Elemente de matematică aplicate în biologie

Motto
Matematica se bucură de o poziție specială în raport cu celelalte ştiințe pentru că legile ei sunt absolut certe şi indiscutabile
(A. Einstein, Geometry and experience, Sidelight on Relativity, Dover Publication, New York, 1983)

Conf. Univ.Dr. Dana Constantinescu

1. Introducere

Argument

Matematica a câştigat şi şi-a menţinut o poziţie excepţională între ştiinţe pentru că rezultatele sale sunt obţinute dintr-un număr mic de axiome (mai mult sau mai puţin evidente) printr -un lanţ de raţionamente. Deoarece e bazată pe o logică impecabilă, matematica furnizează ştiinţelor naturale un grad înalt de securitate (şi claritate) care altfel nu poate fi atins. Din acest motiv, tratarea riguros matematică a acestora este de dorit şi se realizează ori de câte ori e posibil. Mai mult decât atât, matematica este un mijloc de comunicare între oameni de ştiinţă şi ingineri de diverse specialităţi, Ca rezultat, dacă o anumită ramură a ştiinţei este prezentată în formă riguros matematică, accesibilitatea şi audienţa ei sporeşte.

(I. D. Mayergoyz, Mathematical Models of hysteresis and their applications, Elsevier Science Inc. New York, 2003)

Deşi dezvoltarea biologiei nu a fost influenţată în mod esenţial de dezvoltarea matematicii, în ultimele decenii este recunoscută importanţa completării studiului descriptiv al unor fenomene sau mecanisme biologice cu aspecte legate de prelucrearea şi interpretarea datelor obţinute.
Cea mai avansată formă a folosirii matematicii în biologie este biologia matematică. Ea îşi propune modelarea matematică a proceselor biologice şi studiul modelelor folosind metode specifice matematicii. Pentru construirea şi validarea modelelor matematice se pot folosi cercetări statistice.

Statistica dezvoltă tehnici şi proceduri de înregistrare, descriere, analiză şi interpretare a datelor experimentale sau a rezultatelor obţinute din observarea unui proces social, economic, biologic etc., precum şi vizualizarea datelor folosind softuri dedicate acestui scop. Cunoaşterea unor elemente şi principii de bază ale statisticii este importantă în momentul actual, permiţând realizarea unor analize corecte a datelor şi evitarea erorilor de interpretare a acestora. Strâns legată de statistica inferenţială este teoria probabilităţilor, care furnizează metode şi tehnici pentru stabilirea unor previziuni (inferenţe statistice) referitoare la caracteristicile unei populaţii pornind de la rezultatele obţinute din observarea unui eşantion al acesteia.

Biostatistica (combinaţie de cuvinte între biologie şi statistică) este aplicarea statisticii într-un număr mare de domenii ale biologiei.
Biostatistica are drept obiectiv şi fundamentarea teoretică a proiectării şi controlului experimentelor biologice, mai ales în medicină şi agricultură, deoarece ea analizează şi interpretează date concrete şi realizează inferenţe asupra acestora.
Se consideră că principalii beneficiari ai biostatisticii sunt

- Sănatatea publică (studiul aspectelor epidemiologe, legate de nutriţie, corelarea stării de sănătate şi proprietăţile mediului înconjurător, organizarea serviciilor de studiu al sănătăţii populaţiei)
- Ecologia şi previziunile ecologice (studiul inflenţei diverşilor factori asupra dinamicii populaţiilor)
- Statistica genetică (studiază legătura între variaţiile genotipului şi ale fenotipului). Studiul genetic al populaţiilor este folosit în agricultură pentru îmbunătăţirea soiurilor de plante şi animale, iar în genetica umană studiul statistic ajută la identificarea cauzelor care influenţează predispoziţia la anumite afecţiuni)
- Analiza secvenţelor biologice (secvenţe AND, secvenţe de peptide…)

In cele ce urmează prezentăm unele aplicaţii directe ale statisticii matematice şi ale teoriei probabilităţilor în descrierea unor fenomene simple ce apar în biologie şi agricultură. Asocierea celor două domenii beneficiare ale matematicii nu este întâmplătoare, agricultura fiind în bună măsură biologie aplicată.

2. Aplicaţii ale statisticii descriptive în biologie şi în agricultură

Statistica matematică se ocupă cu descrierea şi analiza numerică a fenomenelor (sociale, economice, ştiinţifice etc). Statistica operează cu date care se pot colecta din surse existente sau se pot obţine prin observaţii şi studii experimentale.

Datele statistice sunt în fapt observaţii codificate realizate asupra unei mulţimi de elemente de aceeaşi natură, mulţime care se numeşte ***populaţie statistică.*** O populaţie poate fi finită sau infinită. Numărul de elemente al unei populaţii finite se numeşte *volumul populaţiei.*

Elementele populaţiei (indivizii) sunt purtătoare de informaţii. Indivizii pot fi persoane (de exemplu formând populaţia unei localităţi), agenţi economici, obiecte (de exemplu mijloacele fixe ale unui agent economic, piese produse sau comercializate), evenimente (de exemplu operatiuni bancare), opinii (relative la servicii, calitatea unui produs), etc.

Caracteristica populaţiei este trăsătura comună a elementelor sale care este supusă studiului statistic. In statistica matematică ea este cuantificată prin valori numerice. Deoarece o caracteristică variază de la individ la individ, ea poate fi considerată ca o funcţie $X : P \to R$, unde P este populaţia statistică.

O caracteristică poate fi discretă (dacă valorile sale formează o mulţime finită) sau continuă (în cazul când caracteristica poate lua orice valoare reală).

De exemplu, caracteristica ce indică numărul de piese defecte din fiecare lot este o discretă, în timp ce profitul unei firme sau volumul încasărilor pot fi interpretate ca şi caracteristici continue.

Un fenomen deosebit de important este cuantificarea fenomenelor sociale, adică transpunerea în limbaj numeric a caracteristicilor acestor fenomene pentru a înlesni compararea, analiza şi sinteza lor, precum şi pentru a face prognoze asupra lor.

Problema cuantificării fenomenelor sociale este o problemă de bază a ştiinţelor sociale, în condiţiile creşterii exigenţelor faţă de determinările ştinţifice ale acestora.

Există *fenomene sociale măsurabile prin natura lor*, de exemplu fenomenele demografice, fenomenele economice, diverse fenomene politice sau culturale

Fenomenele sociale măsurabile cu aproximaţie se referă în special la opiniile şi comportamentele colectivităţilor umane. În acest caz măsurarea nu poate fi efectuată decât prin compararea intensităţilor cu care se manifestă acestea la diverse persoane, adică prin realizarea unei scări de mărimi numită *scalogramă*.

Un exemplu de scalogramă care reprezintă intensitatea opiniilor este cea care conţine trei niveluri: cu totul de acord, de acord, nu sunt de acord.

Statistica matematică operează cu fenome cuantificabile numeric, deci fiecărui element al unei scalograme i se asociază un număr.

Demersul statistic are două niveluri: descrierea statistică (statistica descriptivă) şi inferenţa statistică (statistica inferenţială).

Statistica descriptivă se ocupă cu înregistrarea, gruparea, prelucrarea şi prezentarea datelor obţinute prin investigaţie şi pe această bază descrie fenomenul studiat. În studiul statistic descriptiv toate elementele populaţiei sunt luate în consideraţie. Scopul statisticii descriptive este îndepărtarea detaliilor neimportante şi focalizarea atenţiei asupra unor aspecte de interes şi anume:

- precizarea valorii în jurul căreia sunt centrate datele
- descrierea împrăştierea acestora în jurul valorii centrale
- vizualizarea datelor cu ajutorul histogramelor
- analiza corelaţiei între fenomene

Statistica inferenţială are ca obiect de studiu investigarea prin sondaj: din întreaga populaţie se selectează un eşantion reprezentativ asupra căruia se fac măsurători sau observaţii legate de o anumită caracteristică a populaţiei. Pe baza rezultatelor obţinute se fac *inferenţe statistice* (adică se formulează concluzii) asupra parametrilor populaţiei. Statistica inferenţială foloseşte deci informaţia rezultată din studierea unui eşantion pentru a obţine concluzii referitoare la întraga populaţie din care a fost selectat eşantionul. Aceste concluzii nu sunt de tip determinist ci se obţin folosind metode şi tehnici ale teoriei probabilităţilor, teorie ce conţine mecanisme de măsurare şi analiză a incertitudinii

legate de evenimentele viitoare. Această incertitudine este exprimată cu ajutorul nivelelor de încredere.

In realizarea unei cercetări statistice se parcurg de obicei urmatoarele etape:
- *colectarea datelor* care se realizează prin metode specifice obiectivului şi condiţiilor cercetării. In funcţie de tipul de analiză folosit (descriptivă sau inferenţială) se foloseşte întreaga populaţiei sau doar un eşantion.
- *procesarea datelor* înseamnă cuantificarea lor numerică şi obţinerea seriilor de date.
- *analiza datelor* se realizează prin metode şi tehnici specifice statisticii matematice. Această etapă necesită o cunoştere profundă a filosofiei ce stă în spatele fiecărei metode deoarece este posibil să se obţină rezultate nesemnificative statistic atunci când ipotezele de lucru sau condiţiile de aplicare a metodelor nu sunt îndeplinite.
-*interpretarea rezultatelor* este diferită în statistica descriptivă şi în cea inferenţială. In primul caz se obţin informaţii concrete şi clare despre populaţia studiată, în al doilea caz validarea rezultatelor obţinute este realizată prin compararea cu ce se ştia sau se bănuia în domeniul respective. In unele situaţii analiza statistică dezvăluie corelaţii între fenomene, legături care ar fi fost greu sau chiar imposibil de observat fără eficientul mecanism statistico-matematic.

In momentul de fată există o vastă informaţie statistică la nivel global, datorată în principal dezvoltării continue a tehnologiei calculatoarelor. Realizarea şi folosirea corectă a bazelor de date reprezintă o preocupare importantă în mediul economic si nu numai. Soft-urile statistice joacă un rol important în analiza datelor. Ele îmbină proceduri statistice clasice şi moderne cu tehnici de grafică interactivă. Multe soft-uri au două versiuni: una profesională şi una academică. Literatura de specialitate califica drept foarte performante, printer altele, următoarele pachete de programe:
- S-PLUS (http://www.insightful.com/products/splus/)
- XploRe (http://www.xploretech.com/index.pl)
- Statistica (http://www.statsoft.com/)
- SPSS (http://www.spss.com/)

2.1. Serii de date şi distribuţii de frecvenţe

Considerăm o populaţie statistică P finită de volum N pentru care o caracteristică C este codificată de valorile numerice $x_1, x_2, ..., x_N$, nu neapărat diferite.
Sirul finit de numere se notează

$$X : x_1, x_2, ..., x_N$$

şi se numeşte ***serie de date.***

Exemplu: $X : 0,1,0,0,2$ este o serie de date care poate fi interpretată o funcţie $X : \{a,b,c,d,e\} \to \{0,1,2\}$, unde $X(a) = 0$, $X(b) = 1$, $X(c) = 0$, $X(d) = 0$, $X(e) = 2$.
In acest caz populaţia este $P = \{a,b,c,d,e\}$. Deoarece identitatea indivizilor din populaţie nu este interesantă din punct de vedere statistic, aceasta este neglijată în etapele următoare.

Definiţie: Distribuţia de frecvenţe (sau variabila statistică) asociată caracteristicii C a populaţiei P de volum N este

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_k \\ n_1 & n_2 & n_3 & n_k \end{pmatrix}$$

unde x_j, $j \in \{1,2,...,k\}$ sunt valorile diferite înregistrate pentru caracteristica C iar n_j, $j \in \{1,2,...k\}$ reprezintă numărul indivizilor populaţiei caracterizaţi de valoarea x_j.
Numărul n_j se numeşte frecvenţa absolută de apariţie a valorii x_j.
Observaţii: 1. Din definiţia frecvenţelor relative rezultă că

$$\sum_{j=1}^{k} n_j = n_1 + n_2 + ... + n_k = N .$$

2. Unei caracteristici i se poate asocia şi **distribuţia frecvenţelor relative**

$$X_r = \begin{pmatrix} x_1 & x_2 & x_3 & x_k \\ f_1 & f_2 & f_3 & f_k \end{pmatrix}, \quad f_j = \frac{n_j}{N} .$$

În acest caz $\sum_{j=1}^{k} f_j = 1$. Frecvența relativă f_j poate fi interpretată ca fiind probabilitatea ca valoarea x_j să fie luată de caracteristica *C*, iar distribuția frecvențelor relative este în fapt o variabilă aleatoare.

Exemplu: Pentru seria de date $X: 0,1,2,5,2,3,3,2$

distribuția de frecvențe este $X = \begin{pmatrix} 0 & 1 & 2 & 3 & 5 \\ 1 & 1 & 3 & 2 & 1 \end{pmatrix}$ iar cea a frecvențelor relative este $X_r = \begin{pmatrix} 0 & 1 & 2 & 3 & 5 \\ 1/8 & 1/8 & 3/8 & 2/8 & 1/8 \end{pmatrix}$

2.2. Reprezentări grafice

Graficul corespunzător unei serii statistice se numeşte diagramă. Cazul seriilor pentru care caracteristica este măsurată cantitativ (şi exprimată prin numere reale) se întâlnesc în mod current următoarele reprezentări grafice:

- reprezentarea cu segmente vericale:
- histograma cu bare
- poligonul frecvențelor
- reprezentarea cu sectoare circulare

a) **Reprezentarea cu segmente verticale (histograma cu segmente)** se foloseşte pentru serii cu un număr redus de date, de obicei numere întregi.

Pentru distribuția de frecvențe $X_r = \begin{pmatrix} x_1 & x_2 & x_3 & x_k \\ n_1 & n_2 & n_3 & n_k \end{pmatrix}$, histograma cu segmente, sau reprezentarea cu segmente, este familia de segmente verticale ce unesc punctele de coordonate $(x_i, 0)$ şi (x_i, n_i) unde $i \in \{1,2,...,k\}$

Exemplu: Pentru $X = \begin{pmatrix} 1 & 3 & 2 & 4 & 5 \\ 3 & 2 & 4 & 3 & 1 \end{pmatrix}$ reprezentarea cu segmente verticale este prezentată în figura 2.1.

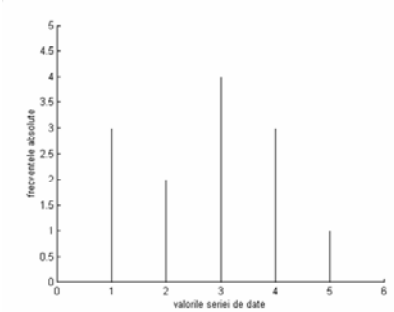

Figura 2.1. Histograma cu segmente

b) **Histograma cu bare** se foloseşte pentru seriile cu un număr mare de date ce nu sunt neapărat numere întregi. Ea se realizează astfel:

- se determina valoarea minimă, $x_{\min}$ şi valoarea maximă $x_{\max}$ a seriei de date

- se divide segmentul $[x_{\min}, x_{\max}]$ prin puncte echidistante cu pasul $h = \frac{x_{\max} - x_{\min}}{n}$, unde n este numărul de intervale ales de analistul seriei. Punctele de diviziune sunt $x_j = x_{\min} + j \cdot h$, unde $j \in \{0,1,2,...,n\}$
- se calculează câte valori ale seriei aparţin fiecărui interval $I_j = [x_j, x_{j+1})$. Acest număr, notat n_j, se numeşte frecvenţa clasei I_j.
- Deasupra fiecărui interval I_j se trasează un dreptunghi cu baza I_j şi înălţimea proporţională cu n_j. Pentru determinarea înălţimii dreptunghiului se poate folosi formula $H_j = \frac{n_j}{h \cdot N}$.

Obiecul grafic rezultat din alăturarea acestor dreptunghiuri se numeşte *histograma cu bare a seriei de date* sau histograma distribuţiei de frecvenţe, pentru că ilustrează modul în care sunt distribuite datele. Un exemplu de histogramă cu bare este dat in Figura 2.2.

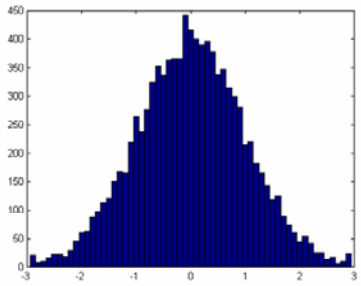

Figura 1.2. Histograma cu bare

O problemă legată de generarea histogramelor este legată de precizarea numărului de intervale de diviziune. In perioada de început a statisticii computaţionale numărul de intervale era proporţional cu $\sqrt{N}$. In unele programme statistice el este ales proporţional cu $\log_2 N$. Cea mai bună idée este să generăm histograme corespunzătoare mai multor numere de intervale şi să le comparăm.

c) Poligonul frecventelor se obţine unind vârfurile segmentelor verticale în cazul reprezentării cu segmente. In cazul reprezentării din Figura 2.1, poligonul de frecvenţe, A, B, C, D, E este dat în figura 2.3.

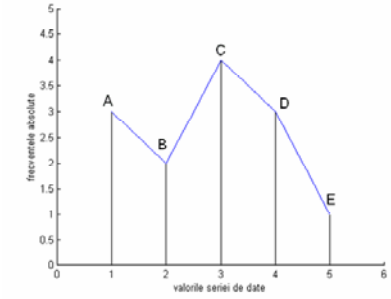

Figura 2.3. Poligon de frecvenţe

d) Reprezentarea cu sectoare circulare este folosită pentru obţinerea rapidă a unei viziuni globale asupra importanţei relative a diverselor clase ale statisticii, interpretarea lor fiind uşurată de colorarea diferită a diverselor clase. In general această reprezentare este folosită pentru seriile cu un număr mic de clase.

Reprezentarea se realizează astfel:

- se determină clasele seriei şi numărul de valori ale seriei din fiecare clasă (frecvenţele absolute ale claselor)
- pe un cerc se consideră sectoare circulare proporţionale cu frecvenţele fiecărei clase. Unghiul la centru corespunzător clasei cu frecvenţa absolută n_j este $\theta_j = \frac{n_j}{360 \cdot N}$.

e) **Reprezentarea polară** se foloseşte atunci când caracteristica statistică prezintă o anumită periodicitate. De exemplu date inregistrate calendaristic (numarul de nasteri inregistrate în fiecare lună) sau date referitoare la aspecte geografice (intensitatea vântului ce bate din anumite direcţii).

Ea se construieşte astfel: pe semidrepte cu aceeaşi origine şi care impart planul într-un număr de sectoare egale (acest număr se stabileşte în funcţie de caracterul seriei statistice) se consideră segmente ce pornesc din origine, proporţionale cu frecvenţele absolute ale claselor şi se unesc extremităţile acestoe segmente. Se obţine un poligon închis în care clasele cu frecvenţă mai mare sunt reprezentate prin vârfuri aflate la distanţă mai mare faţă de origine.

2.3. Indicatori statistici

2.3.1. *Indicatori de poziţie (de nivel, de localizare)*

a) **media aritmetică** $\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_k n_k}{N}$

Media aritmetică este sensibilă faţă de valorile extreme ale seriei, ea devenind nereprezentativă dacă termenii seriei sunt foarte împrăştiaţi. Omogenitatea colectivităţii este o condiţie a reprezentativităţii, pentru orice tip de mărime medie.

b) **media armonică** $x_{arm} = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + ... + \frac{n_k}{x_k}}$

Media armonică este influenţată de prezenţa valorilor individuale mici şi de frecvenţa acestora.

Media armonică se utilizează pentru exprimarea tendinţei centrale în funcţie de scopul cercetării şi mai ales în funcţie de natura obiectivă dintre valorile variabilei numerice observate.

In economie este folosită la calculul productivităţii, pentru calculul indicelui (sintetic) al preţurilor mărfurilor şi tarifelor serviciilor (care sintetizează indicii individuali ai acestor preţuri şi tarife).

c) **media geometrică** $x_g = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot .... \cdot x_k^{n_k}}$

Media geometrică este folosită mai rar ca indicator statistic, îndeosebi când termenii prezintă o evidentă concentrare către valorile cele mai mici sau când se urmăreşte să se acorde o importanţă deosebită valorilor individuale reduse.

Dacă cel puţin o valoare individuală este nulă sau negativă, calculul mediei geometrice este lipsit de sens. Ea nu poate fi folosită dacă în cadrul seriei există cel puţin un termen negativ, deoarece expresia devine imaginară.

Media geometrică mai este denumită şi *medie de ritm*, fiind folosită pentru calculul ritmului mediu de crestere. Un exemplu de folosire a mediei geometrice ca indicator statistic este dat în exemplul următor:

Exemplu O colonie de microorganisme a fost studiată pe parcursul a două zile. S-a constatat ca masa sa iniţială era 10 g, după o zi era 20 g iagr a treia zi era 160 . Să se calculeze ritmul mediu de creştere al coloniei.

Masa coloniei s-a dublat în prima zi şi s-a multiplicat de 8 ori în a doua zi. Dacă se calculează rapid media aritmetică se constată că, în medie, ritmul de creştere este $\frac{2+8}{2} = 5$.

Acest rezultat în mod evident este incorect deoarece, în acest caz după o zi colonia ar avea , $10 \cdot 5 = 50g$, iar după două zile ar avea $50 \cdot 5 = 250g$, ceea ce nu este adevărat

Dimpotrivă, dacă indicele mediu de dinamică se determină ca media geometrică a dinamicilor individuale se obţine următoarea valoare: $\overline{x_g} = \sqrt{2 \cdot 8} = 4$. Acesta este un rezultat mult mai corect decât cel anterior deoarece pornind de la 10 g colonia ar avea $10 \cdot 4 = 40g$ după prima zi (ceea ce nu e adevărat) şi $40 \cdot 4 = 160g$ după a doua zi. Acest rezultat verifică datele problemei, deci ritmul mediu de creştere este egal cu media geometrică a ritmurilor intermediare de creştere, adică este 4.

d) **mediana** seriei de date $X : x_1, x_2, ..., x_N$ cu termenii ordonaţi crescător este numărul

$$me=\begin{cases} x_{\frac{N+1}{2}} & daca\ \ N\ \ este\ \ impar \\ \dfrac{x_{N/2}+x_{1+N/2}}{2} & daca\ \ N\ \ este\ \ par \end{cases}.$$

Mediana este o valoare ce caracterizează "centrul" seriei de date. În cazul când *N* este par mediana nu este obligatoriu valoare a seriei de date.

Are proprietatea că suma frecvențelor valorilor mai mici ca *me* este egală cu suma frecvențelor mai mari ca *me.*

Este utilizată în studiul fertilității, mortalității, determinarea duratei de viață.

e) **modul** (moda su dominanta) este valoarea cu cea mai mare frecvență de apariție (care este la modă). Există repartiții unimodale (cu un singur mod), bimodale (cu două moduri) etc.

Valoare modală este influențată de mărimea valorilor din centrul seriei (la distribuțiile unimodale) sau din centrul îngrămădirii de observații (la distribuțiile plurimodale). Celelalte valori nu au nici o influență asupra ei.

Distribuțiile bimodale (cu două frecvențe maxime) reprezintă o situație rar întâlnită, care impune separarea unităților colectivității în două distribuții de frecvențe.

2.3.2. Indicatorii variației (împrăştierii)

Indicatorii tendinței centrale nu dau nici o explicație asupra împrăstierii, respectiv a modului în care termenii seriei se abat între ei sau de la medie. Astfel, apare necesitatea calculării unor noi indicatori care rezolvă:

- verificarea reprezentativității mediei ca valoare tipică a seriei de distribuție;
- verificarea gradului de omogenitate al seriei;
- verificarea sistematizării informațiilor prin gruparea statistică;
- caracterizarea gradului si formei de variație a unei variabile statistice.

Aceşti indicatori care dau o caracterizare precisă a unei serii statistice prin care se poate cunoaşte *variația* valorilor individuale (cum se grupează aceste valori în jurul valorii medii, dacă sunt apropiate sau îndepărtate de această valoare), se numesc *indicatorii variației.* Ei sunt:

a) **amplitudinea** este diferența dintre cea mai mare şi cea mai mică valoare a seriei de date (sau a distribuției de frecvențe)

b) **abaterea medie absolută** $e_X=\dfrac{1}{N}\sum_{i=1}^{k}n_i\left|x_i-\bar{x}\right|$

c) **varianța (dispersia)** $s^2=\dfrac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2$

d) **abaterea medie pătratică (standard)** $s=\sqrt{\dfrac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}$

Propoziție Dispersia şi abaterea medie pătratică ale unei distribuții de frecvențe $X=\begin{pmatrix} x_1\ x_2\ x_3\ \ldots x_k \\ n_1\ n_2\ n_3\ \ldots n_k \end{pmatrix}$**, unde** $\sum_{i=1}^{k}n_i=N$ **se calculează folosind formulele**

$s^2=\dfrac{\sum_{i=1}^{k}x_i^{\,2}\cdot n_i}{N}-\left(\dfrac{\sum_{i=1}^{k}x_i\cdot n_i}{N}\right)^2$**, respectiv** $s=\sqrt{\dfrac{\sum_{i=1}^{k}x_i^{\,2}\cdot n_i}{N}-\left(\dfrac{\sum_{i=1}^{k}x_i\cdot n_i}{N}\right)^2}$.

Dispersia este un indice de variație ce dă indicații privind împrăştierea valorilor seriei în jurul valorii medii. Cu cât este mai mică dispersia, cu atât valorile seriei statistice se grupează mai mult în jurul valorii medii. In acest caz media este un indicator statistic relevant pentru studiul seriei. O dispersie mare arată că elementele eşantionului au o împrăştiere mare şi valoarea medie nu dă informații relevante despre serie.

Dispersia este influențată de mărimea valorilor din seria de date. Dacă valorile sunt mari, dispersia poate fi si ea mare, dar cazul seriilor de date cu valori mici dispersia poate avea valori mici chiar dacă datele nu sunt grupate în jurul mediei. De aceea, pentru studiul împrăştierii se foloseşte coeficientul de variație care nu este influențat în mod esențial de mărimea termenilor seriei de date.

e) **coeficientul de variație** $CV = \frac{s}{\bar{x}}$

Coeficientul de variație are valori cuprinse în intervalul $[0,1]$. El este cel mai sintetic indicator al împrăştierii.

Cu cât coeficientul de variație e mai aproape de 0, cu atât seria este mai omogenă şi media este mai reprezentativă. Dacă este mai apropiat de 1, împrăştierea valorilor este mare şi media nu este un indicator reprezentativ.

Practica utilizării coeficientului de variație a stabilit pragul de trecere de la starea de omogenitate la cea de eterogenitate: În literatura de specialitate se avansează nivelul de 35 - 40 % ca limită maximă admisibilă pentru coeficientul de variație.

- Dacă $CV \le 0.35$, populația este *omogenă* şi media este un indicator relevant.
- Dacă $CV > 0.35$, populația este *eterogenă* şi media nu este un indicator relevant

In analizele financiare coeficientul de variație este o măsură a riscului relativ.

Exemple

1 Cantitatea de deşeuri organice produse la o ferma în decursul a 100 zile consecutive a fost înregistrată în tabelul de mai jos

Cantitatea de deşeuri produse zilnic "x_i"	Numarul de zile în care s-a produs cantitatea de deşeuri "n_i	Frecvența relativă $n_i/100$
0	5	
1	15	
2	23	
3	22	
4	16	
5	9	
6	5	
7	5	

a) Să se completeze coloana frecvențelor relative;
b) Să se deseneze histograma cu segmente verticale asociată datelor din table.
c) Să se calculeze indicatorii de pozitie (media, mediana, modul) şi indicatorii de împrăştiere (dispersia, abaterea standard şi coeficientul de variație)
d) Să se interpreteze datele obținute

Rezolvare: a) $X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \frac{5}{100} & \frac{15}{100} & \frac{23}{100} & \frac{22}{100} & \frac{16}{100} & \frac{9}{100} & \frac{5}{100} & \frac{5}{100} \end{pmatrix}$

b) Histograma cu segmente este

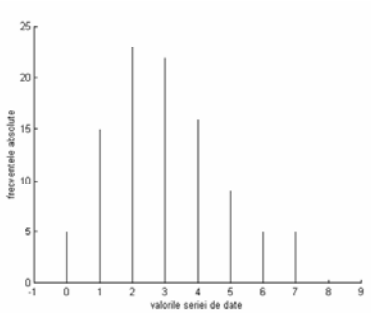

Figura 2.4 Histograma cu segmente a seriei de date din exercițiul 1

c)Indicatorii de poziție sunt:

- media $\bar{x} = \frac{0\cdot 5 + 1\cdot 15 + 2\cdot 23 + 3\cdot 22 + 4\cdot 16 + 5\cdot 9 + 6\cdot 5 + 7\cdot 5}{100} = 3.01$

- mediana se calculează tinând cont ca sunt 100 termeni în serie. Dacă scriem termenii seriei în ordine crescătoare, repetându-i de atâtea ori cât indică frecvența absolută obținem $x_{50} = x_{51} = 3$ Deci $me(X) = \frac{x_{50} + x_{51}}{2} = \frac{3+3}{2} = 3$.

-modul este $mo(X) = 2$ pentru că această valoare are cel mai mare număr de apariții.

Indicatorii de poziție sunt

-dispersia: $s^2 = \frac{1}{100}\cdot[5\cdot 0^2 + 15\cdot 1^2 + 23\cdot 2^2 + 22\cdot 3^2 + 16\cdot 4^2 + 9\cdot 5^2 + 5\cdot 6^2 + 5\cdot 7^2 - \frac{2.85^2}{100}] = 3.0499$

- abaterea standard $s = \sqrt{s^2} = \sqrt{3.0499} = 1.7463$

-coeficientul de variație $CV = \frac{s}{\bar{x}} = 0.58$.

2.Vârsta persoanelor dintr-o comunitate a fost înregistrată şi datele au fost grupate în tabelul de mai jos.

Vârsta (în ani)	Număr persoane
[0,5)	5
[5,10)	12
[10,15)	33
[15,20)	71
[20,25)	119
[25,30)	175
[30,35)	185
[35,40)	158
[40,45)	122
[45,50)	69
[50,55)	35
[55,60)	11
>=60	5
Total	1000

a) Să se deseneze histograma cu bare a acestei serii de date (vârstele mai mari de 60 ani se identifică cu intervalul [60,65)**.**

b) Identificând fiecare interval cu mijlocul său, să se constituie seria statistică a vârstelor celor 1000 de persoane din comunitate. Să se determine media, mediana şi dispersia acestei serii.

Rezolvare:

a) histograma este

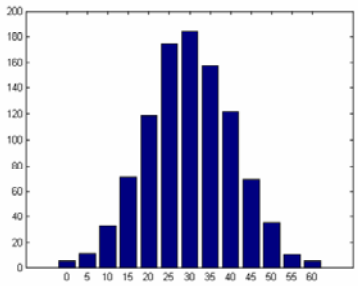

Figura 2.5 Histograma cu bare a seriei de date din exercițiul 2

b) Seria de date este

$$X = \begin{pmatrix} 2.25 & 7.25 & 12.25 & 17.25 & 22.25 & 27.25 & 32.25 & 37.25 & 42.25 & 47.25 & 52.25 & 57.25 & 62.25 \\ 5 & 12 & 33 & 71 & 119 & 175 & 185 & 158 & 122 & 69 & 35 & 11 & 5 \end{pmatrix}$$

Media este $\bar{x} = (2.25 \cdot 5 + 7.25 \cdot 12 + 12.25 \cdot 33 + 17.25 \cdot 71 + 22.25 \cdot 119 + 27.25 \cdot 175 + 32.25 \cdot 185 + 37.25 \cdot 158 + 42.25 \cdot 122 + 47.25 \cdot 69 + 52.25 \cdot 35 + 57.25 \cdot 11 + 62.25 \cdot 5)/1000 = 32.18$

Mediana este $me = \frac{x_{500} + x_{501}}{2} = \frac{32.25 + 32.25}{2} = 32.25$

Dispersia este $s^2 = 114.4950$.

Abaterea standard este $s = \sqrt{s^2} = 10.7002$

3. Statistica naşterilor înregistrate lunar într-o localitate este prezentată în tabelul următor

Luna	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
Nr. Naşteri	**8**	**9**	**13**	**18**	**15**	**20**	**24**	**19**	**12**	**11**	**6**	**5**

a) Să se reprezinte seria de date cu ajutorul histogramei

b) Să se calculeze indicatorii seriei de date şi să se interpreteze rezultatele

a) Histograma este

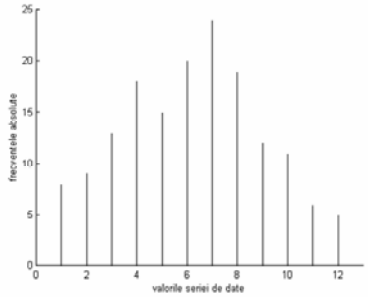

Figura 2.6 Histograma cu bare a seriei de date din exercițiul 3

b) Media este $\bar{n} = 13,(3)$.

Dispersia este $s^2 = 32.7222$.

Abaterea standard este $s = 5.7203$.

Coeficientul de variație este $CV = \frac{s}{n} = 0.4290$

Deoarece coeficientul de variație este mare rezultă că media nu este un indicator reprezentativ pentru seria de date.

4. Frecvența medie a vântului pe direcțiile principale şi secundare ale punctelor cardinale înregistrate la Stația meteorologică Craiova în perioada 1950-2000 este dată în tabelul următor

Direcția	**N**	**NE**	**E**	**SE**	**S**	**SV**	**V**	**NV**
Frecvența (%)	**5**	**10**	**24**	**7**	**5**	**13**	**27**	**9**

a) Să se reprezinte seria de date cu ajutorul histogramei
b) Să se calculeze indicatorii seriei de date şi să se interpreteze rezultatele

a) Histograma este

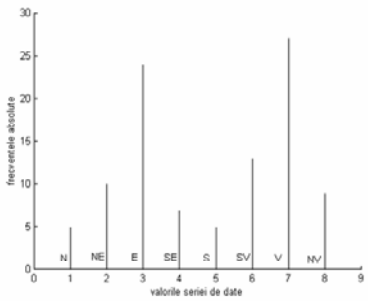

Figura 2.7 Histograma cu segmente a seriei de date din exercițiul 4

b) Media este $\bar{n} = 12.5$

Interpretarea sa este: in medie vântul a bătut din fiecare direcție 12,5% din timp

Dispersia este $s^2 = 63$

Abaterea standard este $s = 7.9362$

Coeficientul de variație este $CV = 0.6349$

Deoarece coeficientul de variație este mare rezultă că media nu este un indicator statistic relevant.

5. Măsurătorile efectuate prin sondaj aleator asupra înălțimii a 50 de spice dintr-un lot de orz indică următoarele valori (în cm.) date în tabelul de mai jos:

Nr. crt	Înălțime	Nr. crt	Înălțime	Nr. crt	Înălțime	Nr. crt	Înălțime	Nr. crt	Înălțime
1	**50,7**	**11**	**50,1**	**21**	**50.0**	**31**	**49,8**	**41**	**49,9**
2	**51,0**	**12**	**50,0**	**22**	**50,0**	**32**	**50,5**	**42**	**50,2**
3	**51,0**	**13**	**50,1**	**23**	**49,9**	**33**	**49,6**	**43**	**49,8**
4	**49,6**	**14**	**50,0**	**24**	**50,2**	**34**	**50,4**	**44**	**49,9**
5	**49,8**	**15**	**49,9**	**25**	**50,0**	**35**	**50,2**	**45**	**50,1**
6	**49,2**	**16**	**50,3**	**26**	**49,7**	**36**	**50,6**	**46**	**50,0**
7	**50,0**	**17**	**50,0**	**27**	**50,3**	**37**	**49,6**	**47**	**49,9**
8	**49,8**	**18**	**50,2**	**28**	**49,2**	**38**	**49,3**	**48**	**49,8**
9	**49,8**	**19**	**49,4**	**29**	**50,0**	**39**	**49,5**	**49**	**50,1**
10	**49,9**	**20**	**49,8**	**30**	**50,1**	**40**	**50,0**	**50**	**50,2**

a). **Să se facă gruparea datelor şi să se determine frecvențele absolute şi relative. Să se facă reprezentarea în batoane.**

b). **Să se reprezinte histograma.**

c). **Să se determine clase de valori de lungime 0.3, să se determine frecvențele absolute ale intervalelor şi să se reprezinte histograma cu bare.**

d). **Să se determine valorile centrale ale claselor, media, valoarea modală, mediana, dispersia şi abaterea medie pătratică.**

a) Distributia de frecvențe a seriei de date este

$$X = \begin{pmatrix} 49.2 & 49.3 & 49.4 & 49.5 & 49.6 & 49.7 & 49.8 & 49.9 & 50 & 50.1 & 50.2 & 50.3 & 50.4 & 50.5 & 50.6 & 50.7 & 50.8 & 51 \\ 1 & 1 & 1 & 1 & 2 & 1 & 6 & 6 & 10 & 5 & 6 & 2 & 1 & 1 & 2 & 1 & 1 & 2 \end{pmatrix}$$

Frecvențele relative sunt date de

$$X_{rel} = \begin{pmatrix} 49.2 & 49.3 & 49.4 & 49.5 & 49.6 & 49.7 & 49.8 & 49.9 & 50.0 & 50.1 & 50.2 & 50.3 & 50.4 & 50.5 & 50.6 & 50.7 & 50.8 & 51 \\ \frac{1}{50} & \frac{1}{50} & \frac{1}{50} & \frac{1}{50} & \frac{2}{50} & \frac{1}{50} & \frac{6}{50} & \frac{6}{50} & \frac{10}{50} & \frac{5}{50} & \frac{6}{50} & \frac{2}{50} & \frac{1}{50} & \frac{1}{50} & \frac{2}{50} & \frac{1}{50} & \frac{1}{50} & \frac{2}{50} \end{pmatrix}$$ b)

Histograma este

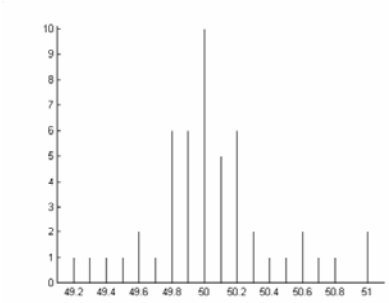

Figura 2.2 Histograma cu segmente a seriei de date din exercitiul 5

c) Clasele sunt date în tabelul următor

Clasa	Frecvența	Valoarea centrală a clasei
[49.2 49.5)	3	49.35
[49.5 49.8)	4	49.65
[49.8 50.1)	22	49.95
[50.1 50.4)	13	50.25
[50.4 50.7)	4	50.55
[50.7 51]	4	50.85

Histograma cu bare este

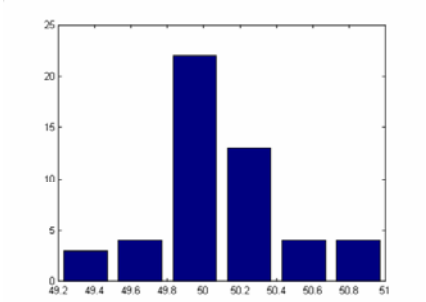

Figure 2.3 Histograma cu bare a grupării de date din exercițiul 5

Distribuția de frecvențe pentru care se calculează indicatorii este

$$X = \begin{pmatrix} 49.35 & 49.65 & 49.95 & 50.25 & 50.55 & 50.85 \\ 3 & 4 & 22 & 13 & 4 & 4 \end{pmatrix}$$

Media este $\overline{X} = 50.0880$.

Modul este $mo = 49.95$.

Mediana este $me = \frac{X(25) + X(26)}{2} = \frac{49.95 + 49.95}{2} = 49.95$.

Dispersia este $s^2 = 0.1264$.

Abaterea medie pătratică este $s = \sqrt{s^2} = \sqrt{0.1264} = 0.3560$

Coeficientul de variație este $CV = 0.0071 \approx 0$, deci datele sunt grupate în jurul valorii medii si media este un indicator relevant.

Exerciții propuse

1. Producția de boabe a 100 de parcele de 6 m^2 cultivate cu un anumit soi de grâu într-un câmp experimental este dată în tabelul:

Nr. crt	Prod.	Nr. crt	Prod.	Nr. crt	Prod.	Nr. crt	Prod.	Nr. crt	Prod
1	2,72	21	2,97	41	3,05	61	3,11	81	3,22
2	2,76	22	2,98	42	3,05	62	3,11	82	3,23
3	2,84	23	2,98	43	3,06	63	3,13	83	3,24
4	2,85	24	2,99	44	3,06	64	3,13	84	3,24
5	2,87	25	2,99	45	3,07	65	3,13	85	3,25
6	2,87	26	3,00	46	3,07	66	3,13	86	3,25

7	2,88	27	3,01	47	3,07	67	3,14	87	3,25
8	2,90	28	3,01	48	3,07	68	3,14	88	3,25
9	2,91	29	3,01	49	3,08	69	3,14	89	3,27
10	2,93	30	3,02	50	3,08	70	3,15	90	3,28
11	2,93	31	3,02	51	3,08	71	3,15	91	3,29
12	2,93	32	3,02	52	3,09	72	3,15	92	3,29
13	2,94	33	3,03	53	3,09	73	3,16	93	3,31
14	2,94	34	3,03	54	3,09	74	3,17	94	3,31
15	2,95	35	3,04	55	3,09	75	3,17	95	3,33
16	2,95	36	3,04	56	3,10	76	3,17	96	3,34
17	2,96	37	3,04	57	3,10	77	3,19	97	3,36
18	2,96	38	3,04	58	3,10	78	3,19	98	3,37
19	2,96	39	3,04	59	3,11	79	3,21	99	3,39
20	2,97	40	3,05	60	3,11	80	3,21	100	3,41

Se cere:

a). Să se facă gruparea datelor pe clase de lungime 0,05, să se întocmească histograma şi să se desenze poligonul frecvenţelor.

b). Să se determine valorile centrale ale claselor, media, valoarea modală, mediana, dispersia şi abaterea medie pătratică.

2. Temperaturile medii înregistrate la Craiova în lunile *mai* ale anilor 1930-1979 sunt date în tabelul de mai jos:

Anul	0	1	2	3	4	5	6	7	8	9
1930…	8,1	4,0	-0,9	3,2	8,2	6,7	8,8	5,6	7,8	4,1
1940…	3,5	6,3	+0,4	4,3	3,8	6,4	6,4	8,2	5,9	0,3
1950…	5,5	6,9	-1,9	5,1	2,1	3,6	0,0	6,2	2,9	6,0
1960…	4,6	8,0	2,3	2,9	3,2	3,7	6,1	6,6	5,5	-0,1
1970…	5,2	3,6	5,5	3,0	4,9	7,7	3,1	7,2	5,8	6,3

a). Să se facă gruparea în clase, de mărime 2^oC cu convenţia ca extremitatea dreaptă a fiecărei clase să nu aparţină clasei (ex. [-2,0;0), [0;2,0), [2,0;4,0), …);

b). Să se completeze tabela obţinută la punctul *a)* cu frecvenţele absolute, cu frecvenţele relative şi cu valoarea centrală a clasei;

c). Să se reprezinte histograma grupării în clase;

d) Să se calculeze indicatorii distribuţiei de frecvenţe a grupării datelor în clase şi să se interpreteze rezultatele.

3. Cantităţile lunare de precipitaţii căzute la Craiova în lunile aprilie ale anilor 1930-1979 sunt date (în litri/m.p.) în tabelul următor

	1930…	1940…	1950…	1960…	1970…
0	55,5	92,0	24,8	39,4	64,4
1	19,6	36,5	40,0	49,4	42,5
2	17,8	33,7	40,8	75,6	16,4
3	7,8	26,9	23,5	33,7	42,6
4	89,0	42,3	52,2	62,6	74,0
5	32,7	35,4	94,3	57,9	43,8
6	22,6	16,3	31,6	65,8	47,1
7	45,3	22,8	65,3	49,5	50,2
8	57,1	37,6	51,4	8,7	31,6
9	28,1	3,9	19,3	31,9	42,7

a). Să se facă gruparea în clase, de mărime 10 litri/mp.

b). Să se completeze tabela obţinută la punctul ***a)*** cu frecvenţele absolute, cu frecvenţele relative şi cu valoarea centrală a clasei;

c). Să se reprezinte histograma grupării în clase şi să se calculeze indicatorii statistici ai grupării în clase şi să se interpreteze rezultatele.

3. Studiul statistic al legăturii dintre fenomene. Aplicații.

Elementele unei populații pot avea diverse caracteristici, fiecare determinând anumite variabile aleatoare *X,Y,...,*acestea având fie un caracter determinist, fie un caracter întâmplător (stochastic) iar între ele putând exista anumite legături.

Legăturile dintre caracteristicile unei populații pot fi foarte strânse, exprimate prin funcții *y=f(x),* numite *funcționale.*

Există însă şi legături în care intervin numeroşi factori sistematici şi accidentali care fac ca două sau mai multe însuşiri (caracteristici) să varieze în strânsă concordanță (nu însă în sens funcțional). Între acestea sunt legăturile dintre fenomenele si procesele economice care apar ca legături statistice (*stochastice)*, a căror particularitate este faptul că rezultatul este determinat ca urmare a influenței unui ansamblu de factori. Legăturile statistice se manifestă, ca tendință valabilă numai la nivelul populației. Dependența de acest tip are caracter întâmplător şi se numeşte *dependență stochastică* sau *corelație.*

In cele ce urmează vom considera fenomene descrise cu ajutorul seriilor de date (exprimate prin numere reale) sistematizate cu ajutorul distribuțiilor de frecvențe (numite şi variabile statistice).

Există două aspecte ale studiului dependenței stochastice între fenomene: analiza de corelație şi analiza de regresie.

Analiza de corelație studiază comportarea fiecărei variabile în funcție de valorile celorlalte variabile, precum şi măsura dependenței dintre variabilele considerate. Se analizează dacă tendința ascendentă a unei variabile implică o tendință ascendentă sau descendentă la cealaltă, sau nici o tendință. Rezultatele se exprimă prin *coeficientul de corelație* sau prin *raportul de corelație.*

Analiza regresiilor constă în determinarea *funcției de regresie* între două variabile. In ipoteza existenței unei legături între variabile se pot prognoza valorile uneia în raport cu valorile celeilalte folosind funcția de regresie.

In paragrafele următoare va fi studiată legatura directă între serii de date (care generează variabilele statistice) care descriu anumite caracteristici ale unei populații.

Pentru simplificare le vom nota

$X : x_{1,}\ x_2, ..., x_n$, respectiv $Y : y_1, y_2, ..., y_n$.

3.1. Analiza corelațiilor

Prin *corelație simplă* se înțelege legătura reciprocă dintre două variabile *X* şi *Y* ale unei populații.

Corelațiile dintre variabile prezintă mare importanță, deoarece cunoscând variația unei însuşiri putem trage concluzii asupra însuşirii sau însuşirilor de care aceasta este legată, fără a recurge la determinări directe.

Corelația poate fi *pozitivă*, atunci când valorile celor două variabile cresc sau descresc în acelaşi timp, sau *negativă,* atunci când valorile unei variabile cresc, iar cele ale celeilalte variabile descresc.

Metodele cele mai simple de constatare a unei corelații sunt *metoda grafică* sau *graficul de corelație (corelograma)* şi *tabela de corelație.*

1.1. Metoda grafică (diagrama de împrăştiere)

Perechile de observații $(x_i,\ y_i)$, $i \in \{1,2,...,n\}$ se reprezintă în planul *Oxy* prin punctele $M_i(x_i,\ y_i)$, $i \in \{1,2,...,n\}$. Se obține « un nor de puncte », numit *corelogramă.* Tendința norului de puncte permite vizualizarea si stabilirea formei analitice a funcției de regresie. Corelograma arată dacă între cele două variabile există o relație şi poate indica şi forma legăturii prin observarea unei densități de puncte care se concentrează în jurul unei anumite curbe, care poate fi liniară sau de altă formă.

Dacă norul de puncte are forma unei elipse alungite există o legătură puternică între variabilele X şi Y . Dacă norul e răspândit în interiorul unui cerc, pătrat variabilele sunt independente.

Exemplul 1: Pentru seriile de date

$X = \{1.2, 0.8, 1, 1.3, 0.7, 0.8, 1.0, 0.6, 0.9, 1.1,\ 0.65, 0.75, 0.85, 0.95, 1.05, 1.1, 1.25\}$

$Y = \{10.1,\ 9.2,\ 11.0,\ 12.0,\ 9.0,\ 8.2,\ 9.35,\ 9.1,\ 10.5,\ 8, 8.5, 9, 8.5, 9.5, 9.5, 10, 11\}$

norul de date este reprezentat in Figura 3.1.

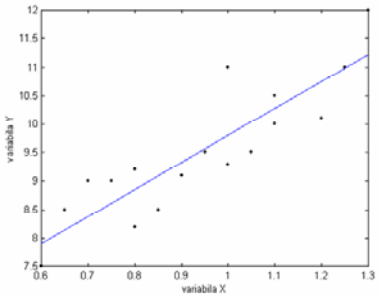

Figura 3.4 Norul de puncte al seriilor de date din Exemplul 1 şi dreapta de corelație

Deoarece forma norului de puncte este apropiată de o elipsă se poate considera că seriile de date sunt puternic corelate.

Exemplul 2 Pentru seriile de date

$X = \{1.2, 0.8, 1, 1.3, 0.7, 0.8, 1.0, 0.6, 0.9, 1.1,\ 0.65, 0.75, 0.85, 0.95, 1.05, 1.1, 1.25\}$

$Y = \{8.1,\ 10.2,\ 7.0,\ 10.0,\ 9.0,\ 7.2,\ 9.3,\ 8.5,\ 8.1,\ 9.5,\ 9.5,\ 8,\ 9.5, 7.5,\ 8.5,\ 10,\ 8\}$

norul de date este reprezentat in Figura 3.2.

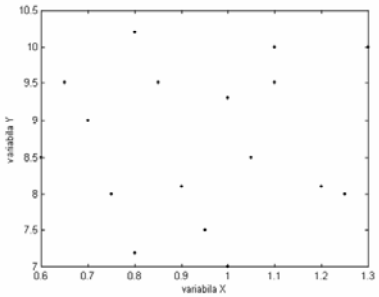

Figura 3.5 Norul de puncte al seriilor de date din Exemplul 2

Faptul că norul de puncte e răspândit în interiorul unui dreptunghi poate fi interpretat ca lipsa unei corelații între cele două variabile.

In cazul probelor cu volum mare de valori observate, pentru cercetarea legăturii dintre variabile se foloseşte *tabelul de corelație* care constă în gruparea pe clase a datelor de observație. In tabelul de corelație termenul $x_{i,j}$ reprezintă numărul de membrii ai populației pentru care variabila X are valoarea x_i iar variabila Y are valoarea y_j

Cu cât valorile individuale din tabelul de corelație sunt mai strâns concentrate în jurul diagonalei cu atât corelația este mai puternică. Cu cât corelația este mai puternică, cu atât valorile din tabelul de corelație sunt mai strâns concentrate în jurul unei diagonale.

Exemplul 3: ([1], pag 284) În tabelul de mai jos sunt trecute datele privind diametrul tulpinii unei plante şi procentul de fibre în funcție de diametru:

y=conținut de fibre %) \ *x= diametrul tulpinii(mm)*	2	3	4	5	6	7	8	
26	2	3	3	2				10
24	4	5	13	7	4			33
22	3	6	18	25	10	2		64
20		1	8	17	18	3		47
18			1	9	8	8	2	28
16				2	3	4	6	15
14						1	2	3
Suma *x*	9	15	43	62	43	18	10	200

Se observă că între cele două caracteristici există corelație pentru că valorile din tabel sunt concentrate în jurul diagonalei secundare. Corelația este negativă deoarece valorilor mai mari ale variabilei X le corespund valori mai mici ale variabilei Y, adică tendința ascendentă a lui X conduce la o tendință descendentă a lui Y.

Aceste observații intuitive reprezintă o informație primară despre corelație, descrierea ei corectă fiind realizată cu ajutorul coeficientului de corelație şi al raportului de corelație.

Pentru seriile de date $X: x_1, x_2, ..., x_n$ şi $Y: y_1, y_2, ..., y_n$ considerăm $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ şi $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Coeficientul de corelație (numit şi coeficientul Pearson) se defineşte prin

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \left(y_i - \bar{y}\right)^2}}.$$

Pentru calcule directe se poate folosi formula

$$r = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\cdot\left(\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}$$

Următoarele observații reprezintă elemente de bază pentru interpretarea coeficientului de corelație

- Coeficientul de corelație r este o mărime adimensională a cărui valoare absolută este subunitară, adică $|r| < 1$.
- Dacă seriile de date X şi Y sunt independente atunci $r = 0$.
- Dacă coeficientul de corelație este nul, seriile statistice nu sunt în mod necesar independente, dar dependența lor nu este liniară, ea putând fi de altă natură.
- Dacă $r \approx -1$ corelația este puternică şi negativă (creşterea valorilor lui X eate asociată cu descreşterea valorilor lui Y)
- Dacă $r \approx +1$ corelația este puternică şi pozitivă (creşterea valorilor lui X eate asociată cu creşterea valorilor lui Y)

Folosirea coeficientului de corelație este recomandabilă îndeosebi atunci când legătura dintre variabile nu se abate mult de la liniaritate, iar populația studiată este de tipul distribuțiilor normale bidimensionale, adică, în cazul când datele studiate aparțin unei distribuții bidimensionale normale şi relația dintre variabile este liniară coeficientul de corelație are un înțeles statistic bine definit. Dimpotrivă, dacă populația pe care o reprezintă datele nu este normală sau dacă din graficul de corelație este evident că relația dintre variabile se abate mult de la liniaritate, coeficientul de corelație r îşi pierde înțelesul său statistic, iar examinarea semnificației sale statistice devine lipsită de sens. Coeficientul empiric de corelație *r* rămâne astfel numai o mărime de calcul şi nu o valoare estimativă.

Pragul de încredere pentru interpretarea coeficientului de corelație este definit prin

$$PI = |r| \cdot \sqrt{n-1}.$$

Se consideră că legătura dintre variabile este sufficient de probabilă dacă $PI \geq 3$ **.**

Pentru prezentarea corelației între două fenomene se procedează astfel

- se realizează diagrama de împrăştiere a norului de puncte şi se observă în mod empiric dacă datele sunt corelate.

- Dacă variabilele sunt corelate şi corelaţia e aproape liniară (norul de puncte se află în interiorul unei elipse alungite) se calculează coeficientul de corelaţie şi pragul de încredere şi se interpretează rezultatele.

Exemplul 4 ([1], pag 288) În urma efectuării a 8 măsurători asupra două caracteristici X şi Y ale unei populaţii, s-au găsit valorile date în tabelul de mai jos:

Proba	1	2	3	4	5	6	7	8
X: x_i	26,9	26,3	23,6	24,8	29,1	19,6	17,9	19,5
Y: y_i	54,0	52,2	55,5	57,1	54,3	63,2	70,1	70,2

Să se determine coeficientul de corelaţie al variabilelor X şi Y.

Norul de puncte corespunzător seriilor de date X şi Y este reprezentat in Figura 3.3.

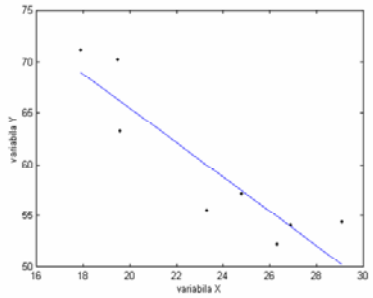

Figura 3.6 Norul de puncte di Exemplul 4 şi dreapta de corelaţie

Configuraţia norului de puncte indică o corelaţie liniară negativă.

Pentru calculul coeficientului de corelaţie aşezăm datele în tabelul de mai jos, pe primele două coloane. Celelalte coloane se completează folosind datele problemei.

x_i (cm)	y_i (%)	x_i^2	y_i^2	$x_i y_i$
26,9	54,0	723,61	2916,00	1452,60
26,3	52,2	691,69	2724,84	1372,86
23,6	55,5	556,96	3080,25	1309,80
24,8	57,1	615,04	3260,41	1416,08
29,1	54,3	846,81	2948,48	1580,13
19,6	63,2	384,16	3994,24	1238,72
17,9	70,1	320,41	4914,01	1254,79
19,5	70,2	380,25	4928,04	1368,90
$\sum x_i$ =187,7	$\sum y_i$ =476,6	$\sum x_i^2$ =4518,93	$\sum y_i^2$ =28766,28	$\sum x_i y_i$ =10993,88
$\overline{x}$ =23,46	$\overline{y}$ =59,57	$\overline{x^2}$ =564,86	$\overline{y^2}$ =3595,78	$\overline{x \cdot y}$ =1374,23

Efectuând calculele necesare obţinem:

$\overline{x} \cdot \overline{y}$ =139751,22 , $\overline{x}^2 = 550{,}3716$, $\overline{y}^2$=3548,5849;

$$s_x^2 = \frac{n}{n-1}\left[\overline{(x^2)} - (\overline{x})^2\right] = \frac{8}{7}(564{,}86 - 550{,}37) = 16{,}56;$$

$$s_y^2 = \frac{n}{n-1}\left[\overline{(y^2)} - (\overline{y})^2\right] = \frac{8}{7}(3595{,}78 - 3548{,}58) = 53{,}94;$$

$$s_{xy} = \frac{n}{n-1}\left(\overline{x \cdot y} - \overline{x} \cdot \overline{y}\right) = \frac{8}{7}(1374{,}23 - 1397{,}51) = -26{,}61;$$

Rezultă $r = \dfrac{s_{xy}}{s_x \cdot s_y}$ =- 0,89, ceea ce indică o corelaţie negativă puternică

Pragul de incredere este $PI = 0.89 * \sqrt{7} = 2.3547 < 3$, deci coe ficientul de corelaţie nu este relevant.

Exemplul 5 (Exemplul 3 continuat) În tabelul de mai jos sunt trecute datele privind diametrul tulpinii unei plante şi procentul de fibre în funcție de diametru:

y=conținut de fibre %) \ *x= diametrul tulpinii(mm)*	2	3	4	5	6	7	8	
26	2	3	3	2				10
24	4	5	13	7	4			33
22	3	6	18	25	10	2		64
20		1	8	17	18	3		47
18			1	9	8	8	2	28
16				2	3	4	6	15
14						1	2	3
Suma *x*	9	15	43	62	43	18	10	200

Cele două variabile statistice pentru care se cere coeficientul de corelație sunt

$X = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 9 & 15 & 43 & 62 & 43 & 18 & 10 \end{pmatrix}$ şi $Y = \begin{pmatrix} 26 & 24 & 22 & 20 & 18 & 16 & 14 \\ 10 & 33 & 64 & 47 & 28 & 15 & 3 \end{pmatrix}$

$$\sum x_i = 2\cdot 9 + 3\cdot 15 + 4\cdot 43 + 5\cdot 62 + 6\cdot 43 + 7\cdot 18 + 8\cdot 10 = 1009$$

$$\sum y_i = 26\cdot 10 + 24\cdot 33 + 22\cdot 64 + 20\cdot 47 + 18\cdot 28 + 16\cdot 15 + 14\cdot 3 = 4186$$

$$\sum x_i^2 = 2^2\cdot 9 + 3^2\cdot 15 + 4^2\cdot 43 + 5^2\cdot 62 + 6^2\cdot 43 + 7^2\cdot 18 + 8^2\cdot 10 = 5479$$

$$\sum y_i^2 = 26^2\cdot 10 + 24^2\cdot 33 + 22^2\cdot 64 + 20^2\cdot 47 + 18^2\cdot 28 + 16^2\cdot 15 + 14^2\cdot 3 = 89044$$

$$\begin{aligned}\sum x_i\cdot y_i = {} & 2\cdot 26\cdot 2 + 3\cdot 26\cdot 3 + 4\cdot 26\cdot 3 + 5\cdot 26\cdot 2 + 2\cdot 24\cdot 4 + 3\cdot 24\cdot 5 + 4\cdot 24\cdot 13 + 5\cdot 24\cdot 7 + 6\cdot 24\cdot 4 \\ & + 2\cdot 22\cdot 3 + 3\cdot 22\cdot 6 + 4\cdot 22\cdot 18 + 5\cdot 22\cdot 25 + 6\cdot 22\cdot 10 + 7\cdot 22\cdot 2 + 3\cdot 20\cdot 1 + 4\cdot 20\cdot 8 + 5\cdot 20\cdot 17 + \\ & + 6\cdot 20\cdot 18 + 7\cdot 20\cdot 3 + 4\cdot 18\cdot 1 + 5\cdot 18\cdot 9 + 6\cdot 18\cdot 8 + 7\cdot 18\cdot 8 + \cdot 18\cdot 2 + 5\cdot 16\cdot 2 + 6\cdot 16\cdot 3 + \\ & + 7\cdot 16\cdot 4 + 8\cdot 16\cdot 6 + 7\cdot 14\cdot 1 + 8\cdot 14\cdot 2 = 20624\end{aligned}$$

Coeficientul de corelație este $r = \dfrac{20624 - \dfrac{1009\cdot 4186}{200}}{\sqrt{\left(5479 - \dfrac{1009^2}{200}\right)\cdot\left(89044 - \dfrac{4186^2}{200}\right)}} = -0.6630$

Coeficientul indică o corelație negativă (confirmând observația intuitivă asupra norului de puncte) puternică (deoarece $|r| > 0.5$).

Pragul de încredere este $PI = 0.6630\cdot\sqrt{199} = 9{,}3527 > 3$, deci coeficientul de corelație este un indicator relevant.

3.2. Analiza regresiilor

In general punctele din norul de puncte asociat seriilor de date nu se găsesc toate pe graficul unei funcții $y = f(x)$, ci sunt mai mult sau mai puțin împrăştiate.

Folosind metoda celor mai mici pătrate se poate determina totusi o funcție față de graficul căreia suma abaterilor valorilor individuale să fie minime. Aceasta este **funcția de regresie.**

Scopul construirii funcției de regresie este prognoza valorilor unei variabile folosind valorile celeilalte variabile.

Regresia liniară

Vom considera cazul cînd punctele corespunzătoare unei serii statistice sunt dispuse aproximativ după o dreaptă, adică variabilele sunt liniar correlate ($r \approx 1$ sau $r \approx -1$). În acest caz legătura cea mai simplă este

cea liniară în care unei creşteri a lui x (care este considerată variabila "predictor") îi corespunde o creştere sau o scădere proporţională a lui y (care este considerată variabila "răspuns").
Această relaţie se numeşte *regresia liniară* şi este dată de ecuaţia

$$y = \alpha \cdot x + \beta.$$

numită *ecuaţia dreptei de regresie.*

Coeficienţii dreptei de regresie se calculează folosind relaţiile

$$\alpha = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\beta = \bar{y} - \alpha \cdot \bar{x} = \frac{1}{n}\cdot\left(\sum_{i=1}^{n} y_i - \alpha \cdot \sum_{i=1}^{n} x_i\right).$$

Regresia liniară se poate folosi dacă sunt îndeplinite următoarele ipoteze:

- valorile variabilei dependente Y trebuie să aibă o repartiţie normală
- Y şi X trebuie să aibă dispersia (sau abaterea standard) asemănătoare
- Legătura dintre variabile trebuie să fie liniară (verificare empirică, pe baza norului de puncte care trebuie să aibă o formă alungită)

Din ecuaţie de regresie se pot determina valorile lui Y dacă se ştiu valorile lui X.
Estimatorul dispersiei lui Y în jurul dreptei de regresie este

$$s^2 = \frac{1}{n-2}\cdot\left[\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\right]$$

Exemplul 6 Producţia de struguri obţinută într-o fermă în mai mulţi ani şi numărul de zile însorite observate de-a lungul anilor sunt înregistrate în tabelul următor. Pe baza datelor din table să se precizeze dacă cele două serii de date sunt corelate.

Producţia de struguri/ha x_i	Numărul de zile însorite y_i	$x_i^{\,2}$	$y_i^{\,2}$	$x_i \cdot y_i$
1.2	101	1.44	10201	1.2120
0.8	92	0.64	8464	0.7360
1	110	1.0	12100	1.1000
1.3	120	1.69	14400	1.5600
0.7	90	0.49	8100	0.6300
0.8	82	0.64	6724	0.6560
1.0	93	1.0	8649	0.9300
0.6	75	0.36	5625	0.4500
0.9	91	0.81	8281	0.8190
1.1	105	1.21	11025	1.1550

$\sum x_i = 9.4$	$\sum y_i = 959$	$\sum x_i^2 = 9.28$	$\sum y_i^2 = 93569$	$\sum x_i \cdot y_i = 924.80$

Norul de puncte corespunzător seriilor de date este prezentat în figura 3.4.

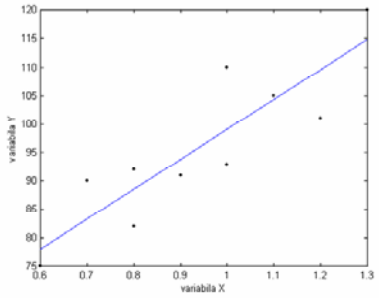

Figura 3.7 Norul de puncte şi dreapta de regresie a serieiilor de date din Exemplul 6

Coeficientul de corelație este

$$r = \frac{924.80 - \frac{9.4 \cdot 959}{10}}{\sqrt{\left(9.28 - \frac{9.4^2}{10}\right) \cdot \left(93569 - \frac{959^2}{10}\right)}} = \frac{23.34}{\sqrt{0.444 \cdot 1600.9}} = 0.8754 .$$

Coeficienții dreptei de regresie sunt dați de

$$\alpha = \frac{924.80 - \frac{9.4 \cdot 959}{10}}{924.80 - \frac{9.4 \cdot 959}{10}} = 118.39 \text{ şi } \beta = 959 - 118.39 \cdot 9.4 = -153.86 .$$

Dreapta de regresie (desenată în figura 4) are ecuația

$$Y = 118.39 \cdot X - 153.86 .$$

Indicatorul dispersiei lui Y în jurul dreptei de regresie este

$$s^2 = 1548.3224 .$$

Interpretarea rezultatului:

- coeficientul de corelație este pozitiv, deci o tendință ascendentă a variabilei *"x"* antrenează o tendință ascendemtă a variabilei *"y"*
- coeficientul de corelație este apropiat de 1, deci corelația este puternică.
- Pragul de încredere este $PI = r \cdot \sqrt{10-1} = 2.6262 < 3$, deci numărul de date nu este sufficient de mare pentru a asigura faptul ca e semnificativ coeficientul de corelație. Aceasta observație este confirmat de faptul că este mare coeficientul de dispersie al lui Y în jurul dreptei de regresie.
- Dreapta de regresie se va folosi cu precauție pentru prognoze, deoarece nu reprezintă o estimare precisă a dependenței dintre seriile de date.

Exemplul 7 Cantitatea de nutreț folosit şi numărul de animale crescute în 14 ferme sunt prezentate în tabelul următor. Pr baza datelor din table să se precizeze dacă există corelații între cele două aspecte ale activității fermei.

	Cantitatea de nutreț x_i	Numărul de animale y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	6.4	380	42.25	144400	2470.0
2	5.2	200	27.04	40000	1040.0
3	0.4	15	0.16	225	6.0
4	1.7	50	2.89	2500	85.0
5	1.9	40	3.61	1600	76.0
6	2.4	40	5.76	1600	96.0
7	3.2	41	10.24	1681	131.2
8	4.7	18	22.09	324	84.6

9	10.1	210	102.01	44100	2121.0
10	12.5	190	156.25	36100	2375.0
11	13.1	200	171.61	40000	2620.0
12	5.5	55	30.25	3025	302.5
13	2.5	38	6.25	1444	95.0
14	1.5	20	2.25	400	30.0
	$\sum x_i = 71.2$	$\sum y_i = 1497$	$\sum x_i^2 = 582.66$	$\sum y_i^2 = 317399$	$\sum x_i \cdot y_i = 11532.3$

In Figura 3.5 este prezentat norul de puncte asociat seriilor de date şi dreapta de regresie.

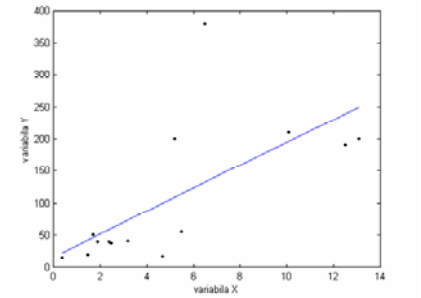

Figura 3. 8 Norul de puncte şi dreapta de regresie asocate seriilor de date din Exemplul 7

Coeficientul de corelație este

$$r = \frac{11532.3 - \frac{71.2 \cdot 1497}{14}}{\sqrt{\left(585{,}66 - \frac{71.2^2}{14}\right) \cdot \left(317339 - \frac{1497^2}{14}\right)}} = 0.6653$$

Dreapta de regresie (desenată în figura 5) are ecuația

$$Y = 17.7685 \cdot X + 16.5626 .$$

Interpretarea rezultatelor

- coeficientul de corelație este pozitiv, deci o tendință ascendentă a variabilei "X" antrenează o tendință ascendemtă a variabilei "Y"
- coeficientul de corelație nu este apropiat de 0, deci deci cele două variabile ar putea fi corelate.
- pragul de încredere este $PI = r \cdot \sqrt{14-1} = 2.3984 < 3$, deci numărul de date nu este sufficient de mare pentru a asigura faptul ca e semnificativ coeficientul de corelație.
- Dreapta de regresie se va folosi cu precauție pentru prognoze, deoarece nu reprezintă o estimare precisă a dependenței dintre seriile de date.

Exemplul 8 L_E reprezintă limita de elasticitate a tulpinei unei plante iar L_R reprezintă limita sa de ruptură. Stiind că raportul $X = \frac{L_E}{L_R}$ este strâns legat de con este strâns legat de conținutul în fibre al tulpinii, notat Y, să se analizeze corelația obținută între cei doi parametrii pe un eşantion de 79 plante, date prezentate în tabelul de mai jos. Numerele întregi din interiorul tabelului reprezintă frecvența de apariție în cele 79 probe a perechilor (X, Y) corespunzătoare.

X\Y	0.5	0.6	0.7	0.8	0.9	Distribuție marginală pentru X
0.5	0	2	0	0	8	10
0.6	0	4	2	9	0	15
0.7	2	12	3	1	0	18
0.8	21	14	0	0	0	35
0.9	1	0	0	0	0	1
Distribuție marginală pentru Y	24	32	5	10	8	79

Variabilele statistice pentru care se studiază corelația sunt

$X = \begin{pmatrix} 0.5 & 0.6 & 0.7 & 0.8 & 0.9 \\ 10 & 15 & 18 & 35 & 1 \end{pmatrix}$ şi $Y = \begin{pmatrix} 0.5 & 0.6 & 0.7 & 0.8 & 0.9 \\ 24 & 32 & 5 & 10 & 8 \end{pmatrix}$, pentru care tabelul interdependenţelor este prezentat anterior.

Pentru calculul coeficientului de corelaţie sunt necesare următoarele rezultate:

$$\sum x_i = 10 \cdot 0.5 + 15 \cdot 0.6 + 18 \cdot 0.7 + 35 \cdot 0.8 + 1 \cdot 0.9 = 49.9$$

$$\sum y_i = 24 \cdot 0.5 + 32 \cdot 0.6 + 5 \cdot 0.7 + 10 \cdot 0.8 + 8 \cdot 0.9 = 55.5$$

$$\sum y_i^2 = 24 \cdot 0.5^2 + 32 \cdot 0.6^2 + 5 \cdot 0.7^2 + 10 \cdot 0.8^2 + 8 \cdot 0.9^2 = 39.93$$

$$\sum x_i \cdot y_i = 0.5 \cdot 0.6 \cdot 2 + 0.5 \cdot 0.9 \cdot 8 + 0.6 \cdot 0.6 \cdot 4 + 0.6 \cdot 0.7 \cdot 2 + 0.6 \cdot 0.8 \cdot 9 + 0.7 \cdot 0.5 \cdot 2 + 0.7 \cdot 0.6 \cdot 12 + 0.7 \cdot 0.7 \cdot 3 + 0.7 \cdot 0.8 \cdot 1 + 0.8 \cdot 0.5 \cdot 21 + 0.8 \cdot 0.6 \cdot 14 + 0.9 \cdot 0.5 \cdot 1 = 34.1400$$

Coeficientul de corelaţie este $r = -0.8194$

Interpretarea rezultatelor:

- coeficientul de corelaţie nu este apropiat de 0, deci variabilele sunt corelate.
- coeficientul de corelaţie este negativ, deci valorilor mari ale lui X le corespund multe valori mici ale lui Y (se confirmă prinpoziţionarea datelor în table)
- pragul de încredere este $PI = 0.8194 \cdot \sqrt{77} = 7.1902 > 3$, desi indicatorul de corelaţie este semnificativ şi poate fi folosit în studiul corelaţiei dintre variabilele X şi Y
- în acest caz dreapta de regresie dă informaţii semnificative asupra valorilor lui Y, dacă se cunosc valorile lui X.

Dreapta de corelaţie are ecuaţia

$$Y = -0.6885 \cdot X + 17.9713 .$$

Folosind această ecuaţie de regresie putem determina (aproximativ) valorile lui Y.

De exemplu, dacă $X = 0.75$ rezultă că $Y = -0.6885 \cdot 0.75 + 17.9713 = 17.4549$.

Exerciţii propuse

1. Pentru a stabili în ce măsură depinde producţia de tulpini de perioada de vegetaţie a diferitelor soiuri de cânepă de fibre, s-au realizat observaţii asupra cinci soiuri de cânepă foarte diferite ca perioadă de vegetaţie. Datele sunt prezentate în tabelul următor (pe orizontală este prezentată perioada de vegetaţie –în zile- şi pe vericală este prezentată producţia de tulpini – în q/ha) pentru cinci ani de producţie.

	55 zile	70 zile	85 zile	100 zile	115 zile	130 zile
An 1	12	18	25	39	48	64
An 2	10	20	27	36	46	57
An 3	14	24	30	34	44	66
An 4	15	22	29	40	54	59
An 5	13	19	26	37	52	62

Să se precizeze dacă cele două caracteristici ale producţiei (perioada de vegetaţie şi producţia obţinută) sunt corelate. $X = 90$ zile

Pe baza ecuaţiei de regresie să se precizez valoarea aproximativă a producţiei Y (per ha) dacă perioada de vegetaţie ar fi $X = 90$ zile.

Pentru stabilirea acţiunii azotului asupra conţinutului de fibre din tulpinile plantelor de cânepă au fost efectuate măsurători în patru ani consecutiv la patru ferme asupra plantelor produse. Rezultatele sunt prezentate în tabelul următor. Pe orizontală este prezentă cantitatea de sulfat de amoniu folosită ca îngrăşământ în cele sinci ferme (în kg/ha) iar pe verticală este prezentat conţinutul de fibre din tulpini (în procente). Pe baza datelor prezentate în tabel să se precizeze dacă între cantitatea de îngrăşământ folosită şi conţinutul de fibre ale tulpinelor există corelaţie.

	Ferma1: 0	Ferma 2: 150	Ferma 3: 300	Ferma 4: 450
An 1	19.0	21.8	22.1	21.8
An 2	18.1	22.5	23.0	22.7
An3	18.9	20.6	22.6	22.4
An 4	19.8	22.0	23.1	20.8

4. Elemente de teoria probabilităţilor aplicate în biologie şi agricultură

4.1. Elemente de analiză combinatorică

Analiza combinatorică se ocupă cu numărarea anumitor grupări ce se pot realiza cu elementele unei mulţimi finite.

Prin cardinalul unei mulţimi finite $A=\{a_1,\ a_2,\ ...,a_n\}$ se înţelege numărul "n" al elementelor sale. Se notează $card(A)=n$.

O grupare care permută elementele mulţimii A este formată din toate elementele muţimii. Două permutări diferă prin ordinea în care sunt scrise elementele. Din punct de vedere matematic, o **permutare** a mulţimii A este o bijecţie de la A la A.

Numărul permutărilor lui A este

$$P_n = 1\cdot 2\cdot 3\cdot ....\cdot n \overset{notatie}{=} n!.$$

O submulţime ordonată de k elemente ale lui A se numeşte **aranjament de ordin** k. Numărul de aranjamente de ordin k ale unei mulţimi cu n elemente este

$$A_n^k = n\cdot(n-1)\cdot ...\cdot(n-k+1)=\frac{n!}{(n-k)!}.$$

El reprezintă numărul aplicaţiilor injective ale mulţimii $\{1,\ 2,\ ...,k\}$ în A.

Submulţimile de câte k elemente ale lui A care nu sunt ordonate se numesc **combinări de ordin** k**.** Numărul acestor combinări este

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n!}{k!\cdot(n-k)!}.$$

Principalele proprietăţi ale combinărilor sunt:

1. $\sum_{k=0}^{n} C_n^k = 2^n$
2. $C_n^k = C_n^{n-k}$
3. $C_n^k = C_{n-1}^k + C_{n-1}^{k-1}$ pentru orice $1\le k\le n-1$ (formula lui Pascal)
4. $(a+b)^n = \sum_{k=0}^{n} C_n^k\cdot a^k\cdot b^{n-k}$ (binomul lui Newton)

Observaţie C_n^k se mai numeşte şi **coeficient binomial**, datorită formulei de dezvoltare a binomului lui Newton .

Dacă $n_1,\ n_2,\ ...n_k$ sunt numere naturale şi $n_1+n_2+...+n_k=n$ se defineste **coeficientul multinomia**l prin

$$C_n^{n_1,n_2,...,n_k} = \frac{n!}{n_1!\cdot n_2!\cdot ...\cdot n_k!}.$$

El are următoarea interpretare: Dacă mulţimea A conţine n elemente, atunci există $C_n^{n_1,n_2,...,n_k}$ partiţii ordonate diferite $\{A_1,\ A_2,\ ...,\ A_k\}$ ale lui A astfel încât fiecare A_i să conţină n_i elemente, $i=1,\ 2,\ ...,k$.

$C_n^{n_1,n_2,...,n_k}$ se numeşte coefficient multinomial pentru că are loc relaţia

$$(x_1+x_2+...+x_k)^n = \sum_{n_1+...n_k=n} C_n^{n_1,n_2,...,n_k}\cdot {x_1}^{n_1}\cdot {x_2}^{n_2}\cdot ...\cdot {x_k}^{n_k}.$$

In încheiere amintim principiul (regula) produsului:

Dacă o operaţiune O_1 poate fi efectuată în n_1 moduri diferite, operaţiunea O_2 poate fi executată în n_2 moduri diferite, etc., operaţiunea O_k poate fi executată în n_k moduri diferite, atunci cele k operaţiuni pot fi executate una după alta în $n_1\cdot n_2\cdot ....\cdot n_k$ moduri diferite.

4.2. Introducere euristică în teoria probabilităţilor

In cele ce urmează prezentăm într-o formă simplă noţiuni de bază ale teoriei probabilităţilor, pornind de la definiţia euristică a probabilităţii de realizare a unui eveniment.

Datele cu care operează teoria probabilităţilor sunt obţinute prin observaţii asupra evenimentelor necontrolate din natură, societate, fie ca rezultat al experimentelor controlate.

Noţiunile primare în teoria probabilităţilor sunt cele de eveniment într-un experiment aleator şi de probabilitate a evenimentului.

Definim un experiment ca fiind procesul prin care efectuăm o observaţie sau o măsurătoare.

Experienţele care pot avea rezultate diferite în funcţie de o serie de circumstanţe întâmplătoare şi rezultatele nu pot fi cunoscute înainea realizării experimentului se numesc experienţe aleatoare.

Rezultatul unui experiment aleator se numeşte **realizare**. Colecţia tuturor realizărilor acoperă orice posibilitate (adică este exhaustivă) şi nici o realizare nu se suprapune peste alta (realizările sunt exclusive).

O colecţie de realizări se numeşte **eveniment,** iar mulţimea tuturor realizărilor formează evenimentul sigur. **Evenimentul sigur** se produce cu certitudine la orice efectuare a experimentului.

Evenimentul care nu se produce ori de câte ori repetăm experienţa se numeşte **eveniment imposibil**.

Evenimentul sigur va fi notat cu *X*, evenimentul imposibil cu Ø, iar evenimentele particulare cu $A, B, C, ...$

Evenimentele compuse se obţin folosind operaţii cu evenimentele simple:

- evenimentul $A \cup B$ se realizează dacă se realizează A sau se realizează B.
- evenimentul $A \cap B$ se realizează dacă se realizează şi A şi B.
- evenimentul $A - B$ se realizează dacă se realizează A şi nu se realizează B.

Unui eveniment A în corespunde evenimentul contrar, notat $C_X(A)$, a cărui producere înseamnă nerealizarea lui A.

Analogia între evenimentele compuse şi teoria mulţimilor este evidentă, un eveniment fiind asociat unei submulţimi a lui X.

Probabilitatea unui eveniment A, notată $P(A) \in [0,1]$ reprezintă şansa pe care o are evenimentul de a se produce.

Dacă experimentul aleator are un număr finit de realizări şi acestea sunt egal probabile (adică nu există un motiv ca o realizare să se producă mai frecvent decât alta) atunci se defineşte probabilitatea unui eveniment ca raportul dintre numarul cazurilor favorabile şi numărul cazurilor posibile, adică

$$P(A) = \frac{numar\ \ cazuri\ \ favorabile\ \ lui\ \ A}{numar\ \ cazuri\ \ posibile}.$$

Observaţie: Dacă experimental aleator are un număr finit de realizări ce nu sunt egal probabile, nu există o modalitate teoretică ce permite calculul probabilităţii cu acurateţe absolută.

Exemplu: Experimentul aleator clasic este aruncarea unui zar cubic, realizat din material omogen . Realizările posibile ale experimentului sunt apariţia feţei cu numărul 1, 2, 3, 4, 5, 6. Evenimentele de apariţie al feţi cu nr "k" se numesc evenimente elementare. Evenimentul sigur este apariţia unei feţe şi este asociat mulţimii $X = \{1,2,3,4,5,6\}$. Alte evenimente sunt reprezentate simbolic prin mulţimi. Spre exemplu apariţia unei feţe pare este reprezentată de mulţimea $A = \{2,4,6\}$.

Probabilitatea de realizarea a lui A este $P(A)=\frac{3}{6}$.

Dacă zarul nu e cubic sau nu este bine centrat, atunci probabilitatea de apariție a unei fețe nu este 1/6.

In unele situații realizarea unui eveniment este condiționată de realizarea prealabilă a altui eveniment. Ideea care conduce la definiția **probabilității condiționate** este următoarea: știm că evenimentul B s-a produs, deci cazurile posibile pentru $A\cap B$ sunt cazurile favorabile pentru B, adică

$$P_B(A)=\frac{nr\ cazuri\ favorabile\ si\ pentru\ A\ si\ pentru\ B}{nr\ cazuri\ favorabile\ pentru\ B}=\frac{\frac{nr\ cazuri\ favorabile\ si\ pentru\ A\ si\ pentru\ B}{nr\ cazuri\ posibile}}{\frac{nr\ cazuri\ favorabile\ pentru\ B}{nr\ cazuri\ posibile}}$$

Probabilitatea unui eveniment A, condiționată de evenimentul B, cu $P(B)\neq 0$, se definește prin

$$P_B(A)=\frac{P(A\cap B)}{P(B)}$$

Două evenimente se numesc independente dacă $P(A\cap B)=P(A)\cdot P(B)$.

Dacă două evenimente sunt independente atunci realizarea unuia nu influențează realizarea celuilalt eveniment, adică $P_B(A)=P(A)$

Pornind de la definiția probabilității, se pot demonstra următoarele proprietăți:

Propoziția 1.

1. $0\le P(A)\le 1$, $P(X)=1$ **şi** $P(\emptyset)=0$

2. $P(C_X(A))=1-P(A)$

3. Dacă $A\subset B$ **atunci** $P(A)\le P(B)$

4. Dacă A **şi** B **sunt două evenimente şi** $A\cap B=\emptyset$**, atunci** $P(A\cup B)=P(A)+P(B)$

5. Dacă A **şi** B **sunt două evenimente, atunci** $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

6. Dacă $A_1\cup A_2\cup\ldots\cup A_n=X$ **şi** $A_i\cap A_j=\emptyset$ **pentru** $i\neq j$ **atunci**

$$P(B)=\sum_{i=1}^{n}P(A_i)\cdot P_{A_i}(B)$$ **(formula probabilității totale)**

$$P_B(A_k)=\frac{P(A_k)\cdot P_{A_k}(B)}{\sum_{i=1}^{n}P(A_i)\cdot P_{A_i}(B)}$$ **(formula lui Bayes)**

4.3 Aplicații în biologie

4.3.1 Intr-un organism există genotipurile AA, Aa, aa. Părinții transmit către urmaşi fiecare câte o singură genă. Se presupune că populația parentală este suficient de mare încât încrucişarea să se facă la întâmplare şi că proporțiile genotipurilor sunt respectiv α, 2β, respectiv γ, cu $\alpha>0$, $\beta>0$, $\gamma>0$ şi $\alpha+2\beta+\gamma=1$. De asemenea se presupune că probabilitatea ca un părinte să transmită o genă este 1/2.
Să se precizeze proporțiile genotipurilor după prima generație şi după a doua generație. Să se interpreteze rezultatele.

In prima generație pot să apară tipurile AA, Aa, aa. Pentru fiecare tip tabelul de calcul al probabilităților este prezentat mai jos.

a) pentru tipul AA

Pentru transmiterea genotipului AA este obligatoriu ca cel puțin o genă A să apară în genotipul fiecărui părinte.

Tipul mascul	Tipul femel	Probabilitatea formării cuplului	Probabilitatea transmiterii genotipuluiAA	Probabilitatea existenței genotipului AA la urmaşi
AA	**AA**	$\alpha \cdot \alpha$	$1 \cdot 1 = 1$	α^2
AA	**Aa**	$\alpha \cdot 2\beta$	$1 \cdot \frac{1}{2} = \frac{1}{2}$	$2\alpha\beta$
aA	**AA**	$\alpha \cdot 2\beta$	$1 \cdot \frac{1}{2} = \frac{1}{2}$	$2\alpha\beta$
aa	**Aa**	$2\beta \cdot 2\beta$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	β^2

Deci probabilitatea existenței unor urmaşi de tipul AA la prima generație este

$$P_1(AA) = \alpha^2 + \alpha\beta + \alpha\beta + \beta^2 = (\alpha + \beta)^2 \qquad \textbf{(1)}$$

b) Raționând la fel se obține probabilitatea existenței unor urmaşi de tip aa la prima generație este

$$P_1(aa) = (\beta + \gamma)^2 \qquad \textbf{(2)}$$

c) pentru tipul generic Aa (sau aA) tabelul probabilităților este

Tipul mascul	Tipul femel	Probabilitatea formării cuplului	Probabilitatea transmiterii genotipului Aa	Probabilitatea existenței genotipului Aa la urmaşi
AA	**Aa**	$\alpha \cdot 2\beta$	$1 \cdot \frac{1}{2} = \frac{1}{2}$	$\alpha\beta$
Aa	**AA**	$2\beta \cdot \alpha$	$\frac{1}{2} \cdot 1 = \frac{1}{2}$	$\alpha\beta$
AA	**aa**	$\alpha \cdot \gamma$	$1 \cdot 1 = 1$	$\alpha\gamma$
aa	**AA**	$\gamma \cdot \alpha$	$1 \cdot 1 = 1$	$\alpha\gamma$
Aa	**aa**	$2\beta \cdot \gamma$	$\frac{1}{2} \cdot 1 = \frac{1}{2}$	$\beta\gamma$
aa	**Aa**	$\gamma \cdot 2\beta$	$1 \cdot \frac{1}{2} = \frac{1}{2}$	$\beta\gamma$
Aa	**Aa**	$2\beta \cdot 2\beta$	$2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$	$2\beta^2$

Probabilitatea existenței unor urmaşi de tip Aa la prima generație este

$$P_1(Aa) = 2\alpha\beta + 2\alpha\gamma + 2\beta\gamma + 2\beta^2 = 2(\alpha + \beta) \cdot (\beta + \gamma) \qquad \textbf{(3)}$$

Faptul că rezultatele sunt corecte este reflectat şşi de relația

$$P_1(AA) + P_1(Aa) + P_1(aa) = (\alpha + \beta)^2 + 2(\alpha + \beta) \cdot (\beta + \gamma) + (\beta + \gamma^2) = (\alpha + 2\beta + \gamma)^2 = 1$$

La a doua generație probabilitățile vor fi:

$$P_2(AA) = \left[(\alpha + \beta)^2 + (\alpha + \beta)(\beta + \gamma)\right]^2 = (\alpha + \beta)^2 (\alpha + 2\beta + \gamma)^2 = (\alpha + \beta)^2 = P_1(AA)$$

$$P_2(aa) = \left[(\beta + \gamma)^2 + (\alpha + \beta)(\beta + \gamma)\right]^2 = (\beta + \gamma)^2 (\alpha + 2\beta + \gamma)^2 = (\beta + \gamma)^2 = P_1(aa)$$

$$P_2(Aa) = 2(\alpha + \beta) \cdot (\beta + \gamma) = P_1(Aa)$$

Prin urmare, de la prima generație încolo probabilitățile de menținere a genotipurilor sunt aceleaşi. Se spune că procesul evolutiv este stochastic stabil.

Bibliografie

1. Bălan V., Matematici Superioare Aplicate, Editura Universitaria, Craiova, 2007
2. Petrişor E., Probabilităţi şi statistică, Editura Politehnica, Timişoara 2005
3. Cristea M., Genetica ecologică şi evoluţia, Editura Ceres, Bucureşti, 1991
4. Ştefănescu D.T.,Călin G., Genetica şi cancerul : (Elemente de genetică şi patologie moleculară), Editura Didactică şi Pedagogică Bucureşti, 1996
5. Raicu P. (coordonator), Biologie : Genetică şi evoluţionism : Manual pentru clasa a XII-a, Editura Didactică şi Pedagogică Bucureşti, 1998
6. Biji E. M. (coordonator), Statistica managerială a agentului economic din agricultură, Editura Ceres, Bucureşti, 1998
7. Howitt D. Cramer D. Introducere în SPSS, Editura Polirom, 2006