Curs 4 – Indicatori sintetici ai distribuției

Lect. Adrian Gorbănescu

Statistics Mall se mândrește că toate produsele vândute sunt la prețuri convenabile pentru toți clienții. Dacă vrei să îți cumperi un parfum, o pereche de pantofi sau o poșetă cu siguranță vei găsi tot ce îți dorești la prețul convenabil pentru tine. Proprietarii au observat faptul că oamenii sunt fericiți când găsesc magazine cu prețuri pe măsura puterii lor de cumpărare, iar clienții fericiți mereu vor reveni la ei pentru a-și face cumpărăturile. Cheia succesului pentru Statistics Mall constă în organizarea magazinelor în funcție de prețuri, ceea ce implică calcularea ***prețului mediu*** *și afișarea lui pe ușa magazinului. Acesta este o* ***valoarea reprezentativă*** *pentru magazin și îi ajută pe clienți să intre în magazinul cu prețuri accesibile pentru venitul lor.*

De cele mai multe ori, statistica descriptivă ne oferă un set de indicatori numerici cu rolul de a descrie distribuția pe care dorim să o analizăm. În capitolul anterior ne-am familiarizat cu analiza de frecvențe, dar informațiile obținute dintr-o astfel de analiză nu sunt suficiente. În plus, analiza de frecvențe presupune utilizarea întregii distribuții, iar munca devine tot mai complicată pe măsură ce setul de date se mărește. Așadar, avem nevoie de indicatori care să ne permită conturarea unei imagini cât mai complete despre o distribuție fără a utiliza întreaga cantitate de date. Acesta este rolul **indicatorilor sintetici** ai distribuțiilor statistice. Prin **indicator sintetic** înțelegem un „descriptor numeric care condensează într-o valoare unică o anumită caracteristică a întregii distribuții de valori" (Popa, 2008, p. 58).

În statistică există trei tipuri de indicatori sintetici, în funcție de informațiile pe care le rezumă la nivelul distribuției:

- **Indicatori ai tendinței centrale** – oferă indicatori reprezentativi pentru distribuție.
- **Indicatori ai variabilității (împrăștierii)** – oferă indicatori care ne informează despre cât de împrăștiate (diversificate) sunt valorile din distribuție. Cu alte cuvinte, cât de mult diferă scorurile între ele.
- **Indicatori ai formei distribuției** – ne oferă informații despre forma distribuției. Distribuțiile pot fi diferite ca formă. Unele pot fi simetrice, iar altele pot fi boltite în zona scorurilor mari și aplatizate în zona scorurilor mici. În concluzie, acești indicatori ne oferă informații despre reprezentarea grafică a distribuției.

În cadrul fiecăreia din cele trei categorii se află indicatori specifici, pe care îi vom prezenta în continuare.

4.1 Indicatori ai tendinței centrale

4.1.1 Media aritmetică (m)

Media aritmetică (pe scurt, **media**; lb. engleză, **mean**) este cel mai des utilizată în statistică. Orice analiză pe date cantitative presupune calcularea mediei aritmetice. În literatura de specialitate în limba engleză se utilizează termenul **mean**. Media este reprezentată prin simbolul **m** la nivel de eșantion și prin **μ** (litera grecească **„miu"**) la nivelul populației. Din motive lesne de înțeles (vezi capitolul 1, noțiunile de eșantion și populație), de cele mai multe vom lucra cu media eșantionului. Media se calculează după aceeași formulă atât la nivelul eșantionului cât și la nivelul populației. Astfel, calcularea mediei pentru o distribuție presupune adunarea tuturor valorilor și împărțirea la numărul lor.

$$m = \frac{\sum X}{N}$$

(formula 4.1)

- ∑X – indică suma tuturor valorilor din distribuție.
- N – este volumul eșantionului.

Statisticienii utilizează litere pentru a prezenta numere. De exemplu litera **X** reprezintă fiecare valoare dintr-o distribuție. Astfel dacă avem o distribuție cu șapte valori ele vor fi prezentate prin: X_1, X_2, X_3, X_4, X_5, X_6 și X_7.

Dar dacă nu cunoaștem câte scoruri sunt într-o distribuție? Există o soluție pentru această problemă – vom nota numărul scorurilor cu **n**. Dacă nu știm câte numere sunt într-o distribuție vom spune că sunt **n** și le vom scrie astfel: X_1, X_2, X_3 …. X_n. În acest caz, **Xn** reprezintă a n-a valoare din distribuție, iar simbolul „…" este un mod de a spune „și așa mai departe".

Exemplu

Profesorului de Statistică îi place foarte mult înghețata și obișnuiește să mănânce cel puțin câte una pe zi. La sfârșitul unei săptămâni, într-un moment în care calculează bugetul familiei, se gândește „*Oare nu am cheltuit cam mulți bani pe înghețată? Câți bani am alocat, în medie, pe zi pentru consumul de înghețată în ultima săptămână?*". Mai jos sunt prezentate sumele cheltuite în fiecare zi din săptămână.

Luni	Marți	Miercuri	Joi	Vineri	Sâmbătă	Duminică
10	15	13	13	12	15	13

$$m = \frac{\sum X}{N} \rightarrow m = \frac{10+15+13+13+12+15+13}{7} \rightarrow m = \frac{91}{7} \rightarrow m = 13.$$

Atunci când calculăm media unei distribuții se întâmplă, de multe ori, ca unele numere să se repete. Dacă analizăm exemplul de mai sus se poate observa că într-o zi s-au cheltuit 10 lei, în alte 3 zile s-au alocat 13 lei, suma de 15 lei se regăsește tot de 2 ori, în timp ce suma de 12 lei s-a cheltuit doar într-o singură zi. Este important să ne asigurăm că în calcul am inclus **frecvența** fiecărui scor din distribuție. Astfel, media din exemplul de mai sus putea fi calculată și astfel:

$$m = \frac{\sum fX}{\sum f}$$

(formula 4.2)

- ∑fx – reprezintă multiplicarea fiecărei valori cu frecvența sa absolută și însumarea rezultatelor obținute după efectuarea înmulțirilor.
- ∑f – este suma frecvențelor absolute.

$$m = \frac{\sum fX}{\sum f} \rightarrow m = \frac{10*1 + 15*2 + 13*3 + 12*1}{7} \rightarrow m = \frac{91}{7} \rightarrow m = 13$$

Să ne întoarcem la Statistics Mall, acolo unde fiecare magazin are afișat pe ușă prețul mediu, acesta fiind un reper pentru clienți. În funcție de prețul mediu, clienții știu dacă în magazin sunt produse pe care și le pot permite sau nu. În magazinul A, unde se vând parfumuri, clienții intră fiind atrași de prețul mediu și nu cumpără aproape nimic. Managerul magazinului nu găsește explicații pentru acest fenomen. În magazin sunt următoarele produse:

- HB la prețul de 20 lei (8 produse)

- C la prețul de 21 lei (5 produse)
- A la prețul de 18 lei (8 produse)
- Urzicescu la prețul de 1 leu (25 produse)
- B la prețul de 15 lei (5 produse)

$$m = \frac{\sum fX}{\sum f} \rightarrow m = \frac{20*8 + 21*5 + 18*8 + 1*25 + 15*5}{51} \rightarrow m = \frac{160+105+144+25+75}{51} \rightarrow m = 9{,}98$$

Calculând prețul mediu al produselor din magazin am obținut o medie de 9,98 lei. Ce este în neregulă cu acest magazin? Observați diferența dintre prețurile produselor? Aproximativ jumătate din produsele magazinului au prețuri cuprinse între 15 și 21 de lei, iar cealaltă jumătate este reprezentată un singur produs care are prețul de 1 leu. În acest caz, produsul cu prețul de 1 leu, reprezintă o **valoare extremă** (lb. engleză, **outliers**). Acesta face ca media prețurilor să scadă. Astfel, clienții intră în magazin fiind atrași de prețul mediu, dar când intră observă că cele mai multe din brandurile de parfum vândute au prețuri superioare mediei. În esență, prin **valoare extremă** *ne referim la valori foarte mici sau foarte mari comparativ cu celelalte scoruri din distribuție*.

Din acest exemplu putem înțelege că media poate fi ușor afectată, atât prin scăderea, cât și prin creșterea ei ca urmare a efectului scorurilor extreme. Acest tip de valori poate fi descoperit cu ajutorul graficului boxplot (vezi capitolul 3).

Miles și Banyard (2007) recomandă calcularea mediei în următoarele situații:

a. Distribuția este simetrică – se traduce prin faptul că valorile se împart omogen de o parte și de alta a mediei. Această asumpție face referire și la lipsa valorilor extreme. Media poate fi calculată și atunci când distribuția nu este simetrică, dar ea va fi greu de interpretat deoarece ar putea fi rezultatul unor valori înșelătoare.
b. Valorile sunt măsurate pe scală de interval/raport. Media poate fi măsurată doar atunci când avem date cantitative. Așa cum am mai amintit și în capitolele anterioare, nu putem calcula media culorii ochilor sau a trăsăturilor de personalitate.

Trebuie să avem în vedere că atunci când utilizăm un soft pentru analiza datelor, el nu poate verifica dacă datele sunt eligibile pentru un calcul sau pentru altul. Astfel, dacă noi atribuim coduri numerice pentru variabile categoriale (de exemplu, genul, orașul natal, profesia etc.) și îi cerem să calculeze media, acesta o va afișa, deși ea este lipsită de logică.

Media nu este doar un concept de bază în statistică sau în științele experimentale, ea apărând de foarte multe ori în viața de zi cu zi a fiecăruia dintre noi. Majoritatea indicatorilor numerici raportați în jurnale științifice sunt medii. Statistica inferențială de cele mai multe ori studiază diferențe dintre medii (Pollatsek, Lima, & Well, 1981). În aceste condiții este important să reținem proprietățile mediei aritmetice:

a. Suma abaterii tuturor valorilor față de medie este egală cu 0. Prin abatere de la medie ne referim la diferența dintre fiecare scor și medie (**X - m**). Să analizăm, din nou, exemplul cu sumele cheltuite într-o săptămână de profesorul de Statistică pe înghețată. În tabelul de mai jos, pe linia *X-m*, sunt prezentate abaterile individuale de la medie (m = 13). Suma abaterilor individuale față de medie este egală cu 0. Cu alte cuvinte, $\sum(X - m) = 0$.
b. Suma pătratului abaterii tuturor valorilor de la medie este mai mică decât suma pătratului abaterilor față de orice valoare din distribuție. Prin ridicarea la pătrat a fiecărei abateri de la medie $(X-m)^2$ și apoi prin adunarea lor vom obține $\sum(X-m)^2$. Dacă vom calcula abaterea fiecărui scor față de o altă valoare din distribuție, de exemplu 12,

$\sum(X-m)^2$ (în cazul nostru, 18) va fi mai mică decât $\sum(X-12)^2$ (25). La fel se va întâmpla și dacă vom calcula suma pătratelor abaterilor față de 10 sau 15.

Zi	**Luni**	**Marți**	**Miercuri**	**Joi**	**Vineri**	**Sâmbătă**	**Duminică**	
Suma cheltuită	**10**	**15**	**13**	**13**	**12**	**15**	**13**	**∑ = 91**
X-m	**-3**	**2**	**0**	**0**	**-1**	**2**	**0**	**∑ = 0**
$(X-m)^2$	**9**	**4**	**0**	**0**	**1**	**4**	**0**	**∑ = 18**
$(X-12)^2$	**4**	**9**	**1**	**1**	**0**	**9**	**1**	**∑ = 25**

c. Adăugarea/scăderea unei constante la fiecare valoare din distribuție determină creșterea/scăderea mediei cu acea constantă.
d. Multiplicarea/împărțirea fiecărei valori a distribuției cu o constantă generează multiplicarea/împărțirea mediei cu acea constantă.

Exemplu de calcul

În magazinul B, din Statistics Mall, se vând poșete. În stoc se află 4 poșete LV (24 lei/buc), 5 poșete He (28 lei/buc), 6 poșete Pr (18 lei/buc) și 5 poșete CC. Care este prețul unei poșete CC, dacă prețul mediu afișat de magazin este de 20 lei? Deoarece nu cunoaștem prețul produselor CC, îl vom nota cu **a**.

Plecând de la formula $m = \frac{\sum fX}{\sum f}$ putem identifica:

- m = 20
- ∑f = Suma frecvențelor absolute ale produselor din magazin. Astfel, ∑f = 4 + 5 + 6 + 5 → ∑f = 20.
- ∑fX = m * ∑f → ∑fX = 20 * 20 → ∑fX = 400.
- ∑fX = 24*4 + 28*5 + 18*6 + **a***5 → 400 = 96 + 140 + 114 + 5**a** → 5**a** = 400 – (96 + 140 + 114) → 5**a** = **400 – 350** → 5**a** = 50 → **a** = 50/5 → **a** =10

În concluzie, prețul unei poșete CC este de 10 lei.

4.1.2 Mediana (Me)

Dacă media devine înșelătoare din cauza valorilor extreme, atunci avem nevoie de un alt indicator care să ne spună cine este valoarea reprezentativă a distribuției. Putem face acest lucru selectând valoarea din mijlocul distribuției. Aceasta este **mediana** (lb. engleză, **median**). Cu alte cuvinte, **mediana** *este valoarea care împarte distribuția în două jumătăți (50% din scoruri sunt mai mici decât ea, iar 50% din valori sunt mai mari)*. Astfel, putem înțelege că *mediana corespunde rangului percentil 50 sau quartilei 2 (Q_2)*. În concluzie, **mediana este quartila 2 (Q2)**.

Pentru a afla mediana în lista cu sumele de bani plătite pe înghețată de profesorul de Statistică, vom ordona distribuția crescător și vom selecta valoarea din mijloc, precum în exemplul de mai jos:

10 12 13 13 13 15 15

Acesta este mijlocul distribuției. Mediana este 13

Așadar, am aflat că mediana din lista cu sumele de bani cheltuite pe înghețată este 13.

Dar dacă în distribuție ar fi fost un număr par de valori? Atunci când în distribuție există un număr par de scoruri, pentru a afla mediana trebuie calculată media aritmetică a celor două valori din mijloc (le adunăm și apoi împărțim la 2).

Exercițiu

În Statistics Mall se află un celebru magazin cu pantofi. Produsele și prețul lor sunt afișate în tabelul de mai jos. Calculați media aritmetică și aflați mediana pentru distribuția de jos.

Brand	V	TH	RL	Pr
Frecvența	4	5	3	2
Prețul	15	18	21	30

În ceea ce privește mediana trebuie să avem în vedere faptul că ea nu reflectă întotdeauna valorile din distribuție. Pot exista situații în care apar scoruri extreme fără ca mediana să își schimbe valoarea. De asemenea, trebuie să ținem cont că mediana nu poate fi utilizată în cazul statisticilor inferențiale.

4.1.3 Modul (Mo)

Pe lângă medie și mediană, mai există un indicator al tendinței centrale – **modul** (lb. engleză, **mode**). Într-o distribuție, **modul** este valoarea cu frecvența absolută cea mai mare. Cu alte cuvinte, modul este valoarea care apare de cele mai multe ori în cadrul unei distribuții. În distribuția de mai jos, **Mo** = 13 (apare de cele mai multe ori).

10 12 13 13 13 15 15

Uneori distribuțiile pot avea mai multe moduri. Aceasta se întâmplă atunci când sunt mai multe scoruri care au cea mai mare frecvență absolută. Dacă sunt două scoruri cu ceva mai mare frecvență de apariție vom spune că distribuția este **bimodală**. Atunci când sunt cel puțin trei valori cu cea mai mare frecvență distribuția este **multimodală**.

În distribuția 5, 6 , 8 , 8, 7, 5, 10, 8, 7, 5 modul este reprezentat de valorile 5 și 8 (apar de câte trei ori) – **Mo** = 5, 8.

Exercițiu

Presa are informații că managerul companiei Statistics Icecream, domnul Pofticiosu, oferă angajaților săi salarii mici, deși aceștia sunt performanți, iar profitul este considerabil în fiecare lună. Pentru a dezminți informațiile, managerul se hotărăște să organizeze o conferință de presă în care să prezinte detalii despre salariile angajaților.

Reporter: Domnule Pofticiosu, sunt adevărate zvonurile despre salariile angajaților din compania dumneavoastră? Care este media salariilor în compania pe care o conduceți?

Dl. Pofticiosu: Informațiile apărute în presă sunt false. Angajații Statistics Icecream sunt foarte bine plătiți, salariul mediu fiind de aproximativ 4550 lei pe lună.

Reporter: Da, este un salariu mediu foarte bun. Ne puteți spune unde se află mediana?

Dl. Pofticiosu: Salariul corespunzător medianei este de 5000 de lei.

Reporter: Datele furnizate până acum ne arată că salariile nu sunt mici deloc. Dar…ne puteți spune care este modul? Vrem să știm care este modul! Avem lista salariilor din compania dumneavoastră!

Angajat 1	2000
Angajat 2	2000
Angajat 3	2000
Angajat 4	2000
Angajat 5	2000
Secretară 1	5000
Secretară 2	5000
Angajat HR	5000
Contabil	7500
Dna Pofticiosu	8000
Dl Pofticiosu	10000

Dl. Pofticiosu: …

1. Identificați modul în distribuția salariilor.
2. Explicați de ce informațiile apărute în presă sunt adevărate.

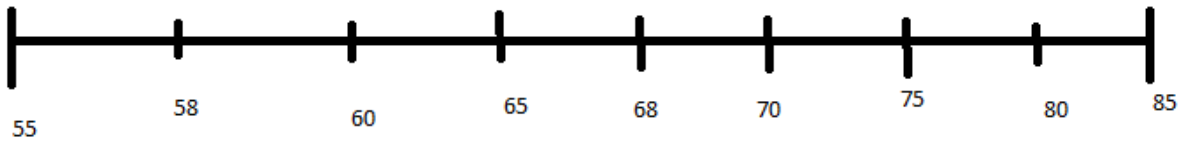
4.2 Indicatori ai împrăștierii

Indicatorii tendinței centrale ne informează în legătură cu valorile reprezentative ale distribuției. Totuși, am observat că aceștia nu pot surprinde întotdeauna realitatea, putând fi modificați de valori extreme sau, în cazul distribuțiilor mici, de adăugarea unor valori noi. Cu alte cuvinte, indicatorii tendinței centrale surprind doar un element al setului de date. Atunci când analizăm o distribuție de scoruri putem fi interesați să observăm cât de diversificate sunt scorurile, cât de mult diferă unele de altele. Indicatorii împrăștierii surprind diversitatea valorilor și cât de mult se îndepărtează ele de tendința centrală (de medie). Asemenea indicatorilor tendinței centrale, și în cazul împrăștierii, există mai mulți indicatori care ne explică diversitatea valorilor.

Magazinele din Statistics Mall promovează o amplă campanie de reduceri pentru toate produsele. Într-un week-end managerul complexului comercial se hotărăște să înregistreze sumele cheltuite de clienți. În acest sens, la ieșirea din Mall clienții care doresc să participe la studiu pot completa o bază de date electronică în care sunt rugați să specifice genul și suma cheltuită. Sumele cheltuite de bărbați sunt cuprinse între 55 și 85 de lei, iar sumele cheltuite de femei variază între 55 și 75 de lei. În medie, bărbații au cheltuit 68 de lei, în timp ce femeile au cheltuit 65 de lei. În concluzie, persoanele înregistrate în baza de date au cheltuit sume cuprinse între 55 și 85 de lei, cheltuindu-se în medie 66,50 lei de persoană.

4.2.1 Amplitudinea (R)

În statistică, **amplitudinea** (lb. engleză, **range**) exprimă diferența dintre valoarea cea mai mare și valoarea cea mai mică din distribuție. Cea mai mică valoare din setul de date se numește *limită inferioară* (lg. engleză, **lower bound**), iar valoarea cea mai mare reprezintă *limita superioară* (lb. engleză, **upper bound**).



Să aruncăm o privire peste sumele cheltuite de clienții magazinelor din Statistics Mall. Pentru a afla amplitudinea trebuie să calculăm diferența dintre valoarea cea mai mare și valoarea cea mai mică.

R = valoarea maximă – valoarea minimă
(formula 4.3)

R = 85 – 55 → R = 30.

Amplitudinea este o cale simplă de a estima cât de împrăștiate sunt datele, oferindu-ne o nouă modalitate de a compara seturile de date.

Exercițiu

Pe baza situației descrise mai sus, calculați amplitudinea pentru distribuția sumelor cheltuite de bărbați, respectiv de femei.

În Statistics Mall există două magazine cu pantofi. Prețurile din cele două magazine sunt prezentate în tabelele de mai jos. Observați diferența dintre prețurile practicate de cele două magazine?

Magazin „**I love shoes**”

Preț	10	12	15	18	20	22	25	27
Frecvență	4	3	6	5	4	3	5	3

Magazin „**All I want is shoes**”

Preț	10	12	15	18	20	22	25	27
Frecvență	3	0	0	0	9	10	13	15

Ambele magazine au aceeași amplitudine (17). Cu toate acestea, prețurile sunt distribuite diferit. Mă întreb dacă amplitudinea ne oferă o imagine adevărată despre împrăștierea datelor dintr-o distribuție?

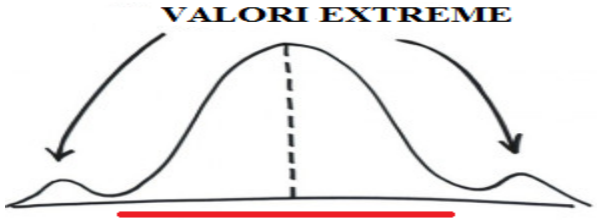
Amplitudinea descrie doar distanța dintre cele două limite ale distribuției, nu și cât de diferite sunt valorile între ele. Cu alte cuvinte, amplitudinea ne prezintă cât de mult se îndepărtează cele două limite, dar este dificil să obținem o imagine despre modul în care sunt datele distribuite. Deși presupune o metodă simplă de calcul, amplitudinea nu este cea mai sigură cale de a afla cât de împrăștiate sunt scorurile. Dacă setul de date cuprinde valori

extreme, utilizarea amplitudinii ca modalitate de analiză a împrăștierii poate fi înșelătoare. În situația de mai sus, ambele magazine au aceeași amplitudine, dar cel de-al doilea magazin are o valoare extremă (10). Uitându-ne peste prețurile din al doilea magazin putem observa că cele mai multe produse au prețul cuprins între 20 și 27 de lei, prețul de 10 lei fiind o valoare extremă. În cazul primului magazin prețurile se împrăștie omogen între 10 și 27 de lei.

Amplitudinea este un indicator simplu de utilizat și foarte ușor de înțeles de majoritatea oamenilor, chiar dacă nu au experiență de lucru în statistică. Dacă discutăm de amplitudinea vârstei unui grup de oameni, toată lumea înțelege la ce ne referim. Cu toate acestea, amplitudinea este un indicator statistic sensibil la valorile extreme.

4.2.2 Abaterea interquartilă (R_Q)

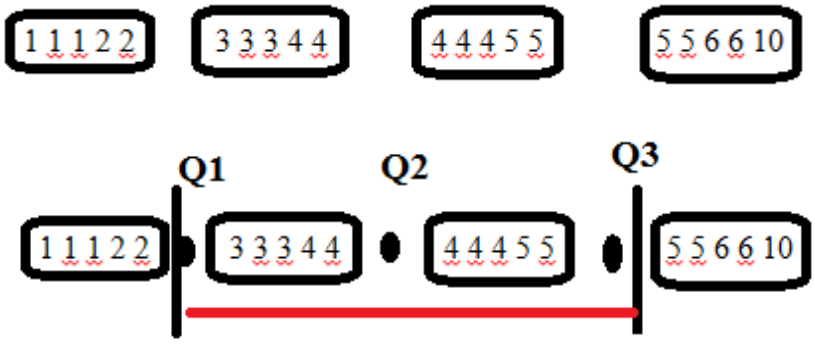
Principala problemă a amplitudinii, dacă ne raportăm la definiția ei, este aceea că include și valorile extreme dacă acestea există în setul de date. De aceea avem nevoie de o cale prin care să eliminăm efectul pe care aceste valori l-ar putea avea astfel încât să putem descrie împrăștierea cât mai fidel.



Să calculăm amplitudine pentru aceste valori

O metodă de a scăpa de efectele negative ale valorilor extreme este reprezentată de calcularea unui fel de *mini amplitudine*, care să facă abstracție de acestea. Putem calcula această mini amplitudine împărțind șirul de date în patru părți egale, astfel încât fiecare parte să conțină un sfert din valorile distribuției. **Quartilele** sunt percentilele care împart distribuția în patru porțiuni egale (a se vedea capitolul 3). Acestea sunt: Q_1 (percentila 25, îi corespunde frecvența cumulată procentual 25), Q_2 (percentila 50) și Q_3 (percentila 75).

Să ne imaginăm următorul șir de numere: 1 1 1 2 2 3 3 3 4 4 4 4 4 5 5 5 5 6 6 10. Mai departe vom împărți acest șir în patru părți egale.



*Calculând diferența dintre aceste valori vom obține ceea ce am numit **mini amplitudine**.*

Diferența dintre valoarea corespunzătoare quartilei 3 și valoarea quartilei 1 se numește **abatere interquartilă (**lb. engleză, **interquartile range)**. Aceasta este o nouă modalitate de estima împrăștierea unui set de date și de a compara diferite distribuții.

Quartilele pot fi obținute prin bifarea opțiunii **Quartiles** din meniul **Statistics** al procedurii **Frequencies** (vezi capitolul 3).

Variabilitatea unei distribuții înseamnă mai mult decât împrăștierea setului de scoruri, referindu-se și la cât de asemănătoare/diferite sunt scorurile. O modalitatea de a studia acest aspect este aceea de a observa cât de mult se îndepărtează valorile față de medie. Dacă vom calcula o distanță medie a scorurilor față de medie am putea obține o imagine a variabilității. Cu cât rezultatul ar fi mai mic cu atât mai apropiate ar fi valorile față de medie, iar scorurile ar fi mai puțin diferite.

4.2.3 Abaterea medie

Diferența dintre un scor din distribuție și medie se numește **abatere de la medie**. Pentru a calcula abaterea medie trebuie să realizăm suma abaterilor individuale ale fiecărei valori de la medie și să împărțim rezultatul obținut la numărul de scoruri. Însă, conform proprietăților mediei, suma tuturor abaterilor individuale față de medie este egală cu 0. În concluzie, și abaterea medie va fi egală cu 0.

$$\sum(X_i - m) = 0 \rightarrow \frac{\sum(Xi - m)}{N} = 0, \text{ unde:}$$

(formula 4.4)

- X_i este reprezentatul fiecărei valori din distribuție.
- m reprezintă media distribuției
- N este numărul de valori.

În tabelul de mai jos, pe coloana **X**, sunt prezentate sumele de bani cheltuite de bărbați la cumpărături în Statistics Mall. Pe coloana $\mathbf{X_i - m}$ sunt notate abaterile individuale ale valorilor de la medie, care însumate sunt egale cu 0. Acest rezultat se obține pentru orice distribuție.

X	$\mathbf{X_i - m}$
55	-13
60	-8
65	-3
55	-13
57	-11
70	2
75	7
85	17
80	12
78	10
$\sum X = 680$	$\sum(X_i - m) = 0$
$m = 68$	

4.2.4 Dispersia (varianța)

Pentru a elimina inconvenientul generat că suma abaterilor individuale ale valorilor de la medie este egală cu 0, trebuie să găsim o modalitate prin care toate acestea să devină pozitive. O soluție este aceea de a ridica abaterile de la medie la pătrat. Această metodă de măsurare a variabilității unei distribuții se numește **dispersie (varianță**; lb. engleză, **variance)**. Aceasta este notată cu simbolul σ^2 (la nivelul populației) și cu s^2 (la nivel de eșantion). Dispersia se calculează după formula:

$$s^2 = \frac{\sum(X_i - m)^2}{N - 1}$$

(formula 4.5)

X	$X_i - m$	$(X_i - m)^2$
55	-13	169
60	-8	64
65	-3	9
55	-13	169
57	-11	121
70	2	4
75	7	49
85	17	289
80	12	144
78	10	100
$\sum X = 680$	$\sum(X_i - m) = 0$	$\sum(X_i - m)^2 = 1118$
$m = 68$		

$s^2 = \frac{\sum(X_i - m)^2}{N-1} \rightarrow s^2 = \frac{1118}{9} \rightarrow s^2 = 124{,}22.$

4.2.5 Abaterea standard

Dispersia distribuției sumelor cheltuite de bărbați la cumpărături este egală cu 124,22. Din cauza ridicării la pătrat putem obține un rezultat ce poate fi mai mare decât amplitudinea, ceea ce face dificilă interpretarea dispersiei. Noi ne dorim un indicator care să ne prezinte împrăștierea și variabilitatea scorurilor în termeni de distanță față de medie, fără ridicare la pătrat. O soluție pentru această dificultate o reprezintă **abaterea standard** (lb. engleză, **standard deviation**). Aceasta este notată cu simbolul **σ** (la nivelul populației) și cu **s** (la nivelul eșantionului). Din această notație putem intui că **abaterea standard** *se obține prin extragerea radicalului* din **dispersie**.

Cu cât abaterea standard este mai mică, cu atât valorile sunt mai apropiate de medie, în timp ce o valoare mare a abaterii standard se traduce prin scoruri mai îndepărtate față de medie.

$$s = \sqrt{\frac{\sum(X_i - m)^2}{N - 1}}$$

(formula 4.6)

Pentru distribuția de mai sus, abaterea standard va fi egală cu:

$$s = \sqrt{124{,}22} \rightarrow s = 11{,}14.$$

Atunci când dorim să calculăm abaterea standard a unei distribuții de valori trebuie să parcurgem următorii pași:

1. Calculăm media aritmetică a valorilor din distribuție.
2. Calculăm abaterea de la medie a fiecărei valori (X-m).
3. Ridicăm la pătrat fiecare abatere de la medie.
4. Adunăm rezultatele obținute în Pasul 3.
5. Împărțim suma pătratelor obținută în Pasul 4 la numărul total de valori minus 1 (N-1). Rezultatul obținut în această etapă reprezintă dispersia (varianța).
6. Extragem radical din rezultatul obținut la Pasul 5 și aflăm abaterea standard.

Cu cât abaterea standard este mai mică, cu atât valorile sunt mai apropiate de medie, în timp ce o valoare mare a abaterii standard se traduce prin scoruri mai îndepărtate față de medie.

Exercițiu

Distribuția prezentată în tabelul de mai jos conține sumele de bani cheltuite de femei în Statistics Mall. Calculați media, dispersia și abaterea standard pentru aceste valori.

X	**X_i - m**	**$(X_i - m)^2$**
55		
56		
60		
75		
70		
65		
70		
72		
67		
60		
∑X =		
N =		
m =		

Abaterea standard este cel mai utilizat indicator al împrăștierii unei distribuții. Din acest motiv este important să cunoaștem cele mai importante proprietăți ale abaterii standard (Bluman, 2007):

a. Abaterea standard este ≥ 0; poate fi egală cu 0 doar atunci când valorile din distribuție sunt identice.
b. Dacă se scade/adună o constantă la fiecare valoarea a distribuției, abaterea standard rămâne neschimbată. Această proprietate este explicată prin faptul că adăugarea/scăderea unei constante la fiecare scor crește/scade media cu acea constantă. În concluzie, abaterea față de medie rămâne neschimbată.
c. Dacă fiecare valoare este multiplicată/împărțită cu o constantă, abaterea standard se multiplică/se divide cu acea valoare.
d. Abaterea standard față de medie este mai mică decât abaterea standard față de orice altă valoare din distribuție.

4.2.6 Coeficientul de variație

Abaterea standard este estimată în unitatea de măsură a variabilei. De asemenea, ea depinde de ordinul de mărime al scorurilor din distribuție. De exemplu, pentru suma de bani cheltuită de bărbați în Statistics Mall am obținut o abatere standard de **11,14** lei. Dacă am administra acestui eșantion și un chestionar de evaluare a satisfacției față de produsele existente în magazine, scorurile obținute au altă unitate de măsură și vor depinde de numărul de întrebări din chestionar. Să presupunem că abaterea standard pentru satisfacția clienților este **15,06**.

Cele două valori ale abaterii standard pot fi comparate? Putem să spunem că distribuția scorurilor la scala de satisfacție dispune de mai multă variabilitate decât distribuția sumelor cheltuite? Având în vedere că cele două distribuții conțin valori care nu pot fi comparate din cauza unității de măsură diferite, se utilizează **coeficientul de variație (cv)** pentru a face o analiză comparativă a celor două abateri standard. Principalul avantaj al coeficientului de variație este acela că nu este dependent de unitatea de măsură. Astfel, coeficientul de covariație este util pentru a compara distribuții cu unități de măsură diferite. Acesta se exprimă ca procent al raportului dintre abaterea standard și media distribuției.

$$cv = \frac{s}{m} * 100$$

(formula 4.7)

În cazul sumelor cheltuite de bărbați avem m = 68 și s = 11,14. Coeficientul de variație pentru această distribuție este calculat astfel:

$$cv = \frac{s}{m} * 100 \rightarrow cv = \frac{11{,}14}{68} * 100 \rightarrow cv = 16{,}38\%$$

Coeficientul de variație are sens doar atunci când valorile sunt măsurate pe scală de interval/raport. De asemenea, el poate fi utilizat doar pentru distribuțiile care conțin valori pozitive. Media tinde să scadă într-o distribuție cu valori negative, ceea ce afectează și valoarea coeficientului de variație. Cu cât valoarea coeficientul de variație este mai mică, cu atât media eșantionului este mai reprezentativă (Abdi, 2010).

Exercițiu

Calculați coeficientul de covariație pe baza mediei și abaterii standard obținute pentru distribuția sumelor cheltuite de femei în magazinele din Statistics Mall.

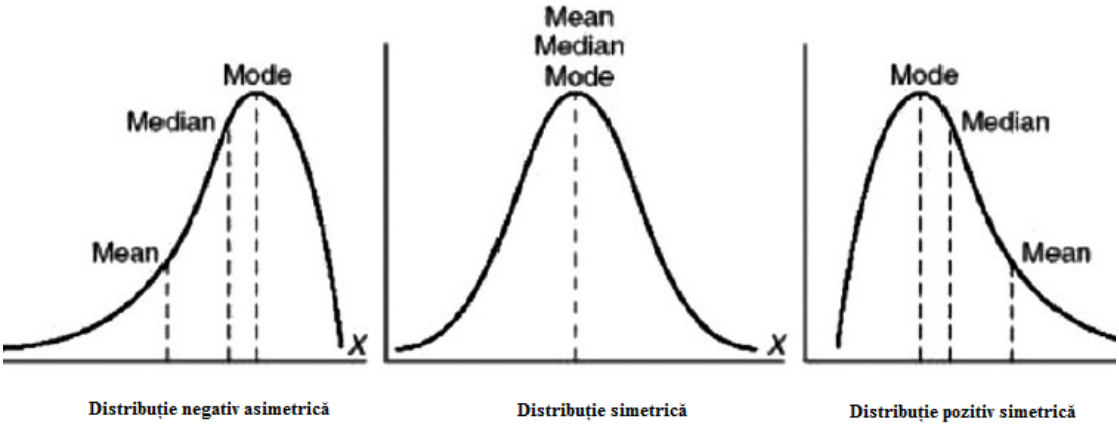
4.3 Indicatori ai formei distribuției

Indicatorii statistici descriși anterior ne oferă informații referitoare la tendința scorurilor de a se asemăna (indicatorii tendinței centrale) sau de a se apropia/îndepărta față de medie (indicatorii împrăștierii). Însă aceștia nu acoperă întreaga descriere a unei distribuții, fiind necesare informații și despre forma pe care o poate lua distribuția. Atunci când ne referim la forma unei distribuții avem în vedere **simetria** și **aplatizarea**.

4.3.1 Simetria (skewness)

O distribuție este **simetrică** atunci când valorile ei se împart în mod egal de o parte și de a alta a mediei. Distribuțiile în care cele mai multe valori se află de o parte sau alta a mediei sunt **asimetrice** (lb. engleză, **skewed**). Atunci când cele mai multe valori se află în **partea stângă a mediei** (în zona scorurilor mici) **distribuția este pozitiv asimetrică**. Dacă cele mai multe scoruri sunt în **partea dreaptă a mediei** (în zona scorurilor mari) distribuția este **negativ asimetrică**. Indicatorul numeric al simetriei se numește **skewness.** În cazul unei distribuții

perfect simetrice, skewness este egal cu **0**. El ia valori pozitive în situațiile când distribuția este pozitiv asimetrică și valori negative atunci când distribuția este negativ asimetrică.

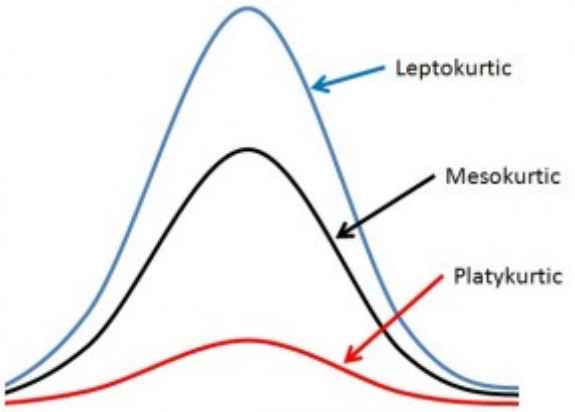


Distribuție negativ asimetrică **Distribuție simetrică** **Distribuție pozitiv simetrică**

În cazul distribuțiilor simetrice modul, mediana și media sunt identice. În situația distribuțiilor negativ asimetrice media are o valoare mai mică decât mediana și modul, iar în cazul distribuțiilor pozitiv asimetrice media are o valoare mai mare decât ceilalți doi indicatori ai tendinței centrale. În concluzie, atunci când distribuțiile sunt asimetrice cei trei indicatori au poziții diferite. După cum se poate observa, mediana este plasată întotdeauna între mod și medie. *Se consideră că o distribuție este simetrică atunci când valoarea indicelui skewness este cuprinsă între -1 și 1* (Popa, 2008).

4.3.2 Aplatizarea (kurtosis)

Indicele de aplatizare (**kurtosis**) ne oferă informații despre aplatizarea/boltirea distribuției. Indicatorul numeric al aplatizării este kurtosis. Atunci când acesta are o valoare pozitivă distribuția are o formă înaltă, iar valorile tind să se apropie de medie (**distribuție leptokurtică**). Valorile negative ale lui kurtosis se traduc într-o distribuție cu o formă aplatizată, în care valorile au tendința de a se îndrepta spre extreme (**distribuție platikurtică**) (deCarlo, 1997). O **distribuție mezokurtică** are o boltire și o îndepărtare a extremelor față de medie moderate, indicele kurtosis fiind egal cu 0. *Se consideră că o distribuție are o formă normală a aplatizării/boltirii atunci când kurtosis are o valoare cuprinsă între -1 și 1* (Popa, 2008).



4.4 Indicatorii statistici în R

Un cercetător este interesat să studieze predictorii comportamentului alimentar nesănătos. În acest sens, selectează un eșantion de 525 de participanți cu scopul de a testa efectul pe care îl au experiențele adverse din copilărie (ACE), alexitimia și stresul asupra comportamentului alimentar (ED). Dintre cei 525 de participanți, 161 sunt de gen masculin, iar 364 sunt de gen feminin. Cercetătorul a codificat genul masculin cu 1 și genul feminin cu 2. Datele colectate sunt introduse într-o bază de date pe care o vom numi **bdcap4**.

4.4.1 Media

Folosind datele din baza de date **bdcap4**, mai jos vor fi prezentate exemple cât mai variate de calculare a mediei aritmetice. Pentru a calcula media aritmetică în R vom folosi funcția **mean**.

Să presupunem că cercetătorul este interesat să calculeze media variabilei ACE. Expresia care calculează media în R este: mean(baza de date$variabila). În. cazul nostru linia de cod necesară pentru a calcula media variabilei ACE devine:

mean(bdcap4$ACE), unde:

- mean – este funcția din R care calculează media aritmetică
- bdcap4 – reprezintă numele bazei de date
- ACE – este numele variabilei analizate.

Rulând linia de cod menționată mai sus se poate observa că media variabilei ACE este 55.45905.

```
mean(bdcap4$ACE)
```

```
[1] 55.45905
```

Este posibil ca cercetătorul să dorească să calculeze media unei variabile doar pe un anumit subgrup din cadrul eșantionului. Mai exact, poate dorește să calculeze media experiențelor adverse din copilărie pentru participanții de gen feminin. Linia de cod utilizată în acest caz este:

mean(bdcap4$ACE [bdcap4$Genul==2]), unde:

- bdcap4$Genul==2 – indică faptul că va fi calculată media doar pentru participanții de gen feminin (codul 2 a fost atribuit participanților de gen feminin).

Astfel, se poate observa că media experiențelor adverse din copilărie pentru participanții de gen feminin este 56.2967. Pentru a calcula media experiențelor adverse din copilărie pentru participanții de gen masculin în linia de cod de mai sus se va scrie Genul==1.

```
mean(bdcap4$ACE [bdcap4$Genul==2])
```

```
[1] 56.2967
```

Mai sus am amintit faptul că media poate fi afectată de valori extreme. Astfel, în R există opțiunea *trim* care ne permite eliminarea unui grup de valori din extremitatea inferioară, respectiv cea superioară. Cele mai multe softuri statistice care au această opțiune elimină 5% din extremitatea inferioară a distribuției și 5% din valorile cele mai mari. Dacă dorim să eliminăm 5% din extremitățile distribuției, în R vom adăuga la funcția *mean* opțiunea *trim=0.05*. Dacă dorim eliminarea a câte 2,5% din extremitățile distribuției vom scrie *trim=0.025*.

mean(bdcap4$ACE, trim=0.05)

După ce rulăm linia de cod de mai sus vom putea observa că media experiențelor adverse din copilărie după ce am eliminat 10% din valori (5% din extremitatea superioară și 5% din extremitatea inferioară) devine 54.90275. Când am luat în calcul toate valorile din distribuție media a fost 55.45905, ceea ce ne arată că valorile din partea superioară au avut un impact mai puternic asupra mediei decât valorile din zona inferioară a distribuției.

4.4.2 Mediana

Mediana reprezintă valoarea din mijlocul unei distribuții. În R **median()** este funcția care permite calcularea acestei valori. Linia de cod cea mai simplă care permite calcularea medianei este prezentată mai jos. După ce îi solicităm softului statistică să ruleze linia de cod observăm că mediana este 54.

median(bdcap4$ACE), unde:

- median – este funcția din R care calculează mediana;
- bdcap4 – reprezintă numele bazei de date;
- ACE – este numele variabilei analizate.

La fel ca în cazul mediei aritmetice, putem calcula mediana doar pentru un anumit grup din cadrul eșantionului. Dacă cercetătorul este interesat să afle care este mediana experiențelor adverse din copilărie pentru participanții de gen masculin la linia de cod de mai sus va adăuga *[bdcap4$Genul==1]* și va observa că mediana este 53.

```
median(bdcap4$ACE)
```

```
[1] 54
```

```
median(bdcap4$ACE [bdcap4$Genul==1])
```

```
[1] 53
```

4.4.3 Modul

Modul este valoarea din distribuție cu cea mai mare frecvență de apariție. Spre deosebire de media aritmetică și de mediană, în R nu există o funcție standard care să calculeze modul motiv pentru care trebuie creată. Mai jos este prezentată funcția care permite calcularea modului pentru variabila ACE. Rulând sintaxa de mai jos vom observa că valoarea 53 are cea mai mare frecvență de apariție în cadrul distribuției ACE.

```
getmode <- function(ACE)
  {
  mod <- unique(ACE)
  mod[which.max(tabulate(match(bdcap4$ACE, mod)))]
  }

getmode(bdcap4$ACE)
```

```
[1] 53
```

Dacă dorim să calculăm modul pentru variabila alexitimie vom păstra aceeași sintaxă și vom înlocui ACE cu Alexitimie. Astfel, scorul 44 este modul în distribuția variabilei alexitimie.

```
getmode <- function(Alexitimie)
  {
  mod <- unique(Alexitimie)
  mod[which.max(tabulate(match(bdcap4$Alexitimie, mod)))]
  }

getmode(bdcap4$Alexitimie)
```

```
[1] 44
```

4.4.4 Amplitudinea

Amplitudinea reprezintă diferența dintre cel mai mare scor și cel mai mic scor din distribuție. În R amplitudinea poate fi calculată cu ajutorul funcției **range()**. În continuare vom exemplifica calcularea amplitudinii folosind variabila *Stres*. După ce solicităm calcularea amplitudinii programul afișează valoarea minimă (0) și valoarea maximă din distribuție (21). În concluzie, amplitudinea distribuției formată din scorurile variabilei *Stres* este 21.

range(bdcap4$Stres), unde:

- range – este funcția din R care calculează amplitudinea;
- bdcap4 – reprezintă numele bazei de date;
- Stres – este numele variabilei analizate.

```
range(bdcap4$Stres)
```

```
[1]  0 21
```

4.4.5 Abaterea interquartilă

Așa cum a fost menționat mai sus, amplitudinea poate fi afectată de valori extreme și ne poate lăsa impresia că scorurile din distribuție sunt diversificate, dar fără ca acest lucru să fie adevărat. Atunci când avem dovezi că distribuția este afectată de valori extreme se recomandă calcula abaterea interquartile (diferența dintre quartila 3 și quartila 1). Quartilele pot fi afișate în R folosind funcția **quantile()**. Programul afișează toate quartilele și se poate observa că quartila 1 este 6 în timp ce quartila 3 este 14. În concluzie, abaterea interquatilă pentru variabila *Stres* este 8.

quantile(bdcap4$Stres), unde:

- quantile – este funcția din R care calculează quartilele;
- bdcap4 – reprezintă numele bazei de date;
- Stres – este numele variabilei analizate.

```
quantile(bdcap4$Stres)
```

```
 0%  25%  50%  75% 100%
  0    6   10   14   21
```

4.4.6 Abaterea standard

Cel mai utilizat indicator statistic al împrăștierii este abaterea standard. Cu cât abaterea standard este mai mare, cu atât valorile din distribuție sunt mai diversificate. Pentru a calcula

abaterea standard vom folosi funcția **sd()**. În cazul în care dorim să calculăm abaterea standard a scorurilor obținute de participanți pentru variabila *ED* putem observa că aceasta este egală cu 5.131521 și o putem aproxima la 5.13.

sd(bdcap4$ED), unde:

- sd – este funcția din R care calculează abaterea standard;
- bdcap4 – reprezintă numele bazei de date;
- ED – este numele variabilei analizate.

```
sd(bdcap4$ED)
```

```
5.131521
```

4.4.7 Skewness

Skewness reprezintă indicele de simetrie, el aratând măsura în care valorile se distribuie similar de o parte și de alta a mediei. Pentru a calcula acest indicator în R vom folosi funcția **skew()**. În cazul în care dorim să calculăm indicele skewness pentru variabila *ACE* se poate observa că acesta este egal cu 1.419507 și poate fi aproximat la 1.42.

skew(bdcap4$ACE), unde:

- skew – este funcția din R care calculează skewness;
- bdcap4 – reprezintă numele bazei de date;
- ACE – este numele variabilei analizate.

```
skew(bdcap4$ACE)
```

```
[1] 1.419507
```

4.4.8 Kurtosis

Kurtosis oferă informații despre aplatizarea distribuției, iar funcția din R care ne ajută în acest sens este **kurtosi()**. De exemplu, pentru variabila *Alexitimie*, indicele kurtosis este egal cu 0.18171 și poate fi aproximat la 0.18.

kurtosi(bdcap4$Alexitimie), unde:

- kurtosi – este funcția din R care calculează kurtosis;
- bdcap4 – reprezintă numele bazei de date;
- Alexitimie – este numele variabilei analizate.

```
kurtosi(bdcap4$Alexitimie)
```

```
[1] 0.1817175
```

4.4.9 Funcția describe()

În cazul în care dorim să avem o privire de ansamblu asupra celor mai importanți indicatori sintetici ai distribuției putem folosi funcția **describe()**. Această funcție afișează cei concomitent cei mai importanți indicatori sintetici. Ea poate fi utilizată pentru o singură variabilă, pentru două sau mai multe variabile, dar și pentru toate variabilele din baza de date. Să presupunem că suntem interesați să aflăm indicatorii sintetici ai variabilei ACE. Folosind această funcție vom putea observa că în distribuție există 525 de scoruri, media este 55.46, abaterea standard este 8.02 ș.a.m.d.

describe(bdcap4$ACE), unde:

- describe – este funcția din R care calculează indicatorii sintetici;
- bdcap4 – reprezintă numele bazei de date;
- ACE – este numele variabilei analizate.

```
describe(bdcap4$ACE)
 vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
    1 525 55.46 8.02     54   54.65 4.45  25  96    71 1.42     4.95 0.35
```

- vars = 1 indică faptul că avem o singură variabilă analizată;
- n = 525 reprezintă numărul de cazuri analizate;
- mean = 55.46 reprezintă media aritmetică;
- sd = 8.02 indică abaterea standard;
- median = 54 indică mediana;
- trimmed = 54.65 arată media obținută după eliminarea 10% din scorurile cele mai mici și 10% din scorurile cele mai mari din distribuție;
- min = 25 indică cel mai mic scor din distribuție;
- max = 96 indică scorul maxim din distribuție;
- range = 71 reprezintă amplitudinea;
- skew = 1.42 afișează indicele skewness;
- kurtosis = 4.95 reprezintă indicele de aplatizare;
- se = 0.35 indică eroarea standard a mediei.

Funcția **describe()** poate fi utilizată și pentru a obține indicatorii statistici ai unor variabile specifice din baza de date. De exemplu, dacă dorim să obținem indicatorii statistici pentru variabilele *Stres* și *ED* putem folosi linia de cod de mai jos:

describe(bdcap4[c(“Stres”, “ED”)])

```
describe(bdcap4[c("Stres", "ED")])
      vars   n   mean     sd median trimmed    mad min max range  skew kurtosis   se
Stres    1 525  10.12   5.45     10   10.11   5.93   0  21    21  0.03    -0.73 0.24
ED*      2 525 157.28 101.06    164  157.38 130.47   1 330   329 -0.08    -1.23 4.41
```

Atunci când dorim să afișăm indicatorii sintetici pentru toate variabilele existente în baza de date vom folosi linia de cod:

describe(bdcap4)

```
describe(bdcap4)
          vars   n   mean     sd median trimmed    mad min max range  skew kurtosis   se
Genul        1 525   1.69   0.46      2    1.74   0.00   1   2     1 -0.84    -1.30 0.02
ACE          2 525  55.46   8.02     54   54.65   4.45  25  96    71  1.42     4.95 0.35
Alexitimie   3 525  45.63  12.62     44   45.04  13.34  20  94    74  0.48     0.18 0.55
Stres        4 525  10.12   5.45     10   10.11   5.93   0  21    21  0.03    -0.73 0.24
ED*          5 525 157.28 101.06    164  157.38 130.47   1 330   329 -0.08    -1.23 4.41
```

În cele din urmă, dacă dorim să afișăm indicatorii statistici în funcție de o variabilă categorială (de exemplu, pentru participanții de gen masculin, respectiv pentru participanții de gen feminin) avem la dispoziție funcția **describeBy()**. Rezultatele pentru participanții de gen masculin sunt prezentate în secțiunea *group: 1*, iar pentru participanții de gen feminin în zona *group 2*.

describeBy(bdcap4, bdcap4$Genul)

```
describeBy(bdcap4, bdcap4$Genul)
group: 1
          vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
Genul        1 161  1.00  0.00      1    1.00  0.00   1   1     0  NaN      NaN 0.00
ACE          2 161 53.57  7.09     53   53.19  4.45  29  85    56 0.97     4.73 0.56
Alexitimie   3 161 46.55 12.84     46   45.87 13.34  24  94    70 0.67     0.88 1.01
Stres        4 161  8.39  5.09      9    8.22  4.45   0  21    21 0.25    -0.55 0.40
ED*          5 161 56.07 37.57     56   55.40 47.44   1 123   122 0.06    -1.20 2.96
------------------------------------------------------------------------------------
group: 2
          vars   n   mean    sd median trimmed    mad min max range  skew kurtosis   se
Genul        1 364   2.00  0.00    2.0    2.00   0.00   2   2     0   NaN      NaN 0.00
ACE          2 364  56.30  8.27   54.0   55.35   4.45  25  96    71  1.52     4.76 0.43
Alexitimie   3 364  45.23 12.53   44.0   44.67  13.34  20  80    60  0.39    -0.22 0.66
Stres        4 364  10.89  5.44   11.0   10.97   5.93   0  21    21 -0.09    -0.71 0.29
ED*          5 364 125.96 79.63  128.5  126.20 103.04   1 261   260 -0.07    -1.27 4.17
```