§ 3.2. Sinteza numerică univariată

Sinteza numerică ţinteşte direct şi indirect aspecte esenţiale de variabilitate. Este un instrument complementar sintezei grafice, care oferă măsuri *obiective* şi *exacte* ale conceptelor sugerate în tabelul de forme tip de distribuţii (vezi 3.1.2. punctul 2°): omogenitate versus eterogenitate (ca aspecte de variabilitate), tendinţă centrală şi alte tipuri de tendinţe etc.

Conceptul statistic de variabilitate poate fi tratat în două moduri provenite din gândirea cantitativă, respectiv calitativă. Gândind cantitativ (eventual chiar geometric continuu), variabilitatea este concepută ca o *împrăştiere,* eventual în jurul unei *tendinţe centrale*. Gândind pur calitativ, variabilitatea "devine" ceea ce se poate denumi *diversitate*.

Într-adevăr, dacă urmărim o caracteristică cantitativă, de exemplu înălţimea la o anumită specie de animale, observăm că exemplarele mature ale acestei specii vor avea o înălţime care se "împrăştie" în jurul unei "tendinţe centrale" pe care o realizează cele mai multe exemplare. Dacă gândim calitativ, putem observa uneori, la un moment dat, o variabilitate atât de mare încât putem vorbi chiar de specii distincte, deci de o "diversitate" de specii. Detalii cu privire la acest mod de gândire familiar biologilor, dar neconceptualizat suficient în statistică vor fi prezentate în paragraful 3.8.

Modul de gândire cantitativ se aplică *variabilelor cantitative,* dar şi celor *calitative binare* sau *binarizate*. Şi pentru acestea, sinteza numerică se face în indicatori (valori tipice) de:

- localizare, poziţionare a *tendinţei centrale,* a tendinţelor extreme şi a tendinţelor intermediare,
- *împrăştiere* (variabilitate, dispersie), de regulă în jurul tendinţei centrale.

Pentru variabilele cantitative continue sau cel puţin comparabile cu variabile continue se calculează şi indicatori de:

- *formă* (pentru compararea cu o distribuţie normală).

Această ultimă categorie care oferă printre altele şi măsuri obiective şi exacte ale asimetriei nu va fi însă tratată în cadrul de faţă.

§ 3.3. Tratarea unei *variabile cantitative* ▸ Indicatori de tendinţă centrală

3.3.1. Condiţiile lui Yule asupra unui indicator de tendinţă centrală

Yule formulează următoarele condiţii pe care ar trebui să le satisfacă un indicator de tendinţă centrală, o "medie" [19]:

1. Să fie definit în mod obiectiv, independent de aprecierea subiectivă a cercetătorului.
2. Să fie expresia tuturor termenilor repartiţiei (seriei).
3. Să posede proprietăţi simple, evidente, făcând posibilă înţelegerea sensului lui general.
4. Să poată fi calculat cu uşurinţă şi rapiditate.
5. **Să se preteze uşor la calculele algebrice ulterioare.**
6. În cazul eşantioanelor, să fie afectat cât mai puţin de fluctuaţiile de selecţie[1] (în particular de *valorile aberante*).

La aceste condiţii ar trebui adăugată specificarea că are sens să vorbim despre tendinţă centrală doar în cazul distribuţiilor unimodale.

Înainte de a prezenta cei mai utilizaţi indicatori de tendinţă centrală, menţionăm că Yule preferă *media* (aritmetică), în special datorită unei proprietăţi numită proprietate de

[1] Sensul acestei cerinţe va putea fi înţeles numai în capitolul de statistică inductivă.

aditivitate. Aceasta va fi prezentată în paragraful dedicat mediei. În general, un concept sau o teorie matematică bogate în proprietăți "frumoase" descoperite de matematicieni, de regulă din motive estetice, se dovedesc, mai devreme sau mai târziu, de cea mai mare utilitate practică în cele mai diverse aplicații. De aceea proprietatea 5 de mai sus este cea mai importantă condiție.

În continuare se vor trata următorii indicatori de tendință centrală: *moda, mediana* şi *media* (aritmetică). Am ales această ordine deoarece primele două concepte, fiind geometrice, au un caracter mult mai intuitiv. Acestea vor fi definite mai întâi în cadrul naturii lor geometrice – curbele de frecvențe – apoi vom scoate în evidență "prețul plătit" pentru a fi definite şi în cadrul discret al seriilor statistice. Media este lăsată la urmă fiind un concept algebric, mai puțin intuitiv. Va fi definită numai în cadrul ei natural – seriile statistice, căci definirea în cadrul geometric este mult mai dificilă fiind nevoie de intermedierea unui instrument matematic neelementar, calculul integral.

3.3.2. Modă

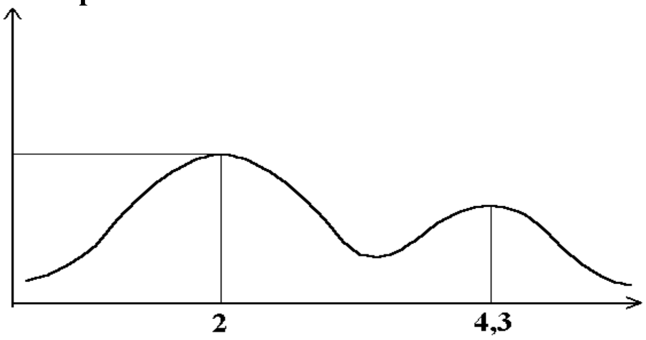
Sinonime: **mod**, **modul**, **dominantă**, **valoare dominantă**, **valoare modală**.

1° Notație: ***Mo*.**

2° Definiții

- În cazul unei *curbe de frecvențe* (distribuție continuă a unei variabile continue), **modă =** punct de maxim local.

Exemplul 3.2.2.



2 şi 4,3 sunt mode pentru distribuția continuă, deoarece sunt puncte de maxim local.

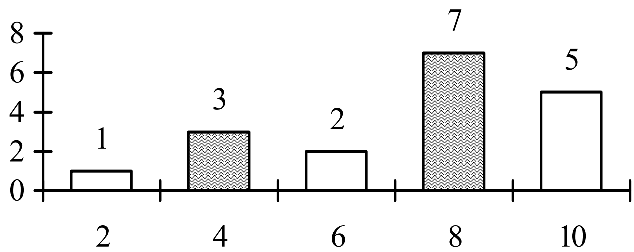
- În cazul *seriilor statistice*, pentru sesizarea modelor, datele trebuie să fie prezentate în *distribuții de frecvențe (negrupate).* În cazul utilizării intervalelor de grupare, obținându-se distribuții de frecvențe grupate, în loc de mode se vorbeşte despre **intervale modale**. În continuare ne vom referi numai la distribuțiile negrupate.

Modă = valoare cu frecvență maximă locală în distribuția de frecvențe.

✓ "Prețul plătit", pentru observarea modelor în acest caz, este gruparea datelor seriilor statistice în distribuții de frecvențe grupate sau nu.

Exemplul 3.2.2'.

x_i	N_i
2	1
→ 4	3
6	2
→ 8	7
10	5



4 şi 8 sunt mode deoarece 3, respectiv 7 sunt frecvenţe maxime locale după cum rezultă şi din diagrama în batoane corespunzătoare.

3° Proprietăţi

Vom enunţa doar următoarele două proprietăţi, cea negativă descalificând moda în teoria şi practica statistică.

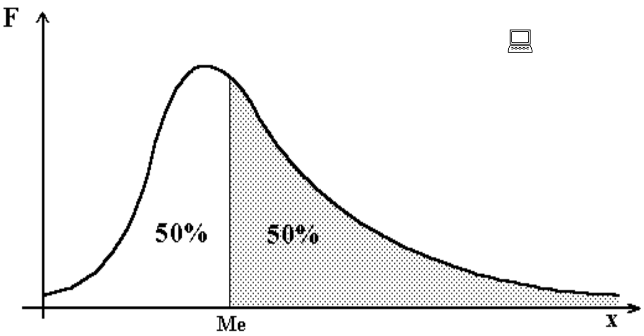
Proprietăţi pozitive:	**Proprietăţi negative:**
1. Modele induc clasificarea în distribuţii unimodale, respectiv multimodale, clasificare esenţială în gândirea statisticii clasice.	2. Nu se pretează la calcule algebrice.

3.3.3. Mediana

1° Notaţii: ***Me*** (pentru populaţii statistice), $\tilde{x}$ (pentru eşantioane).

2° Definiţii

- În cazul unei *curbe de frecvenţe* (distribuţie continuă a unei variabile continue), **mediana** = valoarea care împarte aria de sub curba de frecvenţe în două arii egale (fiecare arie reprezentând 50% din întreaga arie de sub curbă).



- În cazul *seriilor statistice*, **mediană =** o valoare care împarte seria statistică ordonată în două subserii de volume egale, volumele fiind măsurate în număr de unităţi statistice şi, eventual, jumătăţi ale acestora[2].

a) Dacă seria are un număr impar de valori (2*k*+1), mediana este unic determinată de definiţie şi este valoarea x_{k+1} din seria ordonată.

b) Dacă seria are un număr par de valori (2*k*), definiţia este satisfăcută de orice număr cuprins între valorile centrale x_k şi x_{k+1}, din seria ordonată. Pentru unicitatea rezultatului se ia, prin convenţie, drept mediană semisuma valorilor x_k şi x_{k+1}, din seria ordonată.

➢ Incluzând în definiţie şi această convenţie se obţine unicitatea medianei şi pentru serii:

Mediana = termenul din mijlocul seriei statistice ordonate, în caz de volum impar, respectiv semisuma celor doi termeni din mijlocul seriei ordonate, în caz de volum par.

➢ În cazul seriilor statistice "preţul plătit" pentru asigurarea unicităţii medianei a fost introducerea convenţiei suplimentare de mai sus.

În practică: Ordonăm seria de volum *N*. Calculăm rangul medianei *rg*(*Me*) = *N* / 2. Dacă *rg*(*Me*) este număr fracţionar îl rotunjim prin adaos şi *Me* este termenul cu rangul rotunjit. În caz contrar, *Me* este semisuma termenului cu rangul *rg*(*Me*) şi următorul termen.

[2] Precizarea dublu subliniată este propusă de autor. Se va dovedi şi mai utilă în definiţia cuantilelor (vezi 3.4.)

Exemplele 3.3.3.

a) Fie seria ordonată cu 5 (număr impar de) termeni: 1; 3; 7; 8; 12. *Me* = 7 deoarece considerăm că valoarea 7 împarte seria ordonată în subseria formată din 1; 3 şi o jumătate din 7 şi respectiv, cealaltă jumătate din 7; 8 şi 12. **Mai simplu:** *Me* = 7 pentru că se află în mijlocul seriei ordonate de volum impar. **Practic:** *rg* (5/2) = 2,5 (număr fracționar - se rotunjeşte, prin adaos, la 3). Deci *Me* = termenul de rang 3 = 7.

b) Fie seria ordonată cu 4 (număr par de) termeni: 1; 3; 6; 18. Conform definiției orice număr cuprins între 3 şi 6 - de exemplu 3,7; 4,5 sau 5,2 - poate fi mediană. Convenția suplimentară produce rezultatul unic: *Me* = (3 + 6) / 2 = 9 / 2 = 4,5**. Mai simplu:** *Me* = semisuma termenilor din mijlocul seriei ordonate de volum par (3 + 6) / 2 = 4,5. **Practic:** rg(4/2)=2 (număr întreg). Deci *Me* = semisuma termenilor de rang 2 şi 3 = (3 + 6) / 2 = 4,5.

- În cazul distribuțiilor de frecvențe, mediana se poate defini, de asemenea, în mod specific. În cazul distribuțiilor de frecvențe negrupate va fi egală ca valoare cu mediana seriei, iar în cazul celor grupate va fi doar aproximativ egală cu mediana seriei.

3° Proprietăți

Pozitive:	**negative:**
1. Mediana este relativ uşor de observat şi calculat.	
2. Exprimă cel mai bine tendința centrală (în special, la distribuții asimetrice).	
3. Nu ia în considerare valorile seriei decât din punct de vedere al ordinii acestora şi, deci, este practic indicatorul specific al tendinței centrale al variabilelor ordinale. Pentru variabilele cantitative lucrează în sensul statisticii descriptive, adică renunță la o parte din informație pentru câştig în relevanță. Altfel spus, mediana tratează valorile ca pe ranguri.	.
4. Nu este sensibilă la valori extreme (în particular la valori aberante).	
5. Poate fi calculată şi la serii pentru care nu se poate calcula exact media, de exemplu atunci când valorile extreme nu sunt cunoscute, ca în cazul unei măsurători cu un aparat insensibil la valori extreme.	
6. Este element al şirului atunci când şirul are număr impar de termeni.	
7. Poate fi determinată în unele cazuri mult mai uşor prin simple comparații şi ordonare, în loc de măsurători. De exemplu, dacă avem o colecție de crustacee în număr impar, acestea pot fi ordonate după lungime fără a măsura exemplarele, apoi se va măsura doar exemplarul de lungime mediană.	8. Nu se pretează la calcule algebrice

4° Observații

Mediana este întâlnită şi sub următoarele denumiri:

✓ *În toxicologie:* **LD 50** = "**L**ethal **D**ose **50**" = "**D**oza **L**etală **50**" = Doza care omoară 50% din indivizii care au fost intoxicați cu doza respectivă.

✓ *În farmacologie:* **ED 50** = "**E**ffect **D**ose **50**" = **D**oza care are **E**fect asupra a 50% din indivizii tratați cu doza respectivă.

✓ *În biologia populațiilor:* **Media de viață**. Mortalitatea populației în funcție de vârstă pe o curbă de frecvențe (vezi de exemplu distribuția *D2'* de la 3.1.2.) are o mediană care reprezintă vârsta până la care au murit 50% din indivizii populației respective. Această mediană se numeşte în mod impropriu "medie de viață" a populației respective şi reprezintă o situație de excepție în care se calculează un indicator de tendință centrală pentru o curbă multimodală.

3.3.4. Media (aritmetică)

Termenul "medie" este folosit – ce-i drept, mai rar - în sensul general de indicator de tendință centrală şi – cel mai frecvent - în sensul restrâns de *medie aritmetică.* Calificativul "aritmetică" se adaugă în special pentru diferențierea noțiunii față de alte tipuri de medii - *geometrică*, *armonică*, etc. - mai puțin folosite în biostatistică. Utilizat fără calificativ, termenul "medie" înseamnă "medie aritmetică".

1° Notații: M (pentru populații statistice în general), μ (pentru populații statistice teoretice), $\bar{x}$, m (pentru eşantioane).

2° Definiții

(a) În cazul unei *serii statistice* formate din N valori distincte sau nu, $x_1, x_2, \ldots, x_i, \ldots, x_N$, **media** M este suma valorilor seriei împărțită la volumul seriei:

$$M = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Vom numi această formulă, **formula mediei simple** a valorilor (distincte sau nu) x_i.

(b) În cazul unei serii statistice grupate în *distribuția de frecvențe absolute*, (x_j, N_j), ale celor p ($\leq N$) valori distincte, x_j, **media** M va fi dată de formula:

$$M = \frac{\sum_{j=1}^{p} N_j \cdot x_j}{\sum_{j=1}^{p} N_j},$$

pe care o vom denumi **formula mediei ponderate** a valorilor distincte x_j.

Frecvența N_j se va numi **ponderea (absolută a)** valorii x_j, iar $\sum_{j=1}^{p} N_j = N$, volumul seriei.

(c) În cazul unei serii statistice grupate în *distribuția de frecvențe RELATIVE*, (x_j, F_j), ale celor p ($\leq N$) valori distincte, x_j, **media** M va fi dată de formula:

$$M = \sum_{j=1}^{p} F_j \cdot x_j,$$

care este o **SUMĂ ponderată** a valorilor distincte x_j, frecvențele $F_j = \frac{N_j}{N}$ fiind aici **ponderi RELATIVE**, adică $F_j \leq 1$ şi $\sum_{1}^{p} F_j = 1$.

Exemplul 3.3.4.

Fie *seria* de valori de 6 valori:

(a) 1; 4; 2; 2; 1; 2,

respectiv aceeaşi serie prezentată ca *distribuție de frecvențe absolute*:

(b)

x_j	1	2	4	
N_j	2	3	1	$N = 6$

sau *de frecvențe relative*:

(c)

x_j	1	2	4
F_j	2 / 6	3 / 6	1 / 6

Să se calculeze media aritmetică prin formulele corespunzătoare.

Rezolvare:

(a) $M = (1 + 4 + 2 + 2 + 1 + 2) / 6 = 12 / 6 = 2.$

$M = 2$ este media simplă a seriei de valori distincte sau nu 1; 4; 2; 2; 1; 2.

(b) $M = (2 \cdot 1 + 3 \cdot 2 + 1 \cdot 4) / (2 + 3 + 1) = (2 + 6 + 4) / 6 = 12 / 6 = 2.$

$M = 2$ este media ponderată a seriei de valori distincte 1; 2; 4 cu ponderile 2; 3; 1.

Media simplă a acestei noi serii (1; 2; 4) este $M' = (1 + 2 + 4) / 3 \approx 2{,}33$.

(c) $M = (2 / 6) \cdot 1 + (3 / 6) \cdot 2 + (1 / 6) \cdot 4 = 2 / 6 + 6 / 6 + 4 / 6 = 12 / 6 = 2.$

$M = 2$ este suma ponderată a seriei 1; 2; 4 cu ponderile relative 2 / 6; 3 / 6 şi 1 / 6.

+3° Proprietatea de aditivitate a mediei

Să presupunem că o serie statistică a fost separată, din anumite raţiuni, în două sau mai multe grupări de volume v_k pentru care s-au calculat mediile M_k. Atunci media totală, M, se poate calcula din mediile grupărilor prin formula:

(a)

$$\boxed{M = \frac{\sum_{k=1}^{q} v_k \cdot M_k}{\sum_{k=1}^{q} v_k}}$$

. Adică:

"media totală este media ponderată a mediilor grupărilor, ponderile fiind volumele grupărilor".

Această egalitate se numeşte "**proprietatea de aditivitate a mediei**".

(b) Justificarea denumirii proprietăţii provine din următoarea formă de scriere a sa. Pentru aceasta să notăm cu vr_k volumule relative ale grupărilor, adică $vr_k = \frac{v_k}{\sum_{k=1}^{q} v_k}$. Astfel egalitatea de mai sus se poate scrie şi sub forma:

$$\boxed{M = \sum_{k=1}^{q} vr_k \cdot M_k}$$

, adică:

"media totală este suma ponderată a mediilor grupărilor, ponderile fiind volumele relative ale grupărilor".

Prin urmare, aşa cum putem obţine volumul total prin adiţionarea volumelor grupărilor, $N = \sum_{k=1}^{q} v_k$, tot aşa putem obţine şi media totală prin adiţionarea, ponderată ce-i drept, a mediilor grupărilor, ceea ce justifică denumirea de proprietate de aditivitate.

Exemplul 3.3.4'.

Fie seria formată din următoarele două grupări:

0; 2 şi

5; 6; 7,

cu mediile parţiale 1, respectiv 6. *Proprietatea de aditivitate a mediei* ne permite să obţinem *media totală*, 4, pornind de la:

(a) *mediile* şi *volumele* absolute ale grupărilor: $4 = \dfrac{2 \cdot 1 + 3 \cdot 6}{2+3}$ sau de la

(b) mediile şi volumele relative ale grupărilor: $4 = \dfrac{2}{5} \cdot 1 + \dfrac{3}{5} \cdot 6$.

Observație:

(a) Enunțul mai scurt "*media totală este media mediilor parțiale (ale grupărilor)*" reprezintă forma "populară" de exprimare a proprietății de aditivitate a mediei şi este corect doar dacă volumele grupărilor sunt egale.

4° Proprietăți

pozitive:	**negative:**
1. **Proprietatea** algebrică **de aditivitate**.	4. Este relativ dificil de calculat manual.
2. Se pretează la calcule algebrice ulterioare, de exemplu pe baza proprietății de aditivitate.	5. Este sensibilă la valorile extreme (în particular la cele aberante).
3. Media aritmetică ia în considerare toate valorile seriei cu întreaga lor informație.	

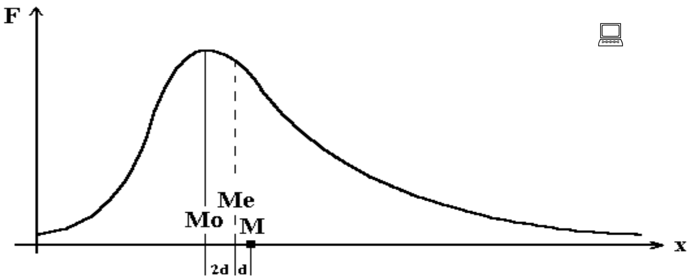
3.3.5. Indicații de preferință între principalii indicatori de tendință centrală

Steinbach [29] formulează următoarele indicații de preferință completate de noi cu aspectele subliniate:

1. Moda unică indică omogenitatea, deci sensul indicatorilor de tendință centrală, în particular al mediei (media nu are sens decât la o distribuție unimodală).
2. O distribuție <u>cvasi</u>simetrică şi cu prelucrări statistice ulterioare cere o medie.
3. Pentru o orientare rapidă la o serie mare folosim moda.
4. Pentru considerarea tuturor valorilor, <u>împreună cu întreaga lor informație</u>, folosim media.
5. Dacă vrem să facem abstracție de valorile extreme se ia în considerare mediana sau moda.
6. Pentru o serie cu volum redus se preferă mediana, pentru că la volum redus nu se "vede" moda. Mediana exprimă cel mai bine tendința centrală la distribuții asimetrice.
7. Dacă cunoaştem doar clasamentul, adică doar rangurile valorilor, atunci folosim mediana.

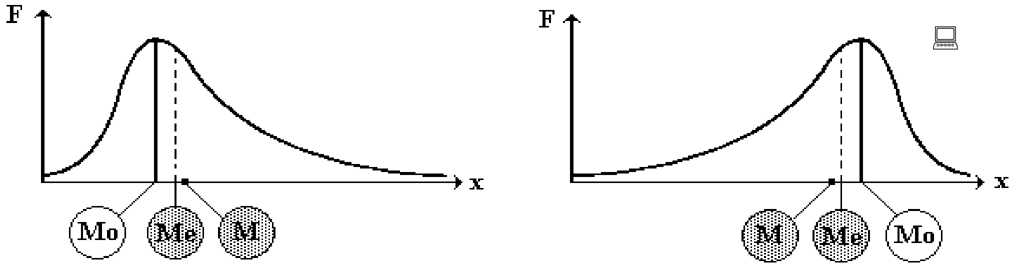
+ Observații:

- ✓ La distribuții unimodale şi simetrice cei trei indicatori coincid.
- ✓ La distribuțiile unimodale uşor asimetrice mediana se află plasată între modă şi medie, distanța acesteia față de modă fiind aproximativ dublul distanței față de medie, ca în figura următoare.



✓ Sensul asimetriei unei distribuţii poate fi detectat în mod <u>aproximativ</u> prin echivalenţele:

(1) asimetrie de stânga ⇔ *Me* < *M*; asimetrie de dreapta ⇔ *M* < *Me*;



(2) asimetrie de stânga ⇔ *Mo* < *Me*; asimetrie de dreapta ⇔ *Me* < *Mo*;

(3) asimetrie de stânga ⇔ *Mo* < *M*; asimetrie de dreapta ⇔ *M* < *Mo*.

Punctele anterioare ne duc la concluzia că prin compararea a doi indicatori de tendință centrală se poate obține o indicație asupra formei distribuției, mai precis a simetriei şi a tipului de asimetrie. Altfel spus compararea medianei cu media este un criteriu aproximativ de detectare a tipului de asimetrie. O măsură exactă poate fi obținută printr-un anumit indicator de asimetrie, care nu va fi însă prezentat în