Lectia 8
Statistică descriptivă

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 caracteristiclor 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 corelației între fenomene

Statistica inferențială are ca obiect de studiu investigarea prin sondaj: din întreaga populație se selectează un eşantion reprezentativ asupra căruia se fac măsurători sau observații legate de o anumită caracteristică a populației. Pe baza rezultatelor obținute se fac *inferențe statistice* (adică se formulează concluzii) asupra parametrilor populației. Statistica inferențială foloseşte deci informația rezultată din studierea unui eşantion pentru a obține concluzii referitoare la întraga populație din care a fost selectat eşantionul. Aceste concluzii nu sunt de tip determinist ci se obțin folosind metode şi tehnici ale teoriei probabilităților, teorie ce conține 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 condițiilor cercetării. In funcție de tipul de analiză folosit (descriptivă sau inferențială) se foloseşte întreaga populației sau doar un eşantion.
- *procesarea datelor* înseamnă cuantificarea lor numerică şi obținerea 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 obțină rezultate nesemnificative statistic atunci când ipotezele de lucru sau condițiile de aplicare a metodelor nu sunt îndeplinite.
-*interpretarea rezultatelor* este diferită în statistica descriptivă şi în cea inferențială. In primul caz se obțin informații concrete şi clare despre populația studiată, în al doilea caz validarea rezultatelor obținute este realizată prin compararea cu ce se ştia sau se bănuia în domeniul respective. In unele situații analiza statistică dezvăluie corelații 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 informație 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/)

8.1 Serii de date şi distribuții de frecvențe

Considerăm o populație 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 funcție $X : \{a,b,c,d,e\} \to \{0,1,2\}$, unde $X(a) = 0$, $X(b) = 1$, $X(c) = 0$, $X(d) = 0$, $X(e) = 2$.
In acest caz populația este $P = \{a,b,c,d,e\}$. Deoarece identitatea indivizilor din populație nu este interesantă din punct de vedere statistic, aceasta este neglijată în etapele următoare.

Definiție: Distribuția de frecvențe (sau variabila statistică) asociată caracteristicii C a populației 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 populației caracterizați de valoarea x_j.
Numărul n_j se numeşte frecvența absolută de apariție a valorii x_j.
Observații: 1. Din definiția frecvențelor relative rezultă că

$$\sum_{j=1}^{k} n_j = N .$$

2. Unei caracteristici i se poate asocia şi **distribuția frecvențelor 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 $X_r = \begin{pmatrix} 0 & 1 & 2 & 3 & 5 \\ 1/8 & 1/8 & 3/8 & 2/8 & 1/8 \end{pmatrix}$

8.2. Reprezentari 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
- reprezentarea polară

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 coordinate $(x_i,0)$ 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

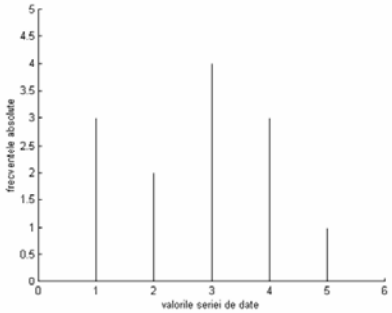


Figure 1 Histograma cu segmente

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 $x_{\max}$ aseriei de date
- se divide segmental $[x_{\min}, x_{\max}]$ 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 aparțin fiecărui interval $I_j = [x_j, x_{j+1})$. Acest număr, notat n_j, se numeşte frecvența clasei I_j.
- Deasupra fiecărui interval I_j se trasează un dreptunghi cu baza I_j şi înălțimea proporțională cu n_j. Pentru determinarea înăltimii 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 distribuției de frecvențe, pentru că ilustrează modul în care sunt distribuite datele.

Un exemplu de histogramă cu bare este dat in Figura 2

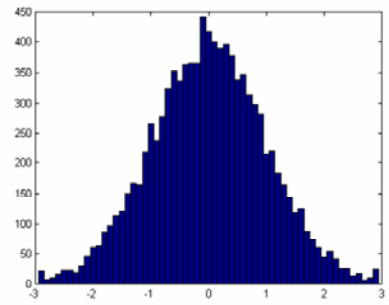


Figure 2 Histograma cu bare

O problemă legată de generarea histogramelor este legată de precizarea numărului de intervale de diviziune. In perioada de început a statisticii computaționale numărul de intervale era proporțional cu $\sqrt{N}$. In unele programme statistice el este ales proporțional 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 obține unind vârfurile segmentelor verticale în cazul reprezentării cu segmente. In cazul reprezentării din Figura 1, poligonul de frecvențe, A, B, C, D, E este dat în figura 3.

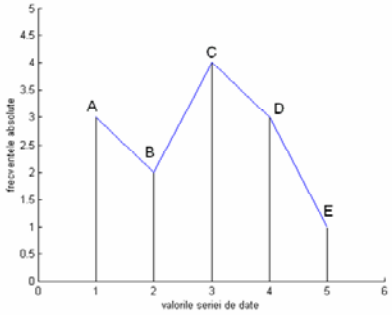


Figure 3 Poligon de frecvențe

d) Reprezentarea cu sectoare circulare este folosită pentru obținerea rapidă a unei viziuni globale asupra importanței 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 (frecvențele absolute ale claselor)
- pe un cerc se consideră sectoare circulare proporționale cu frecvențele fiecărei clase. Unghiul la centru corespunzător clasei cu frecvența 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 direcții).

Ea se construieşte astfel: pe semidrepte cu aceeaşi origine şi care impart planul într-un număr de sectoare egale (acest număr se stabileşte în funcție de caracterul seriei statistice) se consideră segmente ce pornesc din origine, proporționale cu frecvențele absolute ale claselor şi se unesc extremitățile acestoe segmente. Se obține un poligon închis în care clasele cu frecvență mai mare sunt reprezentate prin vârfuri aflate la distanță mai mare față de origine.

8.3. Indicatori statistici

Caracterizarea distribuților de frecvențe se face cu ajutorul unor indicatori.

8.3.1. *Indicatori de poziție (de nivel, de localizare)*

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

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

c) **media geometrică** $x_g = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot .... \cdot x_k^{n_k}}$ este folosită pentru calculu; ritmului mediu de creştere şi determinarea procentului mediu)

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

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

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.

8.3.2. Indicatorii variației (împrăştierii)

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** $e_X = \frac{1}{k} \sum_{j=1}^{k} n_j \left| x_j - \bar{x} \right|$

c) **varianța (dispersia)** $s^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left(x_j - \bar{x} \right)^2$

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

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

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

Regula empirică

Dacă seria de date *X* are media $\bar{x}$ **şi abaterea standard *s* atunci o proporție de cel puțin** $1-\frac{1}{k^2}$ **dintre valorile seriei aparțin intervalului** $\left(\bar{x}-k\cdot s, \bar{x}+k\cdot s\right)$**, pentru** $k>1$

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

Cu cât coeficientul de variație e mai aproape de 0, cu atât seria este mai omogenă ssi media este mai reprezentativă. Dacă este mai apropiat de 1, împrăştierea valorilor este mare şi media nu este un indicator reprezentativ.
In analizele financiare el este o măsură a riscului relativ.

8.3.3 Coeficienți de formă a graficului repartiției frecvențelor

Pentru distribuția de frecvențe $X=\begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_k \\ n_1 & n_2 & n_3 & \dots & n_k \end{pmatrix}$, unde $\sum_{i=1}^{k} n_i = n$ se consideră

media $\bar{x}$, mediana me şi valoarea modală Mo

O repartiție este simetrică dacă $\bar{x} = me = Mo$

Indicele de asimetrie Pearson este $A_s = \frac{\bar{x} - Mo}{s}$

Daca $A_s > 0$, adică $\bar{x} > Mo$, asimetria este de stânga (pozitivă)

Daca $A_s < 0$, adică $\bar{x} < Mo$, asimetria este de dreapta (negativă)

Daca $A_s = 0$, adică $\bar{x} = Mo$, distribuția este simetrică

Exerciţii

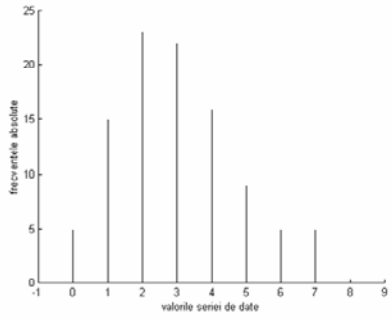
1 Seful departamentului vânzări al unui magazine a înregistrat nivelul cererii zilnice pentru un produs în decursul a 100 zile consecutive. Acesta este prezentat în tabelul de mai jos

Numărul de unităţi, "m" din produsul P cerut zilnic " x_i "	Numarul de zile în care s-au vândut "m" unităţi " n_i	Frecvenţa relativă m/100
0	5	
1	15	
2	23	
3	22	
4	16	
5	9	
6	5	
7	5	

a) Să se completeze coloana frecvenţelor 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 variaţie)
d) Să se interpreteze datele obţinute

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



c)Indicatorii de poziţie sunt:

- media $\bar{x} = \frac{0 \cdot 5 + 1 \cdot 15 + 2 \cdot 23 + 3 \cdot 22 + 4 \cdot 16 + 5 \cdot 9 + 6 \cdot 5 + 7 \cdot 5}{100} = 2.85$

- 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 aceasta valoare are cel mai mare număr de apariții

Indicatorii de poziție sunt

-dispersia: $s^2 = \frac{1}{99} \cdot [5 \cdot 0^2 + 15 \cdot 1^2 + 23 \cdot 2^2 + 22 \cdot 3^2 + 16 \cdot 4^2 + 9 \cdot 5^2 + 5 \cdot 6^2 + 5 \cdot 7^2 - \frac{2.85^2}{100}] =$

$= \frac{1}{99} \cdot 1210{,}91 = 12.23$

- abaterea standard

$s = \sqrt{s^2} = \sqrt{12.23} = 3.49$

-coeficientul de variație $c = \frac{\sigma}{\overline{x}} = \frac{3.49}{2.85} = 1.22$.

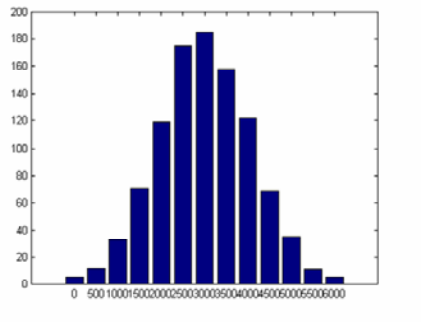
2. Intr-o bancă s-au înregistrat sumele retrase de 1000 clienți în cursul unei luni. Datele au fost grupate în tabelul de mai jos.

Suma retrasă (în euro)	Număr clienți care au retras suma
[0,500)	5
[500,1000)	12
[1000,1500	33
[1500,2000)	71
[2000,2500)	119
[2500,3000)	175
[3000,3500)	185
[3500,4000)	158
[4000,4500)	122
[4500,5000)	69
[5000,5500)	35
[5500,6000)	11
>=6000	5
Total	1000

a) Să se deseneze histograma cu bare a acestei serii de date (sumele mai mari de 6000 se identifică cu intervalul $[6000,6500)$.
b) Identificând fiecare interval cu mijlocul său, să se constituie seria statistică a retragerilor effectuate de 1000 de clienți ai băncii. Să se determine media, mediana şi dispersia acestei serii.

Rezolvare:

a) histograma este



b) Seria de date este

$$X = \begin{pmatrix} 225 & 725 & 1225 & 1725 & 2225 & 2725 & 3225 & 3725 & 4225 & 4725 & 5225 & 5725 & 6225 \\ 5 & 12 & 33 & 71 & 119 & 175 & 185 & 158 & 122 & 69 & 35 & 11 & 5 \end{pmatrix}$$

Media este

$$\bar{x} = (225 \cdot 5 + 725 \cdot 12 + 1225 \cdot 33 + 1725 \cdot 71 + 2225 \cdot 119 + 2725 \cdot 175 + 3225 \cdot 185 + 3725 \cdot 158 +$$
$$+ 4225 \cdot 122 + 4725 \cdot 69 + 5225 \cdot 35 + 5725 \cdot 11 + 6225 \cdot 5) / 1000 = 2981{,}75$$

Mediana este $me = \frac{x_{500} + x_{501}}{2} = \frac{3225 + 3225}{2} = 3225$

Dispersia este

$s^2 = (10384193750 - 2981750\text{^}2/1000)/999 = 1494855.55$.

Abaterea standard este $s = \sqrt{s^2} = 1230$