2. INDICATORII TENDINȚEI CENTRALE, VARIAȚIEI ȘI FORMEI DISTRIBUȚIILOR

- **Indicatorii tendinței centrale**:
  - Mediile aritmetică, armonică, geometrică și pătratică
  - Mediana
  - Modul
- **Indicatorii variației sau dispersiei** (împrăștierii):
  - Amplitudinea variației
  - Varianța și deviația standard
  - Coeficientul de variație
  - Cuartilele, decilele și centilele (diagramele Box Plot)
- **Indicatorii formei distribuțiilor**
  - Coeficientul de asimetrie
  - Coeficientul de boltire
- **Importanța practică a deviației standard** (scorurile z, regula empirică)

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O3-1: calculeze și interpreteze indicatorii tendinței centrale (media aritmetică, media armonică, media geometrică, media pătratică, mediana și modul)

O3-2: calculeze și interpreteze indicatorii variației (amplitudinea variației, varianța și deviația standard)

O3-3: să explice și să aplice corect teorema lui Cebîșev și regula empirică

O3-4: să construiască și să interpreteze cuartilele, decilele și centilele

O3-5: să calculeze și interpreteze coeficienții de asimetrie și de boltire

INTRODUCERE

În capitolul anterior am văzut modalitățile de centralizare a datelor statistice într-o formă clară, sugestivă și ușor de interpretat și ulterior de prezentare a lor sub formă grafică, indiferent dacă variabilele statistice studiate îmbracă o formă calitativă sau cantitativă.

În continuare, vom prezenta trei modalități numerice (indicatori statistici) pentru a descrie variabilele cantitative și anume indicatorii TENDINȚEI CENTRALE (mărimile medii), indicatorii DISPERSIEI sau împrăștierii și indicatorii FORMEI DISTRIBUȚIILOR statistice.

O mărime medie, în general, reprezintă un indicator al localizării care evidențiază valoarea centrală a unei variabile statistice sau mijlocul unei distribuții.

Exemple de mărimi medii întâlnite în practică:

- vârsta medie a populației rezidente a României s-a majorat de la 40,8 ani în 2012 până la 42,3 ani în 2021;
- vârsta medie a mamei la prima naștere în România a crescut de la 22,3 ani în 1990 până la 27,7 ani în 2021;
- câștigul salarial nominal net lunar în România a crescut de la 3217 lei în anul 2020 la 3416 lei în anul 2021 (o creștere de aproximativ 6,2%); ajustat cu rata inflației aferentă perioadei, indicele câștigului salarial real arată o creștere de numai 1,1%;
- numărul mediu de pensionari a crescut de la 3.679.000 persoane în anul 1990 la 5.079.000 persoane în anul 2021 (cu aproximativ 38%).

INDICATORII TENDINȚEI CENTRALE

Într-o cercetare statistică completă imaginea generală conturată prin reprezentarea grafică a datelor este insuficientă, fiind preferată determinarea unui indicator de sinteză, care să reprezinte nivelul general al fenomenului studiat, modelul lui central de dezvoltare – respectiv O VALOARE REPREZENTATIVĂ, în jurul căreia are loc o tendință de concentrare a nivelurilor individuale ale variabilei studiate.

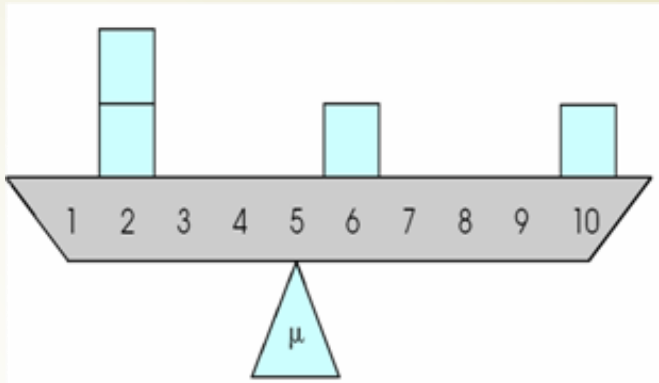
Fenomenele colective supuse analizei statistice sunt purtătoare a mai multor caracteristici, fiecare cu un număr mare de variante ca urmare a îmbinării influențelor diferiților factori sub acțiunea cărora iau naștere și se dezvoltă aceste fenomene. Variația unei caracteristici în cadrul unei colectivități este determinată de acțiunea factorilor esențiali asociați cu factorii neesențiali.

Factorii esențiali acționează asupra tuturor unităților dintr-o colectivitate și determină dezvoltarea lor normală sub aspectul unei anumite caracteristici sau tendința normală, centrală de dezvoltare a unei colectivități. Asupra fenomenelor acționează și factori întâmplători (neesențiali), care determină abateri în ambele sensuri ale nivelurilor individuale ale unei variabile de la tendința normală de dezvoltare.

Mărimile medii evidențiază deci influența factorilor esențiali care au acționat asupra unităților statistice dintr-o colectivitate sub aspectul unei caracteristici care prezintă interes în cercetare.
Influenta factorilor neesențiali sau întâmplători, care determină împrăștierea nivelurilor individuale ale unei variabile în jurul nivelului ei mediu, este măsurată de indicatorii variației (împrăștierii).

1. MEDIA ARITMETICĂ

Media aritmetică (arithmetic mean or average) a unui șir de valori individuale ale unei variabile statistice este egală cu raportul dintre suma și numărul lor (este influențată mai puternic de valorile mari ale variabilei analizate).



Media aritmetică a unei variabile X cu variantele ei x_1, x_2,..., x_i,..., x_n, în cazul unui **eșantion** de volum n extras dintr-o populație generală de volum N, reprezintă o **STATISTICĂ**:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Media aritmetică a unei variabile X cu variantele ei x_1, x_2,..., x_i,..., x_N, în cazul unei **populații** de volum N reprezintă un **PARAMETRU**:

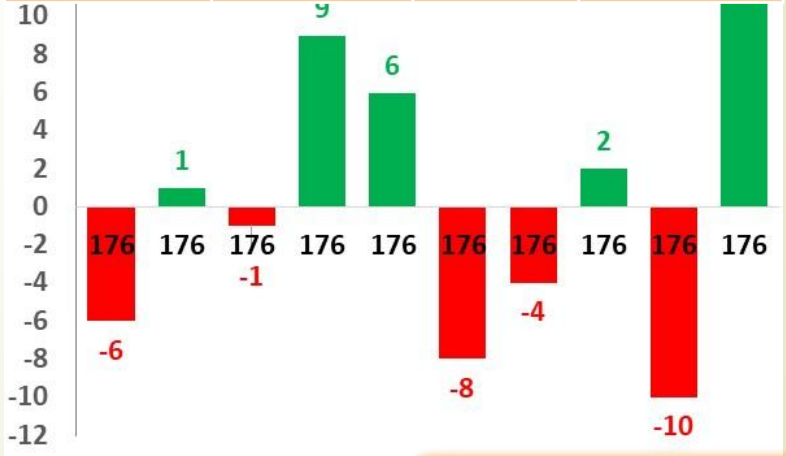
$$\mu = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

În anumite situații particulare, o variabilă poate înregistra una sau mai multe valori neobișnuit de mari sau de mici, care să influențeze într-un mod nedorit valoarea mediei, astfel încât aceasta să nu mai poată servi drept indicator corespunzător al tendinței centrale. Pentru a elimina efectul negativ al acestor valori extreme putem înlătura un procent redus (de exemplu 5%) al acestora din cadrul setului de date. Media valorilor rămase ale variabilei studiate poartă denumirea de **medie aritmetică ajustată** (trimmed mean).

Proprietățile mediei aritmetice

- variabilele utilizate în calculul mediei trebuie să fie de tip interval sau raport
- în calculul mediei aritmetice sunt incluse toate variantele variabilei studiate
- media aritmetică este unică pentru fiecare variabilă statistică în parte, fiind cuprinsă între valorile minimă și maximă ale acesteia
- media aritmetică a unui șir de valori constante este egală cu valoarea lor comună
- media aritmetică este asociativă (dacă se înlocuiesc mai multe variante ale variabilei cu media lor, atunci media seriei modificate este egală cu media seriei inițiale)
- media este translativă (dacă se mărește sau se micșorează fiecare nivel individual al variabilei cu o cantitate constantă, atunci media se mărește sau se micșorează cu acea constantă)
- suma tuturor abaterilor fiecărei valori individuale a unei variabile de la nivelul mediu al acesteia este zero.

PERSOANA	ÎNĂLȚIMEA	MEDIA	DIFERENȚE
1	170	176	-6
2	177	176	1
3	175	176	-1
4	185	176	9
5	182	176	6
6	168	176	-8
7	172	176	-4
8	178	176	2
9	166	176	-10
10	187	176	11
TOTAL	1760	1760	0
MEDIA	176	176	0



În situația în care variabila statistică X, care intră în calculul mediei, prezintă mai multe observații x_i care au aceeași valoare n_i (ponderi sau frecvențe de apariție), vom calcula **MEDIA ARITMETICĂ PONDERATĂ:**

$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{\sum_{i=1}^{k} n_i}$ pentru un eșantion de volum k $\mu = \frac{\sum_{i=1}^{N} x_i \cdot n_i}{\sum_{i=1}^{N} n_i}$ pentru o populație de volum N

În cazul unei serii cu frecvență având **variabila distribuită pe intervale** este necesar să se determine mai întâi câte o valoare reprezentativă a fiecărui interval. Admițându-se convențional că, în cadrul fiecărui interval, variabila studiată se distribuie în mod liniar uniform, mijlocul intervalului respectiv x_i' este valoarea reprezentativă a fiecărui interval. În această situație vom aplica formulele mediei ponderate prezentate anterior, utilizând mijloacele intervalelor de variație ale variabilei statistice.

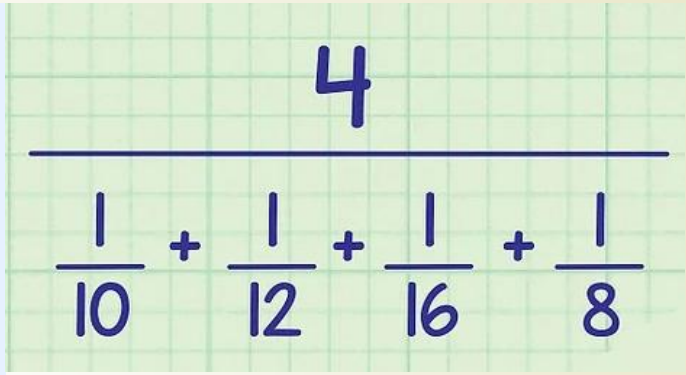
EX: o farmacie a vândut 10 medicamente, 3 din sortimentul A cu prețul unitar de 125 lei, 5 din sortimentul B cu prețul unitar de 95 lei și 2 din sortimentul C cu prețul unitar de 145 lei. Care este prețul mediu de vânzare ale celor 10 medicamente vândute:

$$\overline{x} = \frac{125 + 125 + 125 + 95 + 95 + 95 + 95 + 95 + 145 + 145}{10} = \frac{1140}{10} = 114\ lei$$

$$\overline{x} = \frac{3 \cdot 125 + 5 \cdot 95 + 2 \cdot 145}{10} = \frac{1140}{10} = 114\ lei$$

2. MEDIA ARMONICĂ

Media armonică (harmonic mean) este o formă transformată a mediei aritmetice și se determină ca o medie aritmetică inversă, calculată din mărimile inverse ale variantelor variabilei (se utilizează în calculul valorii medii în cazul mărimilor relative, fiind mai puternic influențată de valorile mici ale variabilei). Formula se poate aplica numai atunci când toate variantele variabilei studiate sunt valori pozitive ($x_i \neq 0$).



$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$ pentru un eșantion n sau $$\overline{\mu}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$ pentru o populație N

În situația în care variabila statistică X, care intră în calculul mediei, prezintă mai multe observații x_i care au aceeași valoare n_i (ponderi sau frecvențe de apariție), vom calcula **MEDIA ARMONICĂ PONDERATĂ:**

$$\overline{x}_h = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$$ pentru un eșantion n sau $$\overline{\mu}_h = \frac{\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} \frac{n_i}{x_i}}$$ pentru o populație N

EX 1: **Presupunem că un anumit medicament își modifică prețul în timpul unui trimestru de 3 ori. Astfel, în prima lună cu 500 de lei se puteau cumpăra 26 cutii, în a două lună cu aceeași sumă 22 cutii, iar în cea de a treia doar 18 cutii. Se cere să se determine cantitatea medie vândută și prețul mediu al acestui medicament pentru întreg trimestrul considerat.**

Dacă am aplica media aritmetică: $\bar{x} = (26 + 22 + 18)/3 = 22\ cutii$, **iar prețul mediu ar fi:** $500/22 = 22,7\ lei/cutie$. **Acest preț mediu nu ține însă cont de cantitățile reale ale medicamentului care au fost vândute în fiecare lună a trimestrului. Dacă am cunoaște aceste cantități, atunci cantitatea medie și prețul mediu ar putea fi determinate prin aplicarea mediei aritmetice ponderate.**

Determinarea corectă a numărului mediu de cutii vândute se poate realiza cu ajutorul mediei armonice:

$$\overline{x_h} = \frac{3}{\sum_{i=1}^{3} \frac{1}{x_i}} = \frac{3}{\frac{1}{26} + \frac{1}{22} + \frac{1}{18}} = 21,51\ cutii$$

Deci, numărul mediu de cutii care se puteau cumpăra în timpul trimestrului cu 500 de lei a fost de doar 21,51 cutii (cu jumătate de cutie mai mic decât am obținut prin aplicarea mediei aritmetice inițial), iar prețul mediu pentru o cutie de medicamente a fost de aproximativ $500/21,51 = 23,25\ lei/cutie$.

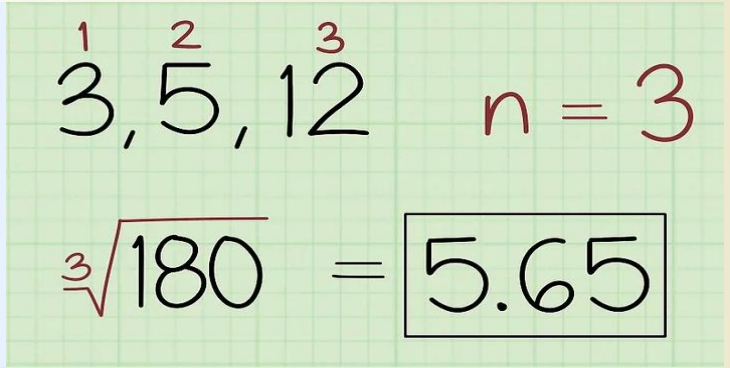
Verificare: prețul mediu de vânzare în prima lună a fost de $500/\ 26 = 19,23\ lei/cutie$, **în a doua lună** $500/22 = 22,72\ lei/cutie$ **și în a treia lună** $500/18 = 27,78\ lei/cutie$. **Calculând media aritmetică a acestor 3 prețuri medii lunare obținem prețul mediu de vânzare pentru întreg trimestrul de 23,25 lei / cutie. În programul Excel putem folosi funcția: =HARMEAN(26,22,18)**

EX2: **Presupunem că ratele de mortalitate (numărul de decese la 1000 de locuitori) în patru localități sunt 8‰, 9‰, 11‰ și 12 ‰. Populațiile celor 4 localități sunt 15000 locuitori, 18000 locuitori, 20000 locuitori, respectiv 30000 locuitori. Vom determina rata medie a mortalității pentru regiunea formată din cele 4 localități:**

$$\bar{x}_h = \frac{\sum_{i=1}^{4} n_i}{\sum_{i=1}^{4} \frac{n_i}{x_i}} = \frac{15000 + 18000 + 20000 + 30000}{\frac{1}{8} \cdot 15000 + \frac{1}{9} \cdot 18000 + \frac{1}{11} \cdot 20000 + \frac{1}{12} \cdot 30000} = 10,13\ ‰$$

3. MEDIA GEOMETRICĂ

În cazul fenomenelor care se dezvoltă în progresie geometrică, pentru determinarea nivelului lor mediu, se utilizează media geometrică (geometric mean). Această mărime medie este folosită de exemplu în calcularea ritmului mediu de creștere anuală sau în dinamica sporului natural al populației (este influențată mai puternic de valorile mici ale șirului).



$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$ pentru un eșantion n sau $\overline{\mu}_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$ pentru o populație N

EX: dacă o tulpină de bacterii își mărește populația cu 20% în prima oră, 30% în următoarea oră și 50% în a treia oră, putem calcula o valoare medie orară estimată a creșterii procentuale a populației acestora.

Vom utiliza în acest caz media geometrică în locul mediei aritmetice.

Să presupunem că avem la început o populație de 100 de bacterii, care se va majora pe parcursul primei ore la 120 de bacterii (= 100 + 100 * 20% = 100 * 120%). După cea de a doua oră, numărul lor ajunge la 156 de bacterii (= 120 + 120 * 30% = 120 * 130%), iar la sfârșitul celei de a treia ore populația de bacterii va fi de 234 unități (= 156 + 156 * 50% = 156 * 150%).

Calculăm media geometrică a acestor creșteri:

$$\bar{x}_g = \sqrt[3]{1,2 \cdot 1,3 \cdot 1,5}$$
$$= 1,3276\ sau\ 132,76\%$$

Astfel, rata medie de creștere a populației de bacterii pe o perioadă de 3 ore este de 32,76% (dacă ar fi crescut uniform pe parcursul celor 3 ore, atunci pornind de la 100 de bacterii am ajunge la 234).

Creștere orară	100	Creștere medie orară ($\bar{x}_g$)	100
120%	120	132.76%	133
130%	156	132.76%	176
150%	234	132.76%	234

În programul Excel putem folosi funcțiile:

=GEOMEAN(1.2,1.3,1.5) sau =(1.2*1.3*1.5)^(1/3) sau =POWER((1.2*1.3*1.5),1/3)

În concluzie, de fiecare dată când ne confruntăm cu o creștere procentuală a unui fenomen într-o perioadă de timp, rata medie de variație se va determina prin aplicarea formulei mediei geometrice.

4. MEDIA PĂTRATICĂ

Media pătratică (mean square) se folosește în situația în care fenomenele înregistrează creșteri, aproximativ, după o **lege exponențială** (creșterea este mai lentă la începutul seriei de date statistice și din ce în ce mai pronunțată spre sfârșitul acesteia), fiind utilizată în analiza tendințelor de evoluție neliniare, de tip exponențial (este influențată mai puternic de valorile mari ale șirului). Este folosită și ca model matematic in calculul indicatorilor sintetici ai variației, de exemplu în determinarea deviației standard.

În cazul unei variabile statistice *X*, având variantele x_1, x_2, x_3, ... , x_i, ..., x_n media pătratică se determină după formula:

$$\bar{x}_p = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$ pentru un eșantion *n* sau $$\bar{\mu}_p = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$ pentru o populație *N*

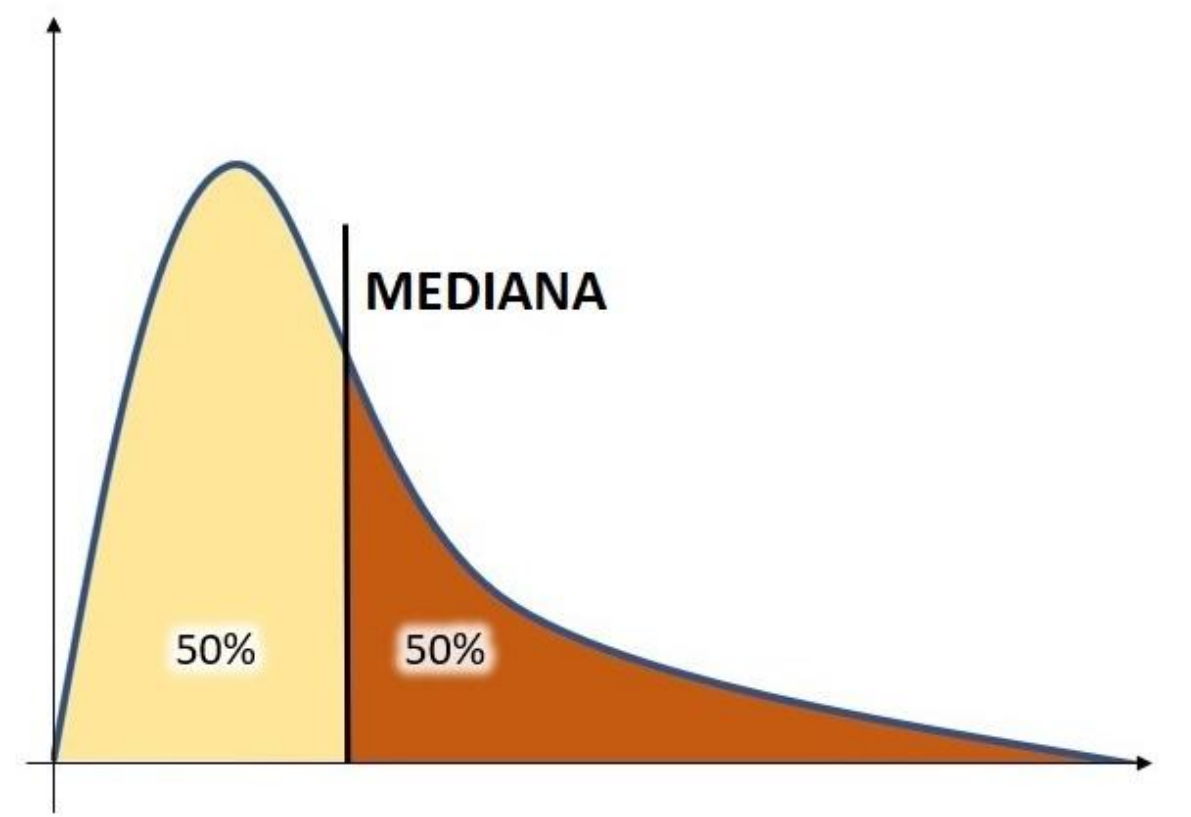
Între mărimile medii prezentate anterior există următoarea relație:

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x} \leq \overline{x}_p$$

Egalitatea între aceste mărimi medii are loc doar în situația când variantele variabilei statistice *X*: $x_1 = x_2 = x_3 = \cdots = x_i = \cdots = x_n = constant$

5. MEDIANA

Mediana sau valoarea mijlocie (median) reprezintă acea valoare a unei variabile statistice care se situează în centrul seriei ordonate crescător sau descrescător, având proprietatea că numărul variantelor mai mari decât ea este egal cu numărul variantelor mai mici.



Mediana poate fi definită ca fiind egală cu varianta x_i a variabilei statistice X, pentru care perpendiculara ridicată pe axa absciselor din punctul x_i împarte suprafața mărginită de curba frecvențelor în două părți egale.

Este un indicator robust al tendinței centrale, mai puțin influențat de valorile extreme comparativ cu media aritmetică și mai stabil la fluctuațiile de selecție. Mediana poate fi calculată și în cazul variabilelor de tip ordinal.

Fie o serie simplă, ordonată de forma: $x_1 \leq x_2 \leq \ldots \leq x_k \leq x_{k+1} \leq x_{k+2} \leq \ldots \leq x_n$.

În determinarea efectivă a medianei ne putem situa în una dintre următoarele situații:

- **Serii cu un număr impar de termeni:** ***n = 2k + 1***

Atunci locul medianei sau numărul de ordine al nivelului variabilei egal cu mediana va fi:

$$U^{me} = \frac{n+1}{2} = \frac{(2k+1)+1}{2} = k+1 \quad \Rightarrow \quad Me = x_{k+1}$$

- **Serii cu un număr par de termeni:** ***n = 2k***

$$U^{me} = \frac{n+1}{2} = \frac{2k+1}{2} = k + \frac{1}{2} \quad \Rightarrow \quad Me = \frac{x_k + x_{k+1}}{2}$$

În acest caz, mediana se găsește în mijlocul intervalului dintre variantele x_p **şi** x_{p+1} **, (zona mediană).**

6. MODUL

Modul (mode) este o mărime medie de poziție, care se poate determina doar în cazul seriilor cu frecvență, și care reprezintă valoarea caracteristicii cu frecvență maximă (care se întâlnește cel mai des).

Valoarea opusă modului, întâlnită în cazul distribuțiilor în formă de J și U, se numește antimod (valoarea cea mai rar întâlnită).

Modul este un indicator corespunzător al tendinței centrale numai în cazul distribuțiilor care se apropie de cea normală, cu un singur punct de maxim (unimodale).

În practică, pentru determinarea modului, deosebim două situații distincte:

a) Serii cu frecvență pe variante ale caracteristicii

Modul este egal cu varianta caracteristicii care are frecvență maximă: dacă considerăm $n_i = n_{max}$ → $Mo = x_i$. În acest caz, modul este bine determinat și este egal cu acea variantă a variabilei care rezultă imediat din citirea tabelului în care este prezentată seria statistică.

b) Serii cu frecvență pe intervale de variație a caracteristicii

Se determină intervalul cu frecvență maximă sau intervalul modal de forma *(xi-1 , xi)*

În cadrul intervalului modal, valoarea căutată a modulului va fi:

$$Mo = x_{i-1} + (x_i - x_{i-1}) \frac{(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

unde: x_{i-1} și x_i sunt limitele inferioară și superioară ale intervalului modal; n_i = frecventa intervalului modal, n_{i-1} = frecventa intervalului anterior și n_{i+1}= frecventa intervalului următor.

INDICATORII VARIAȚIEI SAU DISPERSIEI

Indicatorii variației măsoară influența factorilor neesențiali, care determină împrăștierea nivelurilor individuale ale unei variabile în jurul nivelului ei mediu. Se utilizează pentru caracterizarea omogenității colectivităților (o împrăștiere mai mică determină un grad mai mare de omogenitate și invers).

Gradul de împrăștiere exprimă și măsura în care o medie este reprezentativă pentru o anumită distribuție statistică (o valoare mică a indicatorilor variației înseamnă o concentrare ridicată a valorilor unei variabile în jurul tendinței ei centrale, ceea ce conferă mediei un caracter reprezentativ și invers).

Variația nivelurilor individuale ale unei caracteristici se poate măsura prin abaterea acestora (diferența) față de o mărime medie oarecare (se folosește cel mai frecvent media aritmetică).

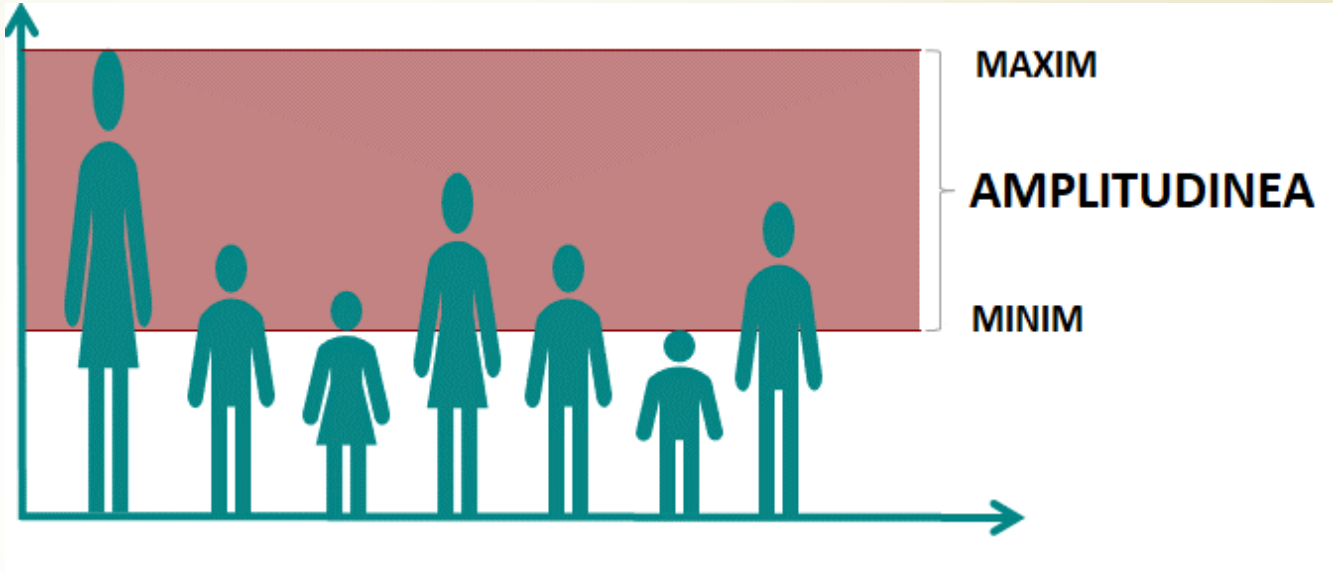
Indicatorii variației exprimă fie distanța dintre două niveluri oarecare ale unei variabile, fie distanța medie dintre toate nivelurile înregistrate, considerate două câte două, fie distanța medie dintre nivelurile caracteristicii tuturor unităților și mărimea lor medie.

a) **Indicatorii simpli ai variației măsoară abaterea fiecărei variante a unei variabile față de nivelul ei mediu sau față de un alt nivel cu o anumită semnificație**
b) **Indicatorii sintetici ai variației măsoară împrăștierea medie a tuturor variantelor unei variabile în jurul nivelului ei mediu. Ei caracterizează complet gradul de împrăștiere a variantelor unei variabile în cadrul distribuțiilor statistice**

1. AMPLITUDINEA VARIAȚIEI

Amplitudinea variației (range) oferă o imagine generală asupra împrăștierii datelor și reprezintă distanța dintre nivelul maxim și nivelul minim al variabilei.

$$A = x_{maxim} - x_{minim}$$



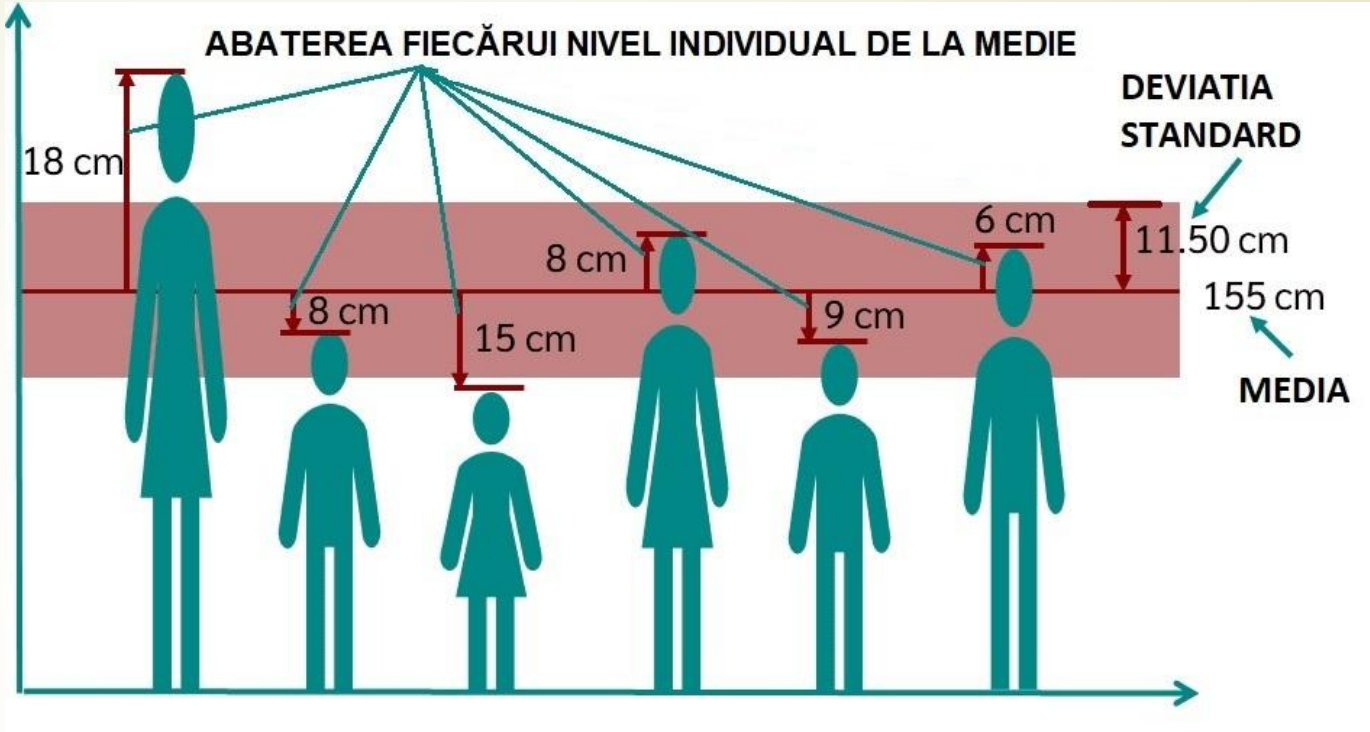
În cazul grupării datelor pe intervale, se determină amplitudinea clasei ca diferență între valorile extreme ale fiecărui interval. Cu cât este mai mică valoarea acestui indicator, cu atât lotul este mai omogen.

Deficiențe:

- **se bazează doar pe două valori (nivelul minim și cel maxim)**
- **depinde de eșantionul de lucru, având variații pentru fiecare eșantion în parte**
- **nu ia în considerare tipul de repartiție al variabilei statistice studiate**

2. VARIANȚA

Varianța sau dispersia (variance) se calculează ca o medie aritmetică a pătratelor abaterilor nivelurilor individuale ale unei variabile de la nivelul lor mediu.



$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ pentru un eșantion *n* sau $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ pentru o populație *N*

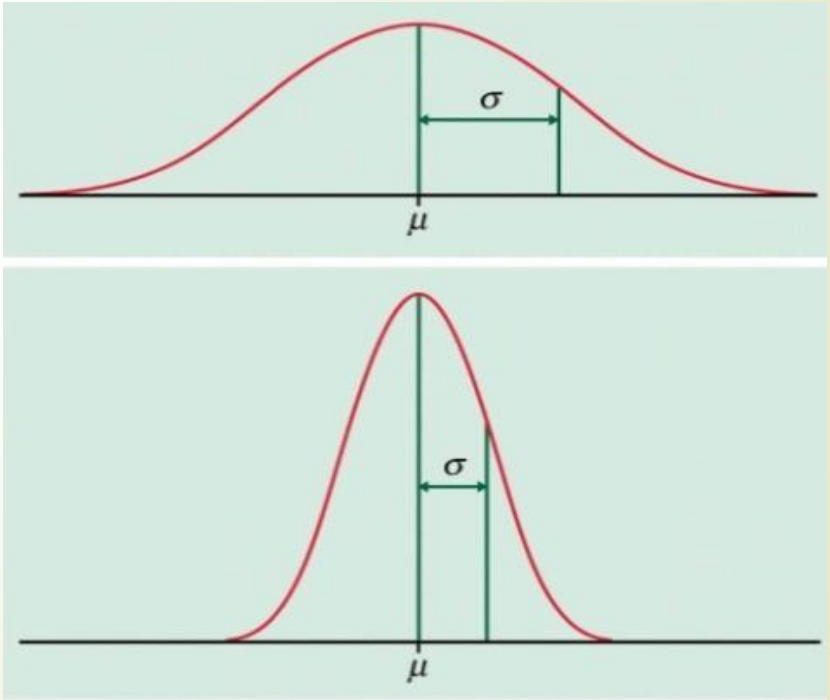
Varianța înlătură deficiența amplitudinii variației, utilizând în cadrul formulei toate variantele variabilei studiate, nu doar valorile minimă și maximă. Deși prezintă o mare importanță în statistică, varianța nu are o interpretare concretă în practică, deoarece ia în calcul pătratele abaterilor variantelor unei variabile față de media lor. Mărimea ei este proporțională cu gradul de variabilitate dintr-o distribuție sau cu gradul ei de împrăștiere.

3. DEVIAȚIA STANDARD

Deviația standard (standard deviation) se determină prin extragerea rădăcinii pătrate din dispersie, reprezentând media pătratică a abaterilor nivelurilor individuale ale caracteristicii față de media lor.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$ pentru un eșantion n sau $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$ pentru o populație N

Abaterea medie pătratică sau deviația standard se interpretează prin raportare la media valorilor variabilei studiate. De exemplu, dacă avem o medie de 100 și o abatere medie pătratică de 5 atunci avem o variație mică (de 5%) in jurul mediei, iar în situația când avem aceeași abatere medie pătratică (5) la o medie de 10, variația este foarte mare (50%). În concluzie, pentru a putea contura o imagine corectă asupra dispersiei datelor se impune raportarea abaterii medii pătratice la valoarea mediei aritmetice.



4. COEFICIENTUL DE VARIAȚIE

Coeficientul de variație se calculează ca raport între abaterea medie pătratică și nivelul mediu al unei variabile și se exprimă în procente, făcând abstracție de unitățile de măsură concrete ale caracteristicilor, putând fi utilizat astfel în comparația gradului de împrăștiere a distribuțiilor cu caracteristica exprimată în unități de măsură diferite.

$v_s = \frac{s}{\bar{x}} \cdot 100$ **pentru un eșantion n sau** $v_\sigma = \frac{\sigma}{\bar{x}} \cdot 100$ **pentru o populație N**

5. CUARTILELE, DECILELE ȘI CENTILELE

Alături de mediană, există și alte mărimi de poziție care împart numărul variantelor ale unei variabile statistice, ordonate crescător sau descrescător, într-un număr de părți egale cu n / k. Aceste mărimi sunt cunoscute sub denumirea de cuantile de ordinul k. Ele se determină în mod asemănător cu mediana.

Dacă: **k = 2** **cuantila poartă denumirea de mediană;**
k = 4 **cuantilele poartă denumirea de cuartile** **(Q_1, Q_2 , Q_3)**
k = 10 **cuantilele poartă denumirea de decile** **(D_1, D_2,…, D_9)**
k = 100 **cuantilele poartă denumirea de centile** **(C_1, C_2,…, C_{99})**

Astfel, există trei cuartile care împart un șir de valori ale unei variabile ordonate crescător în patru părți, în așa fel încât fiecare parte să cuprindă același număr de elemente. Prima cuartilă (Q_1) este acea valoare a unei variabile statistice care împarte o distribuție în două părți, astfel încât un sfert din variantele acesteia au valori mai mici decât Q_1 , iar trei sferturi au valori mai mari decât Q_1. Cuartila a doua (Q_2) împarte numărul variantelor unei VS în două părți egale, astfel încât jumătate dintre variante au valori mai mici decât Q_2 și jumătate au valori mai mari decât Q_2. Q_2 coincide, deci, cu mediana. Cuartila a treia (Q_3) împarte numărul termenilor unei serii în două părți, astfel încât trei sferturi dintre ei au valori mai mici decât Q_3 și un sfert au valori mai mari decât Q_3.

Decilele, în număr de 9, împart un șir de valori ale unei variabile ordonate în 10 părți (intervale) egale. Centilele, în număr de 99, împart variantele variabilei ordonate în 100 de părți egale.

Locația sau poziția unei cuantile k în cadrul șirului ordonat al variantelor variabilei, se determină:

$$L_k = (n+1) \cdot \frac{k}{100}$$

Astfel poziția medianei in cadrul unei serii statistice: $L_{Me} = L_{50} = (n+1) \cdot \frac{50}{100} = \frac{n+1}{2}$

DIAGRAMELE BOX PLOT

o reprezentare grafică care înfățișează forma generală a distribuției unei variabile și este bazată pe 5 indicatori statistici descriptivi: valoarea minimă, valoarea maximă, cuartila 1, cuartila 3 și mediana.

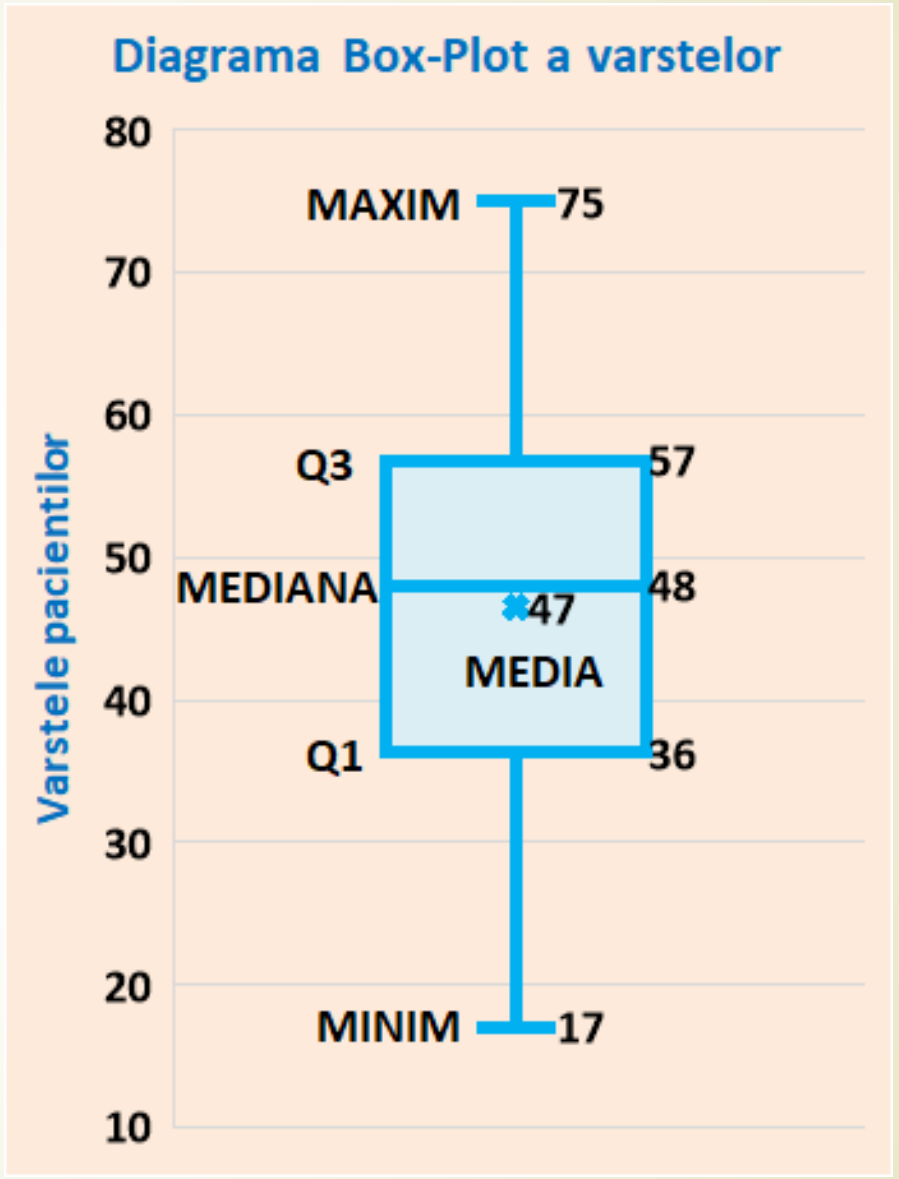
INTERVALUL INTERCUARTILIC

$$I_Q = Q_3 - Q_1$$

Un interval intercuartilic larg este un indicator al unui nivel ridicat de împrăștiere în cadrul intervalului care cuprinde 50% dintre observațiile relevante din centrul seriei statistice.

Este mai utilă însă realizarea unei comparații între mărimea intervalului intercuartilic și amplitudinea variației întregului set de date.

$$I_Q = 57 - 36 = 21\ ani$$



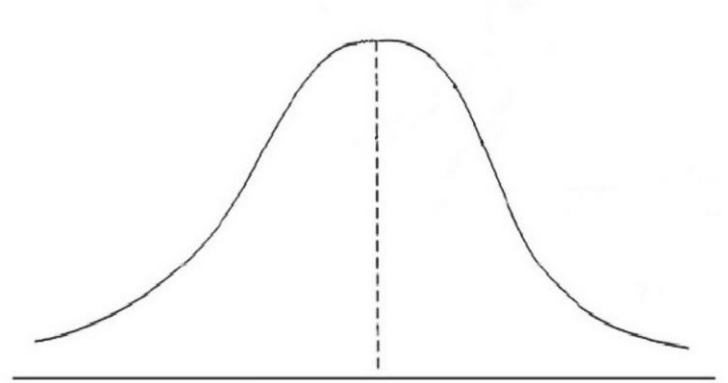
Vârsta mediană a pacienților a fost de 48 de ani, 25% dintre aceștia au sub 36 de ani, iar 25% peste 57 de ani. Jumătate dintre pacienți au vârstele cuprinse între 36 și 57 de ani.

Distribuția vârstelor este aproximativ simetrică, distanța dintre limita inferioară și Q_1 este relativ similară cu cea dintre limita superioară și Q_3. De asemenea, și aria din dreptunghi cuprinsă între Me și Q_1 este egală cu aria dintre Me și Q_3.

Nu este semnalată în cadrul graficului existența unei valori anormale, foarte mici sau foarte mari în comparație cu celelalte valori ale variabilei vârstă, care să fie neobișnuit de departe de restul valorilor seriei.

INDICATORII FORMEI DISTRIBUȚIILOR

Pe lângă tendința centrală și dispersia valorilor unei variabile, o altă caracteristică importantă a unei distribuții statistice este reprezentată de forma acesteia. Un caz particular, având implicații teoretice și practice deosebite în statistică, se referă la situația când media, mediana și modul sunt egale. Este vorba despre bine-cunoscuta CURBĂ NORMALĂ, care are alura unui clopot. Formula matematică a acestei distribuții statistice a fost inițial publicată către matematicianul francez Abraham de Moivre (1667-1754). Contribuții importante a avut și matematicianul german Carl Friederich Gauss (1777-1855), din acest motiv distribuția normală purtând și numele de distribuție gausiană.



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}$$

Această distribuție are mai multe proprietăți, dintre care enumerăm:

- are o repartiție simetrică în jurul mediei aritmetice
- media aritmetică, mediana și modul coincid
- suprafața delimitată de curbă și axa orizontală este egală cu 1
- este complet determinată de 2 parametri (media aritmetică și deviația standard)

1. ASIMETRIA

O distribuție statistică este simetrică, în situația în care jumătatea stângă a graficului său (histograma sau poligonul frecvențelor) reprezintă o imagine în oglindă a jumătății din dreapta. Dacă cele două jumătăți ale graficului nu sunt perfect identice, distribuția respectivă prezintă un anumit grad de asimetrie (skewness).

Atunci când distribuția unei anumite variabile statistice este asimetrică pentru că graficul se extinde mai mult în partea dreaptă comparativ cu cea stângă, aceasta prezintă o asimetrie la dreapta sau pozitivă. În situația opusă, când graficul este mai prelungit în partea stângă decât în dreapta, distribuția este asimetrică la stânga sau prezintă o asimetrie negativă.

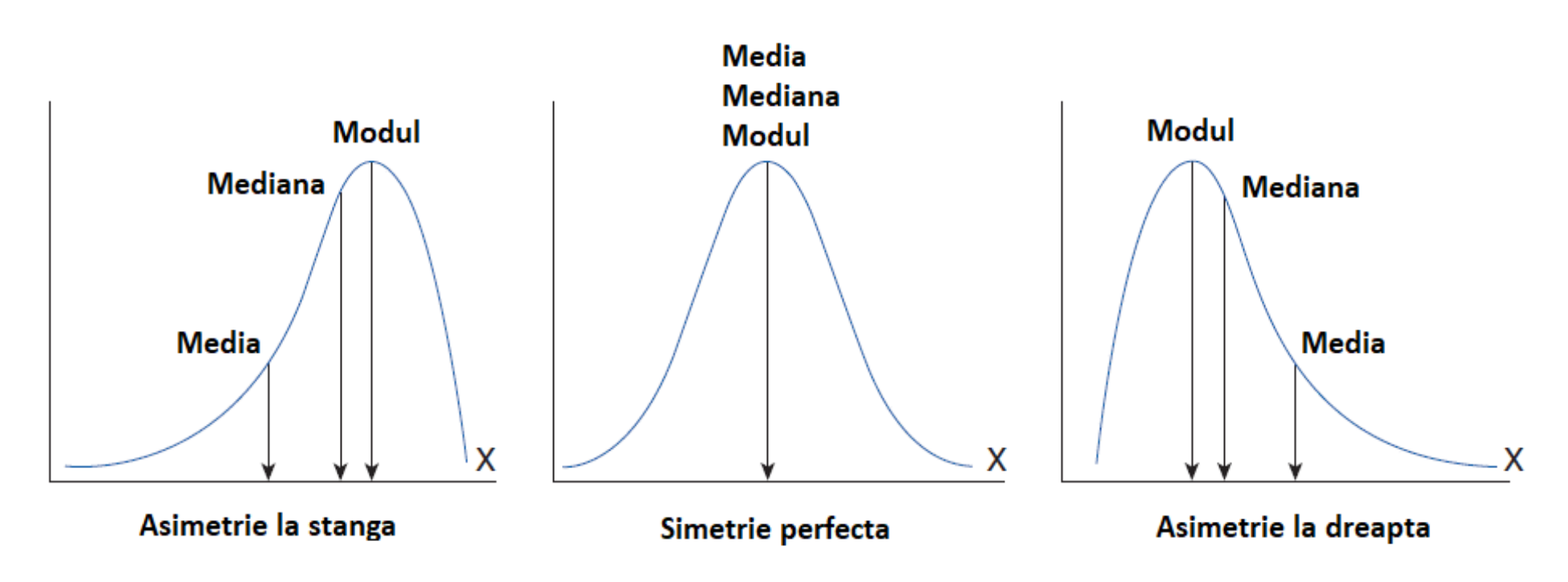
În literatura statistică există mai multe formule de determinare a gradului de asimetrie al unei distribuții statistice. Matematicianul și biostatisticianul englez Karl Pearson (1857-1936) a propus un coeficient bazat pe diferența dintre media aritmetică și mod:

$$S_k = \frac{\bar{x} - Mo}{\sigma} \approx \frac{3 \cdot (\bar{x} - Me)}{\sigma}$$

Coeficientul de asimetrie al lui Pearson poate lua valori între -3 și 3, valoarea zero fiind specifică unei distribuții simetrice, în care media aritmetică, mediana și modul coincid. De exemplu, o valoare de -2,6 indică o asimetrie negativă (la stânga) importantă, iar o valoare de +1,5 arată o asimetrie pozitivă la dreapta (pozitivă) moderată.

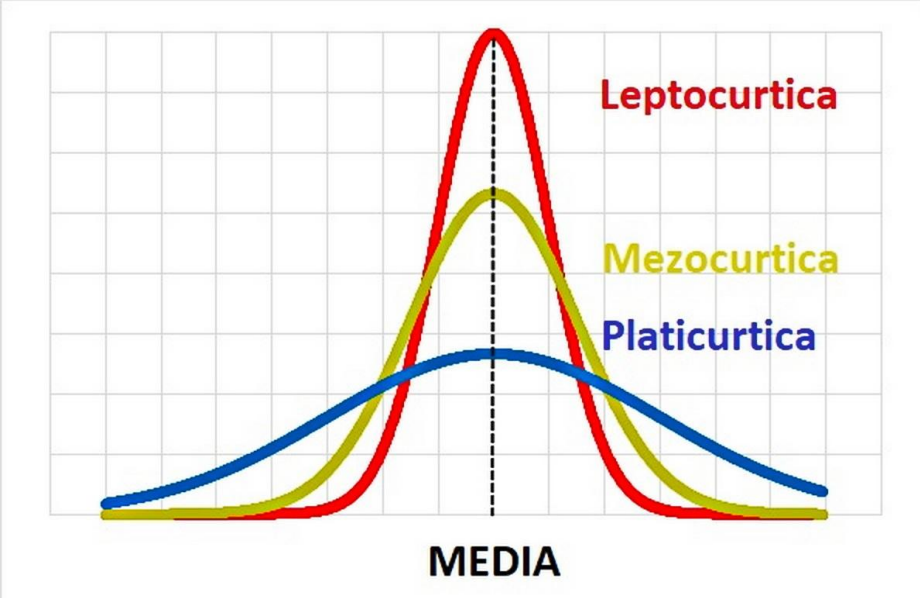
Aplicațiile software din domeniul statisticii utilizează o altă formulă de calcul pentru coeficientul de asimetrie, bazată pe suma abaterilor fiecărei observații de la media aritmetică a variabilei, raportate la deviația standard:

$$S_k = \frac{n}{(n-1)(n-2)} \cdot \sum_{i=1}^{k} \left(\frac{x_i - \overline{x}}{\sigma} \right)^3$$



2. BOLTIREA

Boltirea (kurtosis) reprezintă o măsură a gradului în care o distribuție a unei variabile statistice este mai înaltă sau mai aplatizată în comparație cu distribuția normală, al cărei grafic este de forma unui clopot.



- **Dacă o variabilă statistică prezintă o pondere însemnată a variantelor sau observațiilor în cozile laterale ale graficului (histograma sau poligonul frecvențelor) și implicit un procent mai redus al lor la mijlocul seriei, suntem în situația unei distribuții platicurtice. Este o distribuție în care rezultatele sunt foarte împrăștiate fata de medie, indicând un grad ridicat de eterogenitate a nivelurilor individuale ale variabilei.**
- **În situația în care, o variabilă prezintă o pondere redusă a variantelor sau observațiilor sale în cozile laterale ale graficului și implicit un procent mai însemnat al lor la mijlocul seriei, suntem în situația unei distribuții leptocurtice. O distribuție leptocurtică, ascuțită, arata ca datele sunt foarte grupate si apropiate de medie, lotul de subiecți având un mare grad de omogenitate a valorilor înregistrate.**
- **Distribuția normală gausiană este o distribuție mezocurtică.**

Pentru caracterizarea nivelului de boltire Karl Pearson a propus un coeficient de boltire:

$$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n \cdot \sigma^4} - 3$$

Dacă:

- $\gamma_2 = 0$ **distribuția este mezocurtică**
- $\gamma_2 > 0$ **distribuția este leptocurtică**
- $\gamma_2 < 0$ **distribuția este platicurtică**

IMPORTANȚA PRACTICĂ A DEVIAȚIEI STANDARD

1. SCORURILE Z

Utilizând media și deviația standard putem obține poziția relativă a oricărei valori x_i a variabile statistice $X(x_1, x_2, x_3, \dots, x_i, \dots, x_n)$ supuse analizei. Presupunând că am determinat pe baza datelor din eșantion media aritmetică ($\overline{x}$) și deviația standard (σ), vom avea pentru fiecare variantă x_i a variabilei X o valoare asociată z_i care poartă numele de „scorul z al acesteia”:

$$z_i = \frac{x_i - \overline{x}}{s}$$

Scorul z_i este denumit frecvent „valoarea standardizată a nivelului x_i” al variabilei X și poate fi interpretat ca reprezentând numărul de deviații standard care separă varianta x_i a variabilei X de nivelul mediu $\overline{x}$ al seriei statistice.

Persoana	Înălțimea (x_i)	$x_i - \overline{x}$	$z_i = \frac{x_i - \overline{x}}{s}$
1	170	-6	-0.84
2	177	1	0.14
3	175	-1	-0.14
4	185	9	1.26
5	182	6	0.84
6	168	-8	-1.12
7	172	-4	-0.56
8	178	2	0.28
9	166	-10	-1.40
10	187	11	1.54
SUMA	1760	0	0
MEDIA $(\overline{x})$	176	0	0
DEV STD (s)	7.15	-	1

Persoana nr. 4 din tabel are valoarea z = 1,26, aspect care indică că înălțimea acesteia de 185 cm este mai mare decât înălțimea medie a eșantionului de 176 cm cu 1,26 deviații standard. La polul opus, persoana nr. 9 se situează cu 1,40 deviații standard sub nivelul mediu al înălțimii eșantionului.

Scorul z pentru o variantă a unei anumite variabile poate fi interpretat ca o măsură a poziției relative a acesteia în cadrul seriei. Astfel, variantele a două variabile statistice diferite care au aceeași valoare a scorului z, dețin aceeași poziție în cadrul celor 2 distribuții ale lor, în sensul că sunt la același număr de deviații standard de media fiecărei variabile.

2. TEOREMA LUI CEBÎȘEV

Matematicianul rus Pafnuti Lvovici Cebîșev (1821-1894) a dezvoltat o teoremă care ne permite să determinăm proporția minimă a variantelor unei variabile statistice care se situează într-un anumit număr de deviații standard de la media aritmetică a seriei.

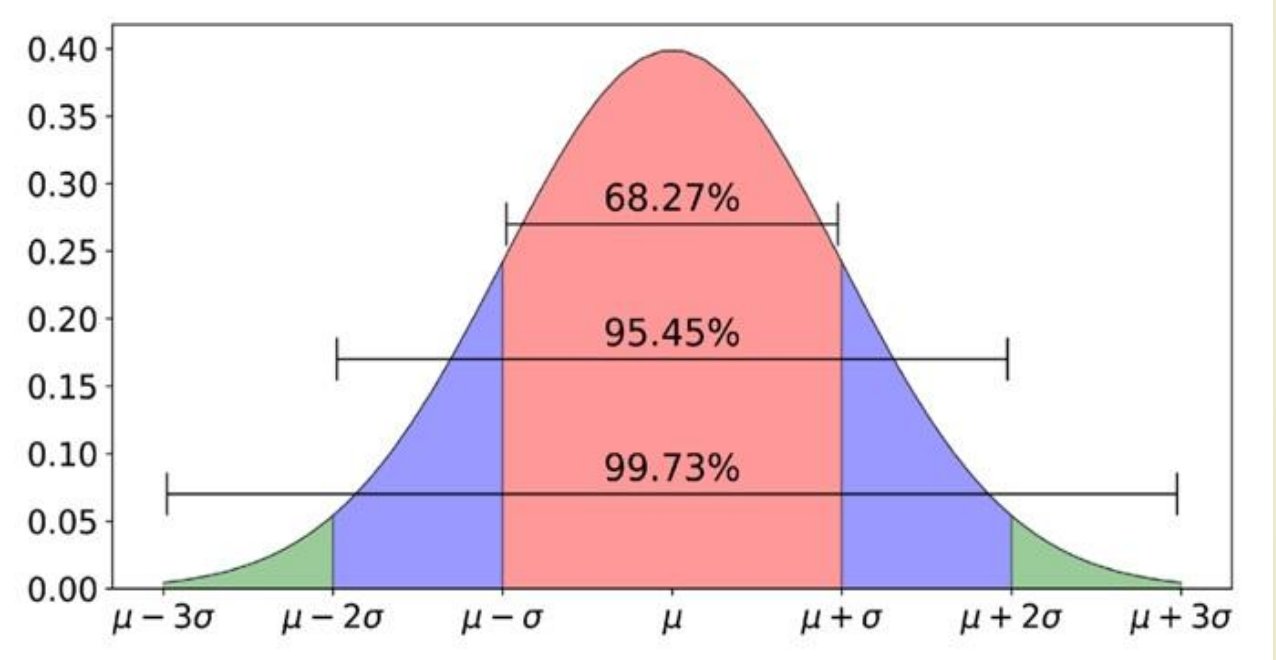
În cazul oricărui set de observații (eșantion sau populație), indiferent de tipul distribuției, proporția valorilor variabilei statistice care sunt cuprinse în intervalul $(\bar{x} - k \cdot \sigma,\ \bar{x} + k \cdot \sigma)$ este de cel puțin $1 - 1/k^2$, unde k reprezintă orice valoare mai mare decât 1.

De exemplu, cel puțin 0,75 sau 75% dintre observațiile unui set de date se vor situa în intervalul $(\bar{x} - 2 \cdot \sigma,\ \bar{x} + 2 \cdot \sigma)$, adică la ± 2 deviații standard în jurul mediei aritmetice.

3. REGULA EMPIRICĂ

Unul dintre avantajele teoremei lui Cebîșev este faptul că nu ține cont de tipul distribuției variabilei studiate. În cazul unei variabile distribuite aproximativ normal, dispersia sau împrăștierea variantelor seriei în jurul mediei aritmetice a acesteia poate fi realizată mult mai precis.

Pentru o distribuție statistică de frecvențe simetrică având alura unui clopot, aproximativ 68% dintre observații vor fi cuprinse între medie și plus minus o deviație standard, aproximativ 95% dintre observații se vor situa între medie și plus minus două deviații standard și aproape toate observațiile (99,7%) se vor situa în intervalul $(\mu - 3 \cdot \sigma,\ \mu + 3 \cdot \sigma)$



Există 95% șanse ca o anumită valoare a unei variabile normal distribuite să se situeze în intervalul

$$(\mu - 1,96 \cdot \sigma,\quad \mu + 1,96 \cdot \sigma)$$