CURSUL AL II-LEA

2. Indicatori statistici

2.1. Serii de valori. Aşa cum s-a văzut în cursul anterior, uneori este necesar să urmărim mai întâi o singură variabilă numerică din multitudinea de variabile înregistrate într-un tabel de date. În acest caz, datele numerice pe care le avem la dispoziţie sunt un simplu **şir de numere** asociate, fiecare din ele, unui individ.

Aceste şiruri de numere rezultate din datele culese le vom numi **serii statistice** sau **serii de date** sau **serii de valori.**

Ceea ce trebuie urmărit în primul rând la o **serie de valori** este modul în care valorile din serie sunt distribuite în plaja de valori între un minim şi un maxim, cum se distribuie în jurul mediei, care este tendinţa centrală a seriei, care sunt valorile cel mai des întâlnite, etc.

Caracterizarea sintetică a unei serii de valori este dată de aşa numiţii **indicatori statistici,** între care media, deviaţia standard, mediana, etc, indicatori pe care îi vom descrie în continuare.

> **Definiţie: Indicatorii statistici** sunt **numere reale**, care sintetizează o parte din informaţia conţinută de o serie de valori, dând posibilitata aprecierii globale a întregii serii, în loc să ţinem cont de fiecare valoare din şir.

Aşa cum se va vedea în acest curs, fiecare indicator urmăreşte să scoată în evidenţă proprietăţi diferite ale şirului de valori.

Astfel, prin combinarea mai multor indicatori, obţinem informaţii relevante şi sintetice despre valorile şirului. Dacă în locul şirului propriu-zis, folosim o serie de indicatori statistici, o parte din informaţie se pierde. Totuşi, de obicei se pierde ceea ce este nesemnificativ, accidental, indicatorii statistici reţinând doar esenţailul. De aici şi utilitatea şi importanţa lor în statistică.

În cele ce urmează, valorile din şirul de numere ce constituie o serie de valori le vom nota cu

$$X:\ x_1, x_2, \ldots\ldots x_n, \text{ sau } Y: y_1, y_2, \ldots . y_n$$

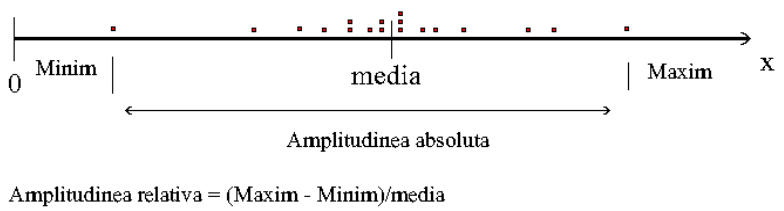
sau notaţii asemănătoare folosind alte litere ale alfabetului.

De exemplu, în loc să spunem că cele 10 valori ale glicemei la cei zece pacienţi dintr-un lot sunt: 88, 97, 103, 89, 93, 105, 98, 105, 88, 103, vom scrie în loc de *Glicemie* litera *X*, şi în locul fiecărui număr din cele zece, simbolurile $x_1, x_2, \ldots x_{10}$. Deci, x_1 ţine locul lui 88, x_2 pe cel al lui 97, etc. Aceste notaţii le folosim pentru a uşura înţelegerea formulelor de calcul pentru unii indicatori.

Valori extreme, amplitudine

Cel mai uşor de căutat şi de înţeles ca semnificaţie sunt indicatorii **Minim** şi **Maxim** care sunt cei ce ne indică *plaja de valori* pe care se întinde seria de valori. Minim este cea mai mică valoare din serie, iar Maxim este cea mai mare.

Amplitudinea absolută, este diferenţa dintre maximul şi minimul unei serii de valori şi ne dă informaţii despre *lărgimea* plajei de valori pe care se întind datele din serie (vezi **figura 1.1**). O serie de valori cu o amplitudine mare indică o plajă de valori întinsă datorată fie unei dispersii sau împrăştieri mari a datelor, fie simplului fapt că sunt multe valori. Dacă două serii de valori au acelaşi număr de valori, dar una are o amplitudine mai mare, atunci valorile ei sunt mai împrăştiate.



Amplitudinea relativa = (Maxim - Minim)/media

Figura 1.1. Indicatorii medie, minim, maxim, amplitudine absolută şi amplitudine relativă.

De cele mai multe ori, valorile minimă şi maximă dintr-o serie nu se înscriu în limitele de normalitate, ceea ce nu înseamnă neapărat că seria conţine valori anormale. Totuşi, de obicei, cele mai îndepărtate câteva valori, atât cele mai mici cât şi cele mai mari trebuie verificate pentru a ne asigura că nu este vorba de date eronate.

De exemplu, deşi se consideră că valorile normale pentru latenţa semnalului nervos pe nervul optic între stimularea retinei şi răspunsul cortical sunt situate aproximativ între 90 ms şi 115 ms, un eşantion de indivizi sănătoşi poate să producă o serie de valori care are şi una sau câteva excepţii. De aceea, din 20 sau 30 de valori, una poate fi 88 ms iar alta 117 ms, majoritatea fiind însă între 90 şi 115 ms.

2.2. Valori medii. **Media aritmetică** a unei serii de valori. Este un indicator simplu şi în acelaşi timp foarte sintetic, fiind un foarte bun indiciu al valorii în jurul căreia se grupează datele. Se notează cu litera *m* sau, dacă seria de valori este notată cu o majusculă ca X sau Y, media se notează cu $\overline{X}$ sau $\overline{Y}$. Formula este cea cunoscută:

$$\overline{X} = \frac{x_1 + x_2 + \ldots\ldots\ldots x_n}{n} = m \qquad (1.1)$$

Definiţie:

Media aritmetică unei serii de valori este raportul dintre suma valorilor seriei şi numărul lor.

Media este indicatorul care arată tendinţa centrală a seriei de valori, şi de obicei arată unde tind datele să se aglomereze. De cele mai multe ori, valorile din serie sunt situate în majoritate în apropierea mediei, iar o mai mică parte din ele sunt situate mult în stânga sau în dreapta mediei. O situare a valorilor din serie faţă de medie se poate observa din aşa-numitul grafic punctual de dispersie, din care este dat un exemplu în **figura 1.2**

Figura 1.2. Cele mai multe valori sunt de obicei mai apropiate de medie.

Dar nu totdeauna datele din seria de valori se situează preponderent în apropierea mediei. Mai rar, şi oarecum mai forţat, ne putem întâlni şi cu situaţii în care datele din serie se situează preponderent în stânga şi dreapta, departe de medie şi doar o mică parte dintre ele se situează aproape de medie, aşa cum se observă în **figura 1.3**.

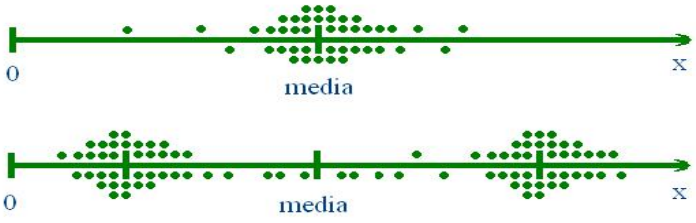


Figura 1.3. Uneori, cele mai multe valori sunt sub medie şi peste medie, destul de departe de aceasta. În seriile de mai sus, avem aceeaşi medie, dar este evident că nu avem aceeaşi situaţie. Valorile din seria de jos sunt mai împrăştiate.

Astfel, dacă în acelaşi lot sunt cuprinşi indivizi hipertiroidieni şi hipotiroidieni, şi se măsoară la fiecare concentraţia hormonului tiroidian T_4, vom observa că hipotiroidienii au preponderent valori în stânga mediei, cei mai mulţi destul de departe de medie, iar hipertiroidienii au preponderent valori în dreapta, tot departe de medie.

De fapt într-un asemenea caz, în zona centrală lipsesc exact ceea ce am spune că sunt normalii, adică indivizi care au valori pentru T_4 uşor peste medie şi uşor sub medie, şi care nu au fost incluşi într-un astfel de lot.

Evident că un eşantion aşa de eterogen nu este folosit prea des în statistică pentru că, aşa cum vom vedea, în acest caz este foarte indicat să se constituie două eşantioane distincte pentru cele două categorii de pacienţi. Totuşi, asemenea situaţii, chiar dacă de obicei nu sunt indicate şi sunt puţin artificiale, există. Situaţia de mai sus este ilustrată în **figura 1.3.**

O formulă simplificată pentru media aritmetică este dată de:

$$\overline{X} = \frac{x_1 \cdot F_1 + x_2 \cdot F_2 + \ldots\ldots\ldots + x_n \cdot F_n}{F_1 + F_2 + \ldots\ldots\ldots + F_n}$$

unde cu *n* am notat numărul de valori diferite din seria de valori, iar $F_1, F_2, \ldots, F_n$ sunt frecvenţele de apariţie în serie ale valorilor $x_1, x_2, \ldots, x_n$.

Această formulă se spune că este formula pentru **media ponderată**. Nu trebuie să credem că media ponderată calculată cu formula de mai sus şi media aritmetică calculată cu formula (1.1), sunt indicatori diferiţi. Ambele medii sunt în realitate identice. Media ponderată se calculează de obicei mai simplu şi deci nu reprezintă decât o formă mai simplă de calcul al mediei aritmetice.

Prin faptul că este un indicator extrem de fidel al tendinţei centrale al unei serii statistice, media este un indicator extrem de mult utilizat în statistică. Media aritmetică are dezavantajul că este sensibilă la valori extreme fie foarte mici, fie foarte mari. Adăugarea unei singure valori (sau a câtorva) mult mai mari decât celelalte, modifică sensibil media aritmetică.

De asemenea, dacă datele sunt distribuite în jurul mediei puternic asimetric, media îşi pierde din puterea de a evoca tendinţa centrală, în aceste cazuri fiind mult mai utilă mediana (vezi mai jos).

2.3. Împrăştiere. Valorile dintr-o serie de valori pot fi mai aglomerate în jurul mediei sau mai dispersate, adică la distanţe mari de medie. Un mod de a măsura aceste abateri de la medie este să se facă diferenţa între toate aceste valori şi media lor. Unele abateri vor fi pozitive, altele negative. Ele nu pot fi adunate, deoarece, prin adunare dau suma 0.

Dispersia. Un mod de a ocoli faptul că suma abaterilor absolute este 0, este ridicarea la pătrat a acestora înainte de a fi adunate, pentru a face să dispară semnele negative la unele şi pozitive la altele.

Suma obţinută, ar trebui împărţită la numărul de abateri pentru a se obţine o medie. În realitate, din motive teoretice foarte bine întemeiate, dar mai greu de explicat în cuvinte simple, împărţirea se face la *n-1* şi nu la *n*. Motivul pentru care se face acest lucru va fi înţeles mai bine în contextul unor noţiuni enunţate la cursul despre teoria estimaţiei. Valoarea care se obţine astfel se numeşte **dispersie** şi este un indicator al gradului de împrăştiere al seriei. Dispersia se notează cu *D* şi are formula:

$$D = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots\ldots + (x_n - \overline{X})^2}{n-1}$$

După cum se observă, numărătorul fracţiei din definiţia dispersiei este cu atât mai mare cu cât abaterile individuale de la medie sunt mai mari şi deci este natural să considerăm că o valoare mare a dispersiei arată o împrăştiere mare a valorilor din serie.

De fapt, este bine de reţinut că:

- La medii aproximativ egale, este mai împrăştiată seria cu dispersia mai mare.
- La dispersii aproximativ egale, este mai împrăştiată seria cu media mai mică.

Dispersia are dezavantajul că se exprimă cu unităţile de măsură ale valorilor din serie, ridicate la pătrat, şi are în general valori foarte mari comparativ cu abaterea medie. De exemplu, dacă valorile din serie se măsoară în mg/l, atunci dispersia se măsoară în mg^2/l^2, ceea ce este în mod evident extrem de nenatural.

În plus, dacă abaterile absolute au o medie, de exemplu în jurul lui 10, dispersia va avea o valoare în jurul lui 100, adică exagerat de mare în comparaţie cu abaterile absolute. De aceea se mai foloseşte un alt indicator, numit abatere standard care este radicalul dispersiei.

Abaterea standard. Se notează cu σ şi are formula:

$$\sigma = \sqrt{D} \text{ sau } \sigma = \sqrt{\frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots\ldots + (x_n - \overline{X})^2}{n-1}}$$

Acest indicator se exprimă cu aceeaşi unitate de măsură ca şi valorile din seria considerată şi este un indicator foarte fidel al împrăştierii seriei. Abaterea standard, nu are dezavantajele dispersiei, adică unitatea de măsură este aceeaşi cu a valorilor din serie, şi, are o valoare comparabilă cu abaterile individuale de la medie.

Exemplu de calcul:

Să presupunem că am măsurat zilnic tensiunea arterială sistolică la doi pacienţi timp de 10 zile, obţinând pentru fiecare următoarele valori:

- 170, 180, 160, 180, 190, 190, 180, 190, 170, 190, pentru primul pacient şi
- 160, 170, 190, 160, 190, 190, 200, 180, 180, 180, pentru al doilea.

Lăsând la o parte studiul modului cum evoluează de la zi la zi tensiunea pacienţilor, care este bineînţeles importantă, să ne propunem să determinăm care are tensiunea cu valori mai împrăştiate, indiferent de evoluţia în timp.

Notând prima serie cu *X* iar pe a doua cu *Y* se constată uşor că ambele au media 180 (datele nu sunt reale, au fost deliberat alese ca să simplifice calculele). Atunci, vom avea pentru abaterile de la medie şi pentru pătratele lor următoarele valori:

- $x_i - \overline{X}$: -10, 0, -20, 0, 10, 10, 0, 10, -10, 10. $\overline{X} = 180$.
- $y_i - \overline{Y}$: -20, -10, 10, -20, 10, 10, 20, 0, 0, 0. $\overline{Y} = 180$.
- $(x_i - \overline{X})^2$: 100, 0, 400, 0, 100, 100, 0, 100, 100, 100.
- $(y_i - \overline{Y})^2$: 400, 100, 100, 400, 100, 100, 400, 0, 0, 0.

Deci vom avea pentru D_x:

$$D_x = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots\ldots + (x_{10} - \overline{X})^2}{10 - 1} = \frac{400 + 6 \cdot 100}{9} = \frac{1000}{9} = 111{,}1$$

şi cu un calcul absolut analog, $D_y = 1600 / 9 = 177{,}7$. Se observă că, în timp ce abaterile de la medie sunt de ordinul zecilor, dispersiile sunt de ordinul sutelor, ceea ce este destul de nenatural, şi în plus, după cum am mai spus, unitatea de măsură este cu totul alta.

Pentru abaterile standard, vom avea:

$$\sigma_x = \sqrt{D_x} = \sqrt{111{,}1} = 10{,}5$$
$$\sigma_y = \sqrt{D_y} = \sqrt{177{,}7} = 13{,}3$$

calculele fiind făcute cu aproximaţie. Deci, este ceva mai împrăştiată seria *Y*.

De fapt, este bine de reţinut că:

- La medii aproximativ egale, este mai împrăştiată seria cu deviaţia standard mai mare.
- La deviaţii standard aproximativ egale, este mai împrăştiată seria cu media mai mică.

Ce se întâmplă însă dacă mediile şi deviaţiile sunt foarte diferite? Atunci o bună apreciere se obţine dacă se foloseşte raportul deviaţiei standard faţă de medie, exprimat în procente, acest raport fiind un alt indicator al împrăştierii valorilor dintr-o serie. Acest indicator se numeşte **coeficient de variaţie**.

Coeficientul de variaţie. Este raportul dintre deviaţia standard şi medie, atunci când media este diferită de 0 şi se exprimă în procente:

$$C.V. = \frac{\sigma}{\overline{X}}$$

Pentru seriile de mai sus, coeficientul de variaţie este mai mare pentru cea mai împrăştiată, adică pentru cea cu deviaţia standard mai mare:

- $C.V._x = 10{,}5 / 180 = 0{,}058 = 5{,}8\ \%$.
- $C.V._y = 13{,}3 / 180 = 0{,}073 = 7{,}3\%$.

Totuşi, seriile de mai sus sunt comparabile cu ajutorul abaterilor standard, deoarece au aceeaşi medie, şi, aşa cum s-a văzut, la medii egale sau aproximativ egale, are valorile mai împrăştiate seria cu abaterea standard mai mare.

Aprecierea cu ajutorul coeficientului de variaţie se face mai ales atunci când două serii de valori au medii mult diferite şi deviaţiile standard pot să nu ne dea o indicaţie suficient de utilă. De exemplu, măsurând ***latenţa* şi *amplitudinea* semnalului electric pe nervul optic** la 120 de pacienţi cu *scleroză multiplă*, s-au obţinut următoarele rezultate:

- *Latenţa* medie: 113,6
- Abaterea standard a *latenţei:* 14,7
- *Amplitudinea* medie: 2,68
- Abaterea standard a *amplitudinii*: 2,03

Dacă dorim să apreciem împrăştierea valorilor din cele două serii, abaterile standard nu ne sunt de ajutor. Într-adevăr, latenţa are o abatere standard mult mai mare decât amplitudinea, dar şi media latenţei este cu mult mai mare decât aceea a amplitudinii. De aceea, în acest caz, doar coeficientul de variaţie ne permite o apreciere corectă a împrăştierilor, în vederea comparării lor:

- Pentru latenţă: $C.V._{latenţa} = \frac{14,7}{113,6} = 0,129 = 12,9\%$
- Pentru amplitudine: $C.V._{amplitudine} = \frac{2,03}{2,68} = 0,757 = 75,7\%$

Se observă că valorile amplitudinii sunt cu mult mai împrăştiate decât cele ale latenţei. Acest fapt se datorează atât unei variabilităţi biologice mai mari la amplitudine decât la latenţă, cât şi unei variabilităţi datorate aparatelor de măsură, care măsoară latenţa cu mai multă precizie, în timp ce la măsurarea amplitudinii, erorile de măsurare sunt mai mari.

Coeficientul de variaţie este cel mai fidel indicator al împrăştierii unei serii statistice, dar are şi el un inconvenient, este cu atât mai fidel cu cât mediile sunt mai depărtate de 0.

La medii foarte apropiate de 0 îşi pierde din fidelitate şi nu este indicat să fie folosit. Acest lucru se întâmplă mai ales atunci când valorile din serie sunt şi negative şi pozitive, şi când, din acest motiv, media poate fi aproape de 0.

2.4. Indicatori de asimetrie. Atunci când valorile unei serii sunt distribuite nesimetric în jurul mediei, acest fapt este imposibil de surprins cu ajutorul indicatorilor de dispersie. De aceea, s-au introdus indicatori care să pună în evidenţă şi acest aspect al seriilor de valori: excentricitatea, sau asimetria. Va trebui să ţinem cont atât de numărul de valori care sunt în stânga şi în dreapta mediei, cât şi depărtarea lor faţă de medie.

Mediana. Este un indicator al tendinţei centrale, şi anume este valoarea de mijloc, într-o serie de valori.

Definiţie:

Mediana este acea valoare dintr-o serie de valori, pentru care exact jumătate din ele sunt mai mici decât ea, iar jumătate mai mari.

Altfel spus, este valoarea măsurată pentru individul din mijloc, dacă indivizii pe care s-au făcut măsurătorile ar fi ordonaţi crescător. Pentru o înţelegere mai uşoară, să luăm un exemplu cu numai 10 înregistrări: tensiunea arterială maximă la un bolnav în 10 zile:

150, 160, 160, 170, 160, 170, 150, 160, 170,160.

Dacă se aşază aceste valori într-un şir crescător, obţinem:

150, 150, 160, 160, 160, 160, 160, 170,170,170.

În acest caz, mediana se ia între a cincia şi a şasea valoare din acest şir ordonat, adică 160. Dacă aceste două valori de mijloc diferă, se ia media lor aritmetică. Dacă numărul de măsurători este impar atunci madiana este chiar valoarea de mijloc, care în acest caz este unică.

De fapt, mediana este importantă în primul rând la serii de valori cu foarte multe înregistrări, caz în care se poate lucra direct pe tabelul de frecvenţă, sau chiar pe tabelul pe clase.

Pentru a exemplifica modul cum se caută mediana pe tabelul de frecvenţă, vom lua **tabelul 1.3**, în care sunt centralizate vârstele a 234 de pacienţi, fiecare valoare a vârstei având o anumită frecvenţă absolută F_i, o frecvenţă relativă f_i şi o frecvenţă relativă cumulată crescător, f_{icc} (vezi mai sus, pentru amănunte).

Valoarea medianei se culege din coloana întâi, a vârstelor, dar pentru a şti care valoare trebuie aleasă, trebuie să privim pe ultima coloană, a frecvenţelor cumulate, f_{icc}, în dreptul frecvenţei cumulate de 50%.

Se observă că, pe coloana frecvenţelor cumulate, nu există frecvenţa de 50%, dar, există frecvenţa de 47,9%, care este prea mică, şi frecvenţa de 53,8%, care este prea mare. În acest caz, mediana se citeşte din dreptul primei frecvenţe cumulate crescător care depăşeşte 50%, în cazul nostru, în dreptul frecvenţei de 53,8%, şi pe coloana ***Vârsta*** citim 55 ani. Deci, vârsta mediană este 55 ani.

Tabelul 1.3. Vârstele a 234 de pacienţi centralizate într-un tabel de frecvenţă

Varsta	Fi	fi	ficc
26	1	0.4%	0.4%
28	1	0.4%	0.9%
29	1	0.4%	1.3%
30	2	0.9%	2.1%
31	2	0.9%	3.0%
32	1	0.4%	3.4%
35	3	1.3%	4.7%
36	2	0.9%	5.6%
37	3	1.3%	6.8%
38	1	0.4%	7.3%
40	3	1.3%	8.5%
41	5	2.1%	10.7%
42	1	0.4%	11.1%
43	4	1.7%	12.8%
44	10	4.3%	17.1%
45	6	2.6%	19.7%
46	6	2.6%	22.2%
47	5	2.1%	24.4%
48	13	5.6%	29.9%
49	2	0.9%	30.8%
50	4	1.7%	32.5%
51	6	2.6%	35.0%
52	9	3.8%	38.9%
53	10	4.3%	43.2%
54	11	4.7%	47.9%
55	14	6.0%	53.8%
56	11	4.7%	58.5%
57	9	3.8%	62.4%
58	19	8.1%	70.5%
59	5	2.1%	72.6%
60	9	3.8%	76.5%
61	13	5.6%	82.1%
62	5	2.1%	84.2%
63	4	1.7%	85.9%
64	4	1.7%	87.6%
65	6	2.6%	90.2%
66	2	0.9%	91.0%
67	4	1.7%	92.7%
68	3	1.3%	94.0%
69	4	1.7%	95.7%
70	1	0.4%	96.2%
71	2	0.9%	97.0%
72	2	0.9%	97.9%
74	1	0.4%	98.3%
77	2	0.9%	99.1%
78	2	0.9%	100.0%
Total	234	100.0%	

Deci, vom spune că jumătate dintre pacienţi au vârstele cuprinse între 26 şi 55 ani şi jumătate au vârstele mai mari decât 55 ani. Această alegere este permisă în cazul acesta al vârstelor care se înregistrează cu valori întregi.

Mediana este un indicator al tendinţei centrale, ca şi media, dar oferă mai puţină informaţie decât aceasta din urmă. La distribuţiile echilibrate, la care valorile din serie se dispun aproximativ simetric în stânga şi în dreapta mediei, media şi mediana sunt foarte apropiate, deci folosirea medianei este superfluă. Dacă însă mediana este mult în stânga sau în dreapta mediei, distribuţia se zice că este excentrică.

De exemplu, venitul median este mai informativ decât venitul mediu deoarece distribuţia veniturilor într-o populaţie este foarte excentrică, fiind foarte mulţi indivzi cu salarii foarte mici şi foarte puţini indivizi cu salarii foarte mari.

Cuartilele. În mod asemănător cu căutarea medianei, se poate pune problema căutării unor valori pentru care să avem un sfert din valorile seriei mai mici şi respectiv, mai mari.

Definiţie:

> **Cuartila Q_1 este acea valoare dintr-o serie de valori, pentru care 25% din valorile seriei sunt sub Q_1 şi 75%, peste**.

Pentru tabelul de frecvenţe 1.3, cuartila Q_1 se caută în dreptul frecvenţei relative cumulate crescător de 25%. În tabel găsim procentul de 24,4% şi în dreptul lui vârsta de 47 de ani, precum şi frecvenţa de 29,9 şi în dreptul ei vârsta de 48 de ani. Vom lua tot vârsta care corespunde primului procent peste 25%, adică 48 de ani.

Definiţie:

> **Cuartila Q_3 este acea valoare dintr-o serie de valori, pentru care 75% din valorile seriei sunt sub Q_3 şi 25%, peste**.

Pentru tabelul 1.3, cuartila Q_3 se ia din dreptul frecvenţei relative cumulate crescător de 75%. Poate fi luată cu aproximaţie, 60 ani.

Care este utilitatea medianei şi cuartilelor în aprecierea simetriei distribuţiei? Pentru a sublinia utilitatea indicatorilor Q_1 şi Q_3, să considerăm şirul vârstelor:

- cel mai tânăr pacient,
- Q_1,
- Vârsta mediană,
- Q_3,
- cel mai în vârstă pacient.

Pentru **tabelul 1.3**, obţinem şirul: 26 ani, 48 ani, 55 ani, 60 ani, 69 ani.

- Se observă că sfertul (25%) pacienţilor cei mai tineri este situat în zona 26 - 48 de ani adică într-o plajă de 22 de ani.
- Sfertul următor, este intre 48 şi 55 de ani, adică pe un interval de doar 7 ani.
- Al treilea sfert este situat între 55 şi 60 de ani, adică pe 5 ani,
- Cei mai în vărstă 25 % din pacienţi sunt între 60 şi 69 de ani, pe un interval de 9 ani.

Putem să spunem că vârstele pacienţilor se distribuie **uşor asimetric**, deoarece:

1. Sfertul cel mai tânăr se distribuie pe o plaja de 22 de ani, iar cel mai în vârstă pe o plajă de doar 9 ani.
2. Sfertul al doilea se distribuie pe 7 ani, iar al treilea doar pe 5 ani.

În cadrul laboratorului, alte exemple vor arăta utilitatea acestor indicatori.

Să mai observăm că mediana este într-un fel "cuartila de 50%", adică Q_2. Se spune că există trei cuartile: Q_1, mediana, Q_3.

Decile Uneori, loturi mai mari de multe sute de indivizi trebuie urmărite foarte atent în ceea ce priveşte modul cum sunt distribuite valorile şi de aceea s-au introdus indicatorii **decile,** care sunt de o acurateţe mai bună decât cuartilele. Sunt 9 decile, fiecare corespunzând unui procent de 10%, 20%, ... 90% din lot, asemănător cu cuartilele. Decila 5, sau de 50%, este de fapt mediana.

Centilele (percentilele) sunt mai rar folosite, în studii pe mii de cazuri, de obicei de un interes mai larg, naţional, internaţional, în studii epidemiologice, şi sunt corespunzătoare procentelor de 1%, 2%,...99% din lot. Centila de 25% este cuartila Q_1, cea de 50% este mediana, iar cea de 75% este cuartila Q_3. Centilele de 10%, 20%,….90%, sunt cele nouă decile. Centilele dau o imagine destul de exactă a distribuţiei valorilor dintr-o serie de valori foarte mare. Nu are rost să calculăm centile pentru serii cu câteva sute de valori, pentru că erorile sunt prea mari şi imaginea obţinută este deformată.

Modul. Dintre frecvenţele absolute apărute într-un tabel de frecvenţe, una este maximă. Clasa sau valoarea corespunzătoare acestei frecvenţe maxime se numeşte **mod.** Modul este de obicei un indicator al tendinţei centrale. În tabelul 1.2. modul este clasa de la 55 la 60 de ani, cu frecvenţa absolută 53. De obicei, frecvenţele absolute au tendinţa de a creşte către mod, după care urmează o descreştere continuă. Modul este deci o indicaţie relativă la maximul frecvenţelor absolute. Sunt însă distribuţii la care se înregistrează creşteri şi descreşteri astfel încât pot apare două moduri sau chiar mai multe. Aceste distribuţii sunt mai rare şi au un caracter cu totul special. Ele se numesc distribuţii **bimodale** sau **multimodale**, după caz.

Este un indicator care poartă în el puţină informaţie despre datele seriei. Modul este mult influenţat de fluctuaţii aleatoare şi nu este prea recomandat pentru a aprecia tendinţa centrală a valorilor dintr-o serie. Mai mult, unele distribuţii pot fi multimodale, caz în care modul nu mai indică prea mult despre tendinţa centrală.

Excentricitate. (Engl. Skew, Skweness). Este un indicator al asimetriei şi este luat de diverşi autori cu diverse formule. Distribuţiile cu excentricitate pozitivă sunt mai des întâlnite decât cele cu excentricitate negativă. În medicină, parametrii fiziologici sunt în majoritate modificaţi în diverse afecţiuni în sensul că au valori peste normal. Astfel, tensiunea arterială o vom întâlni la valori normale, crescute sau scăzute. Cum indivizi cu valori foarte mari, vom întâlni cu atât mai rar cu cât valoarea este mai mare, distribuţia va avea o coadă spre dreapta. La fel la mulţi alţi parametric cum ar fi bilirubina, transaminazele, colesterolul, lipemia, etc.

Totuşi, vom întâlni şi parametri care se distribuie cu asimetrie stânga în patologii: hemoglobina, calcemia, sodiul ionic, etc. Hemoglobina, de exemplu, se poate distribui cu frecvenţă mai mare la valori relativ normale şi cu frecvenţe din ce în ce mai mici pe măsură ce coborâm la valori mai mici. Chiar dacă avem o patologie de tip anemie, ne aşteptăm ca frecvenţa în jurul a 9-10 să fie mai mare decât frecvenţa în jurul a 7-8, frecvenţă care ne aşteptăm să fie foarte mică.

Excentricitatea unei serii de valori x_1, x_2,…..x_n, se calculează cu formula:

$$sk = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^3}{n\sigma^3}$$

Cu cât o distribuţie este mai simetrică cu atât *sk* tinde la 0. Ca o regulă generală, **la distribuţiile cu excentricitate pozitivă, media este mai mare decât mediana**. Evident, media este mai mică decât mediana la distribuţiile cu excentricitate negativă. Există cazuri rare în care regula de mai sus nu este valabilă.

Sunt multe alte formule pentru alţi coeficienţi de excentricitate şi când vorbim despre excentriciatte, trebuie să menţionăm la ce coeficient de excentricitate ne referim. Uneori se foloseşte un coeficient de asimetrie care măsoară diferenţa dintre medie şi mediană, eventual raportată la abaterea standard sau la intervale intercuartilice($Q_3 - Q_1$). Indiferent ce formulă se foloseşte, o excentricitate egală cu zero, sau foarte apropiată de zero, este un indiciu al simetriei repartiţiei valorilor din serie. Din contră, excentricităţi mult diferite de 0, peste 0,15 -0,20, sau mai jos de -0,15 -0,20 sunt indicii ale asimetriei. Dăm mai jos, cu titlu facultativ, câteva formule pentru coeficienţi de excentricitate.

$$sk_1 = \frac{\overline{X} - Mo}{\sigma} \quad sk_2 = \frac{3(\overline{X} - Me)}{\sigma} \quad sk_3 = \frac{2(Q_3 + Q_1 - 2Me)}{Q_3 - Q_1} \quad sk_4 = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

Boltirea (facultativ). Boltirea este un indicator care se bazează pe lungimea cozilor unei distribuţii. Cele cu cozi relativ mari se numesc **leptocurtice** iar cele cu cozi relativ mici se numesc **platicurtice** (vezi **figura 1.4**). Formula de calcul a boltirii este:

$$k = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^4}{n\sigma^4} - 3$$

Aşa cum se va vedea în capitolul despre repartiţii, boltirea este un indicator util în aprecierea apropierii repartiţiei de repartiţia normală. Distribuţiile din **figura 1.4** au aceeaşi medie, aceeaşi dispersie, aproximativ aceeaşi excentricitate dar diferă mult ca boltire.

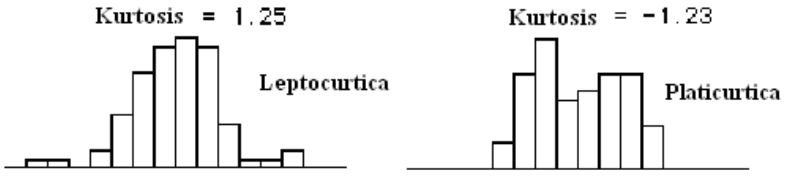


Figura 1.4. Distibuţie leptocurtică şi distribuţie platicurtică.

2.5. Clasificarea indicatorilor

Indicatorii statistici poartă în ei, fiecare, o anumită cantitate de informaţie, din seria de valori pentru care au fost calculaţi. Aşa cum s-a văzut în paragrafele precedente, unii indicatori ne dau informaţii despre tendinţa centrală a valorilor din serie, alţii ne dau informaţii despre împrăştierea valorilor, alţii ne dau indicaţii despre simetria valorilor din serie, boltirea ne dă indicaţii despre lungimea cozilor distribuţiei, etc.

Informaţia oferită de indicatorii statistici este redundantă, în sensul că, de exemplu, împrăştierea valorilor din serie este indicată şi de dispersie şi de abaterea standard şi de amplitudinea absolută şi de coeficientul de variaţie, etc. Totuşi, fiecare din ei aduce o mică informaţie specifică, deci, nu ne putem lipsi de unul sau altul

dintre indicatorii statistici. Uneori trebuie folosiţi unii dintre indicatori, fiind cei mai eficienţi, alteori trebuie folosiţi alţii.

Pentru a avea o ideie despre modul cum trebuie folosiţi indicatorii statistici, ei sunt clasificaţi în câteva categorii mai importante, categorii care vor fi exemplificate mai jos, insistând pe aceia care sunt cei mai importanţi, restul fiind indicatori mai rar folosiţi, numai în cazuri speciale.

Indicatori ai tendinţei centrale. Cei mai importanţi indicatori ai tendinţei centrale sunt **media**, **mediana** şi **modul**. Media indică tendinţa centrală atunci când seria de valori este repartizată simetric în jurul ei şi când valorile nu au o dispersie exagerat de mare. În cazul seriilor de valori distribuite foarte asimetric, tendinţa centrală nu mai este indicată de către medie, ci de către mediană.

Modul, este un indicator al tendinţei centrale, la seriile unimodale, adică atunci când în tabelul de frecvenţe există un singur maxim. Dacă avem o serie multimodală, modul îşi pierde calitatea de indicator al tendinţei centrale.

Indicatori ai împrăştierii. Folosiţi mai des în practică, şi deci mai importanţi, sunt **dispersia, abaterea standard** şi **coeficientul de variaţie**.

Abaterea standard este indicatorul folosit cel mai des pentru aprecierea împrăştierii, dar atunci când mediile diferă mult, este mai util coeficientul de variaţie. Dispersia este folosită ca măsură a împrăştierii în testele statistice (vezi capitolul dedicat testelor statistice).

Indicatori ai asimetriei. **Mediana** şi **cuartilele** sunt cel mai mult folosite pentru aprecierea asimetriei valorilor dintr-o serie. De fapt, mediana se foloseşte în combinaţie cu media pentru aprecierea asimetriei. O mediană mult diferită de medie indică asimetrie puternică, iar o mediană foarte apropiată de medie indică o tendinţă spre simetrie.

Cuartilele, se folosesc în combinaţie cu mediana şi indicatorii minim şi maxim, pentru aprecierea simetriei.

Indicatorii statistici fundamentali. Sunt indicatorii care poartă în ei cea mai mare cantitate de informaţie din informaţia conţinută de seria de valori.

La seriile de valori distribuite relativ simetric, indicatorii statistici fundamentali sunt **media** şi **deviaţia standard**. În capitolul dedicat repartiţiilor, se va vedea că, dacă o serie de valori are o repartiţie normală şi are suficient de multe valori, cei doi indicatori, poartă în ei aproape toată informaţia. Astfel, dacă o serie de valori de acest tip are media $\overline{X}$ şi deviaţia standard σ, scrierea încetăţenită este $\overline{X} \pm \sigma$

La seriile distribuite asimetric, deşi se consideră ca indicatori fundamentali tot media şi deviaţia standard, sunt mai utile **mediana** şi **cuartilele**. În acest caz, este încetăţenită scrierea medianei *M* şi a cuartilelor Q_1 şi Q_3 în forma $M\ [Q_1;\ Q_3]$. De exemplu, dacă o serie puternic asimetrică are mediana 2,45, iar cuartilele sunt Q_1=1,54 şi Q_3=6,23, acest fapt se precizează astfel: 2,45 [1,54; 6,23].

3. Chestiuni de examen:

1. Definiţia si formula mediei

2. Formula deviaţiei standard si a coeficientului de variaţie

3. Definiţia medianei si a cuartilelor Q1, Q3

4. Media unei serii de valori numerice este:
   - **A. Suma valorilor împărţită la numărul lor**
   - **B. Mai mare decât valoarea minimă din serie**
   - **C. Mai mică decât valoarea maximă din serie**
   - **D. Un indicator al tendinţei centrale a valorilor seriei**

5. Media unei serii de valori numerice are următoarele proprietăţi:
   - A. Este egală cu cea mai mică valoare din serie
   - **B. Dacă schimbăm o valoare din serie, mărind-o, media se schimbă, mărindu-se**
   - C. Dacă schimbăm o valoare din serie, mărind-o, media se schimbă, micşorându-se
   - D. Dacă ştergem o valoare din serie, media râmîne nemodificată

6. Media unei serii de valori numerice este un indicator al:
   - **A. Tendinţei centrale a valorilor seriei**
   - B. Împrăştierii valorilor seriei
   - C. Plaja de valori între care sunt cuprinse valorile seriei
   - D. Media nu este indicator statistic

7. Dispersia unei serii de valori numerice este un indicator al:
   - A. Tendinţei centrale a valorilor seriei
   - **B. Împrăştierii valorilor seriei**
   - C. Plaja de valori între care sunt cuprinse valorile seriei
   - D. Simetriei distribuţiei valorilor seriei în jurul mediei

8. Dispersia unei serii de valori numerice are printre dezavantaje:
   - **A. Se măsoară cu unitatea de măsură a valorilor seriei, ridicată la pătrat**
   - **B. Are valori prea mari, comparativ cu abaterile individuale de la medie**
   - C. Indică şi tendinţa centrală a valorilor seriei
   - D. Nu se poate calcula cu exactitate

9. Abaterea standard unei serii de valori numerice are printre avantaje:
   - **A. Se măsoară cu unitatea de măsură a valorilor seriei**
   - **B. Are valori comparabile cu abaterile individuale de la medie**
   - C. Indică şi tendinţa centrală a valorilor seriei
   - D. Nu se poate calcula dacă dispersia este negativă

10. Dacă două serii de valori au aproximativ aceeaşi medie, atunci:
   - **A. Este mai împrăştiată cea cu dispersia mai mare**
   - B. Este mai împrăştiată cea cu abaterea standard mai mică
   - C. Sunt la fel de împrăştiate
   - D. Nu se pot compara împrăştierile cu ajutorul dispersiei în acest caz

11. Dacă două serii de valori au medii foarte diferite, atunci:
   - A. Este mai împrăştiată cea cu dispersia mai mare
   - B. Este mai împrăştiată cea cu abaterea standard mai mare
   - **C. Nu se pot compara nici cu ajutorul dispersiei şi nici cu ajutorul abaterii standard**
   - D. Au aceeaşi împrăştiere

12. Dacă media unei serii de valori este 10 şi dispersia 4, atunci coeficientul de variaţie este:
   - A. 40%
   - **B. 20%**
   - C. 80%
   - D. 10%

13. Dacă mediile a două serii de valori sunt foarte diferite, iar abaterile standard sunt tot foarte diferite, atunci este mai împrăştiată :
   - **A. Cea cu coeficientul de variaţie mai mare**
   - **B. Cea cu raportul dintre abaterea standard şi medie mai mare**
   - C. Cea cu coeficientul de variaţie mai mic
   - D. Împrăştierile celor două serii de valori nu se pot compara

14. Mediana unei serii de valori numerice este:
   - A. Egală cu media
   - B. Un grafic
   - **C. Un număr**
   - D. Un tabel de frecvenţă

15. Mediana unei serii de valori numerice este:
   - **A. Valoarea pentru care jumătate din valorile seriei sunt mai mari şi jumătate mai mici**
   - B. Valoarea situată la mijloc, între minimul seriei şi maximul seriei

C. Valoarea cea mai frecvent întâlnită printre valorile seriei
D. Un indicator al excentricităţii valorilor seriei

16. Dacă o serie de valori are în componenţă 21 de numere, atunci, pentru aflarea medianei, se ordonează valorile crescător şi se ia:
A. Valoarea a 11-a din şirul ordonat
B. Media între valorile a 10 şi a 11-a
C. Media între valorile a 11 şi a 12-a
D. Valoarea a 10-a din şirul ordonat

17. Dacă o serie de valori are în componenţă 24 de numere, atunci, pentru aflarea medianei, se ordonează valorile crescător şi se ia:
A. Valoarea a 12-a din şirul ordonat
B. Media între valorile a 11-a şi a 12-a
C. Media între valorile a 12-a şi a 13-a
D. Valoarea a 13-a din şirul ordonat

18. Cuartila întâi a unei serii de valori este:
A. Valoarea din seria ordonată situată la 25% din numărul de valori al seriei
B. Valoarea din seria ordonată situată la 75% din numărul de valori al seriei
C. Valoarea numerică pentru care un sfert din valorile seriei ordonate sunt mai mici
D. Valoarea numerică pentru care un sfert din valorile seriei sunt mai mici

19. Cuartila a treia a unei serii de valori este:
A. Valoarea din seria ordonată situată la 25% din numărul de valori al seriei
B. Valoarea din seria ordonată situată la 75% din numărul de valori al seriei
C. Valoarea numerică pentru care un sfert din valorile seriei ordonate sunt mai mici
D. Valoarea numerică pentru care trei sferturi din valorile seriei ordonate sunt mai mari

20. Referitor la indicatorii decile, este adevărat:
A. Avem exact nouă decile
B. Avem exact 99 de decile
C. Decila 50 este mediana
D. Decila a treia este mediana

21. Indicatorii statistici fundamentali sunt:
A. Dispersia şi media
B. Media şi abaterea standard
C. Abaterea standard şi mediana
D. Mediana şi cuartilele

22. Indicatorii de dispersie (sau de împrăştiere) sunt:
A. Amplitudinea, media, dispersia şi mediana
B. Abaterea standard, media, dispersia şi mediana
C. Amplitudinea, media, dispersia şi abaterea standard
D. Abaterea standard, dispersia şi coeficientul de variaţie

23. Care din următorii indicatori statistici ajută la aprecierea asimetriei:
A. Mediana, media şi excentricitatea
B. Mediana, cuartilele şi excentricitatea
C. Mediana, cuartilele şi media
D. Mediana, dispersia şi excentricitatea

24. Indicatorii statistici pentru tendinţa centrală a valorilor unei serii de valori sunt:
A. Media, dispersia şi mediana
B. Media, abaterea standard şi modul
C. Media, dispersia şi excentricitatea
D. Media, mediana şi modul