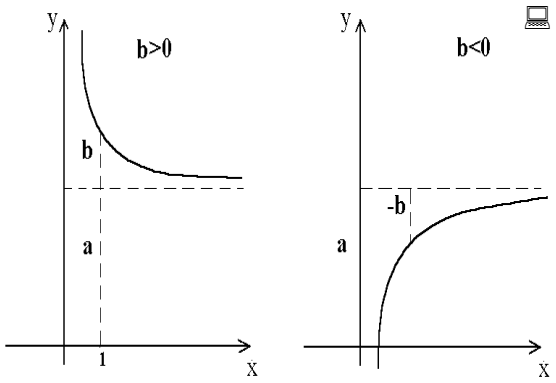
$Y = a + b \cdot x,\ b \neq 0.$

Linia dreaptă neparalelă cu axa *OX* sau *OY*.



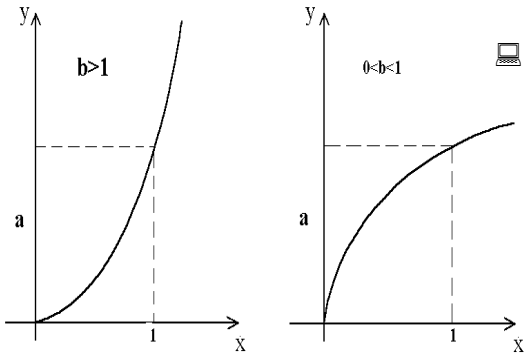
$Y = a + b / x,\ a > 0,\ b \neq 0.$

Hiperbola.



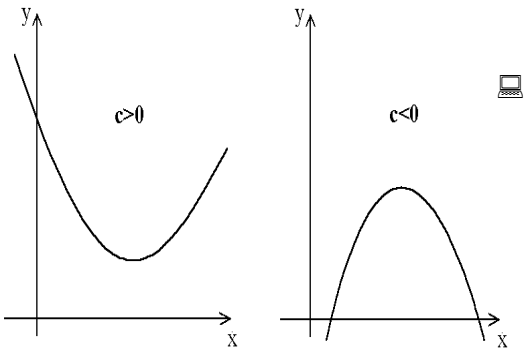
$Y = a \cdot x^{b},\ a, b > 0.$

Funcția putere.



$Y = a + b \cdot x + c \cdot x^{2},\ c \neq 0.$

Parabola de gradul 2.



1° Unele proprietăți ale primelor două forme enumerate

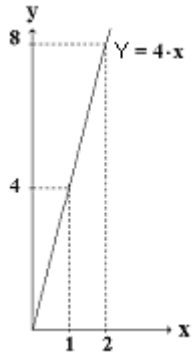
Linia dreaptă sub forma particulară $Y = b \cdot x$, cu $b > 0$ este ascendentă, trece prin origine şi exprimă ideea *dependenței* **direct proporționale**. Aceasta înseamnă, etimologic vorbind, că:

"la o *creştere* a variabilei *x*, de un anumit număr de ori corespunde o *creştere* a variabilei *y*, de acelaşi număr de ori".

Cu alte cuvinte această formă de dependență exprimă ideea de "**efect direct proporțional cu efortul**". (Se înțelege că efectul se manifestă în variabila dependentă *y*, iar efortul în variabila independentă *x*.)

Exemplul 4.3.1.

În desenul alăturat este figurată dreapta de ecuație $Y = 4 \cdot x$. Se observă că dacă $x = 1$ se dublează, devenind 2, atunci şi imaginea lui 1, adică 4, se dublează devenind 8.

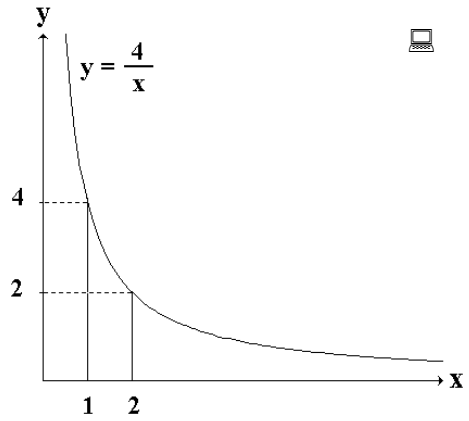


Hiperbola sub forma particulară $Y = \frac{b}{x}$, cu $b > 0$ exprimă ideea *dependenței* **invers proporționale**. Aceasta înseamnă, de asemenea etimologic vorbind, că:

"la o *creştere* a variabilei *x*, de un anumit număr de ori corespunde o *descreştere* a variabilei *y*, de acelaşi număr de ori".

Exemplul 4.3.1'.

În desenul alăturat este figurată hiperbola de ecuație $Y = \frac{4}{x}$. Se observă că dacă $x = 1$ se dublează, devenind 2, atunci imaginea lui 1, adică 4, se înjumătățește devenind 2.



2° Clasificare de lucru a formelor de corelație

Linia dreaptă joacă, în statistica bidimensională, un rol analog distribuției normale în statistica unidimensională.

Amintim că în statistica unidimensională vorbeam de transformări de simetrizare care aplicate anumitor distribuții produc distribuții simetrice cvasinormale. Făceam acest lucru pentru că distribuția normală este cea mai bogată în proprietăți. Prin urmare, distribuțiile pot fi clasificate în: distribuții normale şi alte distribuții, cele din urmă putând fi împărțite în distribuții simetrizabile (de regulă, cele unimodale) şi nesimetrizabile.

În mod analog, formele de corelație pot fi clasificate astfel:

- **corelația liniară;**
- corelații neliniare:
  - **liniarizabile**,
  - neliniarizabile.

Dintre formele enumerate mai sus, *hiperbola* şi *funcția putere* sunt *liniarizabile*. De exemplu hiperbola, în forma sa generală $Y = a + b / x$, $a > 0$, $b \neq 0$, devine, făcând notația (substituția) $X = \frac{1}{x}$, funcția liniară $Y = a + b \cdot X$.

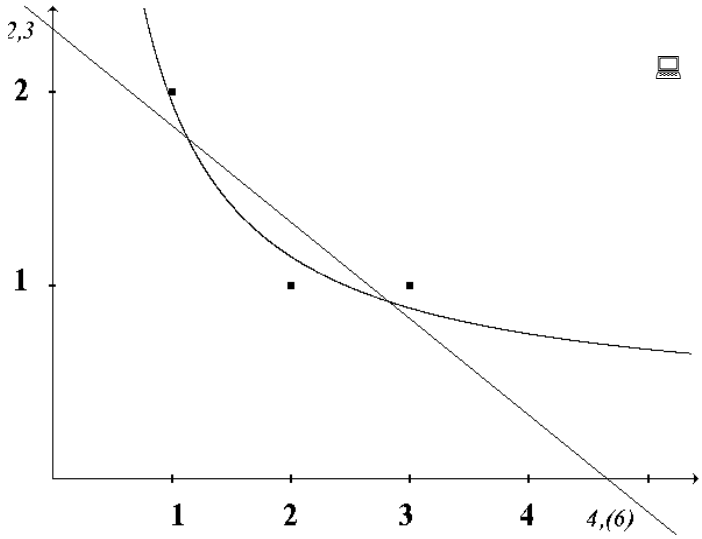
✓ *Funcția putere* precum şi alte funcții, cum ar fi funcția exponențială şi cea logaritmică sunt liniarizabile prin logaritmări [12]. Nu le vom trata însă în acest cadru "de începător". Pentru lucrul cu acestea se recomandă colaborarea cu un biostatistician.

4.3.2. Ajustarea unei anumite forme la datele experimentale

Din punct de vedere teoretic orice formă de corelație poate fi ajustată oricărui nor de puncte. Prin **ajustarea** unei anumite forme înțelegem faptul, că o dată fixată forma de corelație, o singură linie (dreaptă sau curbă) de tipul respectiv se va "potrivi" cel mai bine datelor, în raport cu un anumit *model* şi un anumit *criteriu* dinainte stabilite. Vom vedea la punctul 1° de ce linia respectivă se numeşte, în mod impropriu, **linie** (**dreaptă** sau **curbă**) **de regresie**.

Exemplul 4.3.2.

De exemplu, dacă am stabilit primul *model* şi *criteriu* care vor fi prezentate mai jos şi dacă dorim o *ajustare liniară*, o singură dreaptă va fi cea mai bună ajustare liniară la datele respective. La fel, dacă am ales forma de corelaţie hiperbolică, o singură hiperbolă va fi cea mai bună *ajustare hiperbolică* (vezi desenul alăturat care conţine, pentru norul de trei puncte dat, *dreapta de regresie* şi *hiperbola de regresie*, după un anumit model şi criteriu).



A rămas de precizat ce înseamnă *model de regresie* şi ce înseamnă "cea mai bună" ajustare de forma respectivă, adică de precizat *criteriul*.

Statistica pune la dispoziţia utilizatorilor mai multe modele şi mai multe criterii. Modelul cel mai des folosit este următorul.

1° Modelul 1 - regresia lui *y* în *x* - şi două criterii

Se presupune că variabila *x* este la îndemâna experimentatorului, iar variabila *y* depinde de *x* după o funcţie cu grafic $Y = f(x)$ de o anumită formă, la care se adaugă o variaţie aleatoare ε de medie zero şi dispersie finită şi relativ mică: $y = (Y + \varepsilon) = f(x) + \varepsilon$.

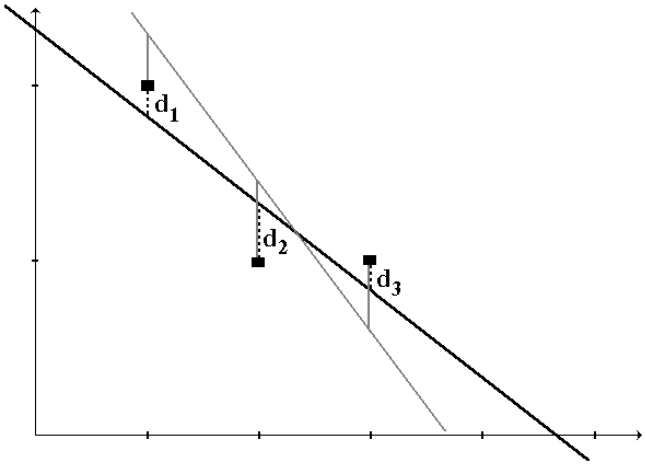
Exemplu

Pentru studiul *in vitro* al tumorilor maligne, se organizează experimente în care concentraţia factorilor de creştere (administraţi sub formă de ser fetal de viţel) este variabila *x*, iar durata ciclului celular (timpul de diviziune) este *y*.

În cazul acestui model, o ajustare (de o anumită formă specificată) care poate fi considerată "cea mai bună" se poate obţine punând condiţia ca suma pătratelor distanţelor măsurate pe verticală între punctele experimentale şi corespondentele lor pe linia de forma specificată să fie minimă. Deoarece ajustarea se determină prin minimizarea unei sume de pătrate, criteriul sau, mai corect, procedura de obţinere, se numeşte **metoda celor mai mici pătrate,** deşi denumirea mai corectă ar fi trebuit să fie "metoda celei mai mici sume de pătrate".

Un alt criteriu ar putea fi minimizarea sumei modulelor distanţelor respective. Procedeul s-ar denumi “metoda celei mai mici sume de module”. Datorită proprietăţilor matematice remarcabile se preferă însă metoda celor mai mici pătrate.

În cazul exemplului 4.3.2, dacă dorim o ajustare liniară ($f(x) = a + b \cdot x$), linia dreaptă "cea mai potrivită" datelor va fi cea pentru care am figurat distanțele d_1, d_2 şi d_3 marcate cu linii întrerupte. În desen este figurată şi o altă dreaptă - cea cu panta mai mare - care nu mai îndeplineşte condiția de minimizare a sumei de pătrate (vezi distanțele figurate prin linii continue).



2° Alte modele de regresie

Un alt model, denumit **regresia lui *x* în *y*,** poate fi obținut dacă schimbăm rolurile între *x* şi *y*. Atunci va trebui minimizată suma pătratelor distanțelor măsurate pe orizontală.

Al treilea model poate pleca de la ideea că ambele variabile sunt supuse fluctuațiilor aleatoare. Un criteriu adecvat pentru acest model ar putea fi minimizarea sumei pătratelor distanțelor de la puncte la dreaptă. Deoarece aceste distanțe se măsoară pe direcțiile perpendiculare la dreaptă (ortogonal), modelul se numeşte **regresie ortogonală**. Lista modelelor de regresie poate continua, precum şi cea a criteriilor.

În continuare ne vom referi doar la *modelul 1* şi la ajustări după diverse forme obținute prin *metoda celor mai mici pătrate*.

4.3.3. Ajustarea liniară - dreapta de regresie

Fie o serie bidimensională (x, y) de volum N, seria unidimensională (x) având media $\bar{x}$, iar seria (y), media $\bar{y}$. **Dreapta de regresie a lui *y* în *x*** are ecuația (graficului) $\boldsymbol{Y = a + b \cdot x}$, în care:

$$b = \boxed{\frac{\sum (x-\bar{x})\cdot(y-\bar{y})}{\sum (x-\bar{x})^2}} = \boxed{\boxed{\frac{N\cdot\sum x\cdot y - \sum x\cdot\sum y}{N\sum x^2 - (\sum x)^2}}}$$

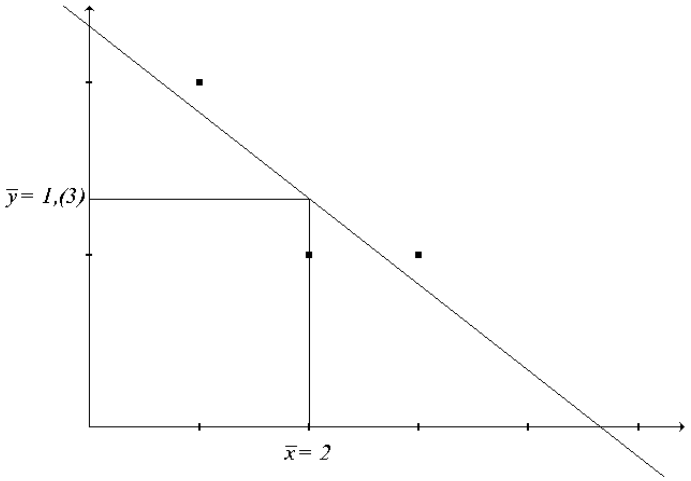
$$a = \boxed{\bar{y} - b\cdot\bar{x}} = \boxed{\boxed{\frac{\sum y - b\cdot\sum x}{N}}}$$

Primele expresii sunt **formulele teoretice**, iar ultimele sunt **formulele de calcul rapid şi exact**.

Coeficientul *b*, adică panta dreptei de regresie (a lui *y* în *x*), se numeşte **coeficientul de regresie (a lui *y* în *x*).**

Formula teoretică a lui *a* (ordonata în origine a dreptei de regresie) provine dintr-o proprietate remarcabilă şi anume:

"Dreapta de regresie trece prin punctul mediu, adică punctul de coordonate $(\overline{x}, \overline{y})$."



1° Proveniența denumirii de regresie

Francis Galton în memoriul *"Regresion toward mediocrity in hereditary stature"* (1886) citat în [19], studiind cum ar putea rămâne în echilibru dinamic o populație dacă generațiile noi ar moșteni caracteristicile măsurabile ale părinților, a observat că fiii (la maturitate) se abat de la înălțimea medie mai puțin decât tații, deci că fiii regresează către medie. De aceea, Galton a denumit linia care leagă înălțimile fiilor de cele ale taților, linie de regresie, iar procesul general de predicție a unei variabile (de exemplu înălțimea copiilor) dintr-o altă variabilă (de exemplu înălțimea părinților) a rămas în literatura statistică sub denumirea de regresie.

Prin urmare sintagma *linie* (*dreaptă* sau *curbă*) *de regresie* este improprie. Ea se păstrează din motive de tradiție, dar cititorul trebuie să se gândească, de fapt, la ideea de **linie de dependență, de corelație, de predicție, de estimație** sau **de tendință**.

4.3.4. Măsurarea gradului de corelație liniară - coeficientul de corelație liniară

Pentru definirea noțiunii de coeficient de corelație liniară este nevoie să introducem mai întâi noțiunile de *covariație* şi *covarianță*. Acestea sunt corespondentele bidimensionale ale conceptelor unidimensionale de *variație*, respectiv *varianță*. Ele exprimă împrăştierea simultană a două variabile în jurul punctului mediu $(\overline{x}, \overline{y})$, aşa cum variația şi varianța exprimă împrăştierea în jurul mediei.

Definiții: Se numeşte **covariația** seriei bidimensionale (*x, y*) de volum *N*, expresia:

$$\sum (x - \overline{x}) \cdot (y - \overline{y})$$

şi **covarianța** seriei (notată ***cov* (*x, y*)**), *covariația* divizată prin volumul *N*.

- ✓ Se observă imediat că variația unei serii unidimensionale (*x*) este covariația seriei bidimensionale (*x, x*) şi că varianța seriei unidimensionale (*x*) este covarianța seriei bidimensionale (*x, x*), adică *var* (*x*) = *cov* (*x, x*).

1° Coeficientul de corelație liniară (Bravais-Pearson)

Este cel mai cunoscut şi utilizat indicator de corelație.

Definiție: Se numeşte **coeficientul de corelație liniară (Bravais-Pearson)** al unei serii bidimensionale (x, y) de volum N, raportul:

$$R = \boxed{\frac{cov(x,y)}{\sqrt{var(x)}\cdot\sqrt{var(y)}}} = \boxed{\boxed{\frac{N\cdot\sum x\cdot y - \sum x\cdot\sum y}{\sqrt{(N\cdot\sum x^2 - (\sum x)^2)\cdot(N\cdot\sum y^2 - (\sum y)^2)}}}}.$$

Prima expresie este **formula teoretică** iar ultima este **formula de calcul rapid şi exact.**

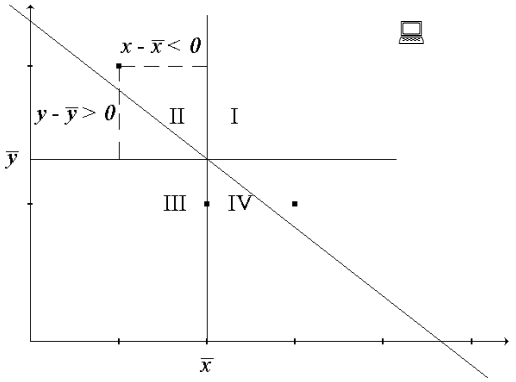
2° Proprietăți ale coeficientului de corelație liniară

1. $-1 \le R \le 1$.

Deci R poate avea şi valori negative. Într-adevăr, semnul este dat doar de numărător deci de *covarianță*, iar covariația, respectiv covarianța (*cov* (x,y)) pot avea şi valori negative.

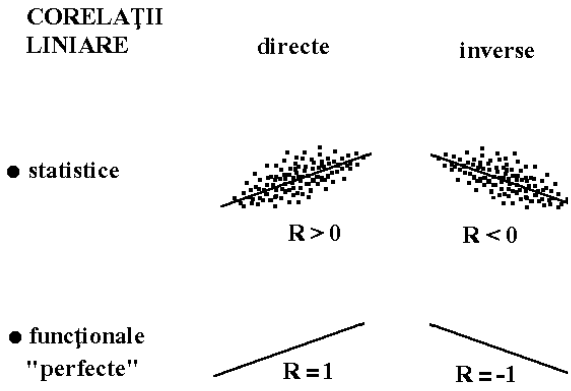
Exemplu

În desenul alăturat se observă că punctul mediu $(\bar{x}, \bar{y})$ este pentru covarianță o nouă origine, generând în plan 4 noi cadrane. Produsele $(x-\bar{x})\cdot(y-\bar{y})$ din covarianță vor avea semnele dictate de cadranul unde se află. De exemplu, pentru primul punct (1,2) produsul abaterilor față de medii va fi negativ.



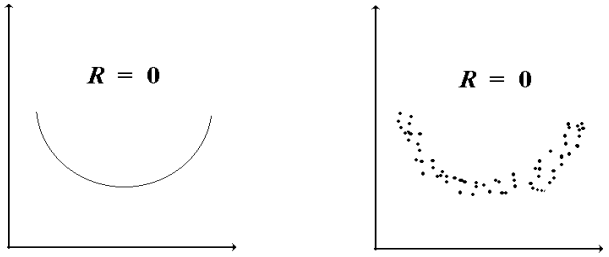
În general, în cadranele I şi III semnul va fi pozitiv, iar în cadranele II şi IV semnul va fi negativ. Dacă luăm în considerație şi mărimile acestor produse de abateri, rezultă că un R pozitiv va indica o preponderență a punctelor din cadranele I şi III, deci o alură ascendentă, iar un rezultat negativ o preponderență a punctelor din cadranele II şi IV, adică o alură descendentă, ca în cazul nostru. De aici rezultă următoarele două proprietăți.

2. $R > 0 \Leftrightarrow$ norul are o tendință ascendentă. În acest caz spunem că există o corelație liniară **directă**[22] (adică "*x* şi *y* variază în acelaşi sens"), iar
3. $R < 0 \Leftrightarrow$ norul are o tendință descendentă. În acest caz spunem că există o corelație liniară **inversă**[23] (adică "*x* şi *y* variază în sens contrar").
4. $R = 1 \Leftrightarrow$ norul se plasează pe o dreaptă ascendentă $\Leftrightarrow$ există o corelație funcțională liniară directă. În acest caz spunem că există o corelație liniară directă perfectă şi
5. $R = -1 \Leftrightarrow$ norul se plasează pe o dreaptă descendentă $\Leftrightarrow$ există o corelație funcțională liniară inversă. În acest caz spunem că există o corelație liniară inversă perfectă.



Ultimele două proprietăți caracterizează cazul dependențelor (liniare) *funcționale*.

$R = 0$ dacă şi numai dacă nu există o corelație liniară. Adică $R = 0$ dacă şi numai dacă variabilele sunt independente (vezi desenele din 4.2.5. punctul 2°) ori există o corelație dar neliniară. **Exemple:** corelația semi-circulară funcțională şi cea semicirculară statistică din figurile alăturate.



6. $|R|$ măsoară gradul de grupare a punctelor în jurul dreptei de regresie. Altfel spus, valoarea absolută a lui R exprimă **calitatea ajustării**.

✓ Una din greşelile cele mai răspândite este calculul coeficientului de corelație liniară şi în cazurile în care dependența nu este exprimată liniar şi nici nu este indicată aproximarea sa liniară. Altfel spus, se consideră în mod eronat că R este un coeficient universal de corelație, neaprofundându-se ideea că o corelație nu are sens decât raportată la o anumită formă. Greşeala provine şi din faptul că în multe lucrări R este denumit coeficient de corelație, fără atributul "liniară".

+3° Coeficientul de determinație

Definiție: Se numeşte **coeficient de determinație** pătratul coeficientului de corelație liniară. Se *notează*, în mod firesc, R^2.

[22] Unii autori o denumesc **corelație pozitivă**, ceea ce nu spune nimic ca interpretare.

[23] A nu se confunda cu dependența, corelația *invers proporțională*, prezentată în 4.3.1.

- ✓ Coeficientul de determinație exprimă proporția variației variabilei *y* care este "explicată" de dreapta de regresie. De exemplu, atunci când $R^2 = 1$ întreaga variație a lui *y* este "explicată" de dreapta de regresie deoarece toate punctele se află pe dreapta respectivă, conform proprietăților 4 și 5 de la punctul 2°.
- ✓ Deoarece $0 \le R^2 \le 1$, coeficientul de determinație, spre deosebire de coeficientul de corelație liniară, "pierde semnul" și deci nu poate indica corelațiile liniare inverse. De aceea acest coeficient este abandonat în favoarea coeficientul de corelație liniară.

4° Calcul simultan rapid și exact al dreptei de regresie și al coeficientului de corelație liniară și al coeficientului de determinație

Pentru o rezolvare manuală vom alcătui un tabel în care în primele două coloane vom pune perechile seriei bivariate (*x, y*), iar în următoarele 3 vom calcula pătratele valorilor *x*, respectiv *y* și produsele $x \cdot y$. Pe ultima linie vom calcula apoi sumele corespunzătoare.

Vom exemplifica acest calcul pe seria bivariată prezentată în majoritatea graficelor din subparagrafele anterioare începând din 4.3.2.

Exemplul 4.3.4.

	x	y	x^2	y^2	$x \cdot y$
	1	2	1	4	2
	2	1	4	1	2
	3	1	9	1	3
Sume:	**6**	**4**	**14**	**6**	**7**

Din tabel preluăm elementele necesare calculului lui *b*, *a*, respectiv *R*, conform formulelor de calcul:

$$b = \frac{N \cdot \sum x \cdot y - \sum x \cdot \sum y}{N \sum x^2 - (\sum x)^2} = \frac{3 \cdot 7 - 6 \cdot 4}{3 \cdot 14 - 6^2} = \frac{21 - 24}{42 - 36} = \frac{-3}{6} = -\frac{1}{2} = -0{,}5$$

$$a = \frac{\sum y - b \cdot \sum x}{N} = \frac{4 - (-\frac{1}{2}) \cdot 6}{3} = \frac{4 + 3}{3} = \frac{7}{3} = 2{,}(3)$$

Prin urmare, ecuația dreptei de regresie este $Y = 2{,}(3) - 0{,}5 \cdot x$. Coeficientul de corelație liniară *R*, este:

$$R = \frac{N \cdot \sum x \cdot y - \sum x \cdot \sum y}{\sqrt{(N \sum x^2 - (\sum x)^2) \cdot (N \sum y^2 - (\sum y)^2)}} = \frac{3 \cdot 7 - 6 \cdot 4}{\sqrt{(3 \cdot 14 - 6^2) \cdot (3 \cdot 6 - 4^2)}} =$$

$$= \frac{21 - 24}{\sqrt{(42 - 36) \cdot (18 - 16)}} = \frac{-3}{\sqrt{6 \cdot 2}} = \frac{-3}{\sqrt{12}} \cong \frac{-3}{3{,}46} \cong -0{,}87 .$$

Interpretare:

Deoarece $R < 0$, corelația liniară este inversă, iar pentru că $|R| = 0{,}87$ este relativ apropiat de 1, calitatea ajustării liniare este satisfăcătoare.

Coeficientul de determinație, R^2, are valoarea:

$$R^2 = (-0{,}87)^2 = 0{,}7569.$$

Interpretare:

75,69 % din variația lui *y* este explicată de dreapta de regresie $y = 2,(3) – 0,5 \cdot x$. Altfel spus, dreapta $y = 2,(3) – 0,5 \cdot x$ explică 75,69 % din variația lui *y*.

4.3.5. Ajustări liniarizabile şi măsurarea gradului de corelație corespunzător

Curbele care pot fi liniarizate se bucură, după transformarea respectivă, de proprietățile remarcabile ale regresiei şi corelației liniare. Una dintre acestea este proprietatea coeficientului de corelație liniară (aplicat datelor transformate) de a indica prin semn sensul corelației (directă sau inversă).

Cu alte cuvinte, în cazul unei forme de corelație liniarizabile, în loc de a calcula direct, dar mai complicat în general, curba respectivă de regresie, se calculează dreapta de regresie pentru datele transformate pentru liniarizare. Apoi se apreciază gradul corelației curbilinii respective prin coeficientul de corelație liniară aplicat datelor transformate.

1° Calculul regresiei hiperbolice şi aprecierea indirectă a corelației respective

Exemplul 4.3.5.

Reluăm exemplul 4.3.4 aplicând transformarea expusă la punctul 2° din 4.3.1. şi anume $X = 1 / x$. Prin urmare, vom calcula dreapta de regresie şi coeficientul de corelație liniară între $X = 1 / x$ şi *y*. Se construieşte din nou tabela anterioară, înlocuindu-se însă valorile lui *x* cu $1 / x$, adică în prima coloană se pun valorile 1, 1 / 2, 1 / 3. Se aplică în continuare exact aceleaşi etape de calcul. În final se obține ecuația $Y = 0,35 + 1,6 / x$ şi R' (coeficientul de corelație liniară a seriei bivariate $(1 / x, y)$) = 0,97.

Observăm că în acest caz, coeficientul de corelație liniară a devenit pozitiv pentru că *y* depinde în mod direct de $1 / x$, ceea ce este firesc din moment ce *y* depinde în mod invers de *x* iar $1 / x$ inversează ordinea.

2° Compararea celor două ajustări

Deoarece, conform proprietății 7 a coeficientului de corelație liniară, calitatea ajustării este dată de valoarea sa absolută vom compara valorile absolute ale celor doi coeficienți.

Exemplul 4.3.5'.

În exemplele 4.3.4. şi 4.3.5. am obținut valorile $R = - 0,87$, respectiv $R' = 0,97$. Cum $|R| < |R'|$ rezultă că *ajustarea hiperbolică* este mai bună decât cea liniară, ceea ce se observă şi vizual pe diagrama de împrăştiere din exemplul 4.3.2.

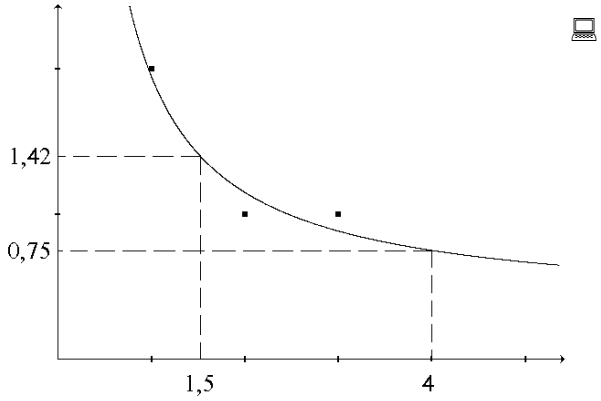
3° Rostul ajustării - prognoza prin ecuația de regresie

O dată determinată ca parametri forma care se ajustează cel mai bine la datele experimentale am obținut aproximarea cea mai convenabilă, printr-o ecuație, a norului de puncte. Astfel, ajungem la posibilitatea prognozei prin interpolare şi prin extrapolare, evident în anumite limite ale fenomenului.

Exemplul 4.3.5''.

Continuăm exemplul anterior. Pentru x = 1,5 (valoare cuprinsă între valorile experimentale, deci) prin interpolare, *hiperbola de regresie* ne va conduce la valoarea 1,42 obținută astfel: 0,35 + 1,6 / 1,5 = 1,42.

Prin extrapolare pentru x = 4 (valoare în afara datelor experimentale), obținem: 0,35 + 1,6 / 4 = 0,75.



✓ În mod evident acestea vor fi doar valori aproximative într-un fenomen biologic, deoarece dependența este doar statistică, nu funcțională, chiar dacă norul de puncte nu conține replicate diferite ale lui y pentru un x fixat, având date puține. Volumul datelor a fost aici intenționat luat mic, pentru a ușura înțelegerea de principiu şi a calculelor.

+4.3.6. Alte ajustări şi validarea lor

Pentru formele de dependență neliniarizabile ecuațiile de regresie se obțin sau se aproximează mult mai dificil.

Validarea ajustărilor cu astfel de curbe este însă mai accesibilă deoarece există un indicator care se poate aplica oricărei forme de dependență. Acesta se întâlneşte în literatură sub denumirea de **indice** (sau raport) de **corelație**, dar, după părerea noastră ar trebui adăugată în titulatură şi un calificativ al formei respective. Adică ar trebui să vorbim de *indicele de corelație liniară, indicele de corelație hiperbolic*ă, *indicele de corelație exponențială* etc.

✓ Astfel se poate înlătura confuzia generată de formulări din literatură de tipul: "raportul de corelație joacă, în cazul corelației neliniare, un rol analog cu cel al coeficientului de corelație în cazul corelației liniare"[19]. Din acest citat rezultă, pentru începător, că există doar două tipuri de corelație: cea liniară şi cea neliniară, când, de fapt există o infinitate de tipuri (forme) de corelație neliniare.

Indicele de corelație liniară este egal cu *coeficientul de corelație liniară* fără semn, ceea ce justifică, încă o dată, preferința pentru coeficientul de corelație liniară.

Aceşti indici de corelație au o formulă unică (de unde provine şi denumirea unică din literatură) dar, evident valori diferite, pentru că fiecare se raportează la o curbă de regresie de altă formă. Formula unică de definiție se bazează pe proprietatea de aditivitate a variației.