

§ 3.6. ► Formule de calcul pentru medie și dispersie

Deși pentru obținerea unui anumit rezultat numeric putem folosi mai multe formule echivalente algebric, vom prefera întotdeauna acele formule care presupun cât mai puține operații – sunt cât mai rapide - sau care, în cazul operării cu valori aproximative, amplifică erorile de aproximare cât mai puțin – sunt cât mai precise. Pentru ultima proprietate va trebui ca operațiile care introduc aproximări - împărțirile și extragerile de radicali - să fie plasate cât mai târziu. Aceste formule se numesc, în literatura statistică, **formule de calcul** sau, mai explicit, *formule de calcul rapid și / sau precis*.

3.6.1. Metodă de calcul rapid al mediei și dispersiei

<p>(a) În cazul unei <i>serii statistice</i> $x_1, x_2, \dots, x_j, \dots, x_N$, respectiv,</p>	<p>(b) în cazul unei serii statistice grupată în <i>distribuția de frecvențe absolute</i> (x_j, N_j) cu $N = \sum N_j$</p>
<p>notând cu $T_1 = \sum x_i$ și $T_2 = \sum x_i^2$,</p>	<p>respectiv $T_1 = \sum N_j \cdot x_j$ și $T_2 = \sum N_j \cdot x_j^2$,</p>

media va fi:

$$M = \frac{T_1}{N}, \text{ iar}$$

dispersia se va putea obține, după calcularea mediei, prin formula:

$$S^2 = \frac{T_2}{N} - M^2.$$

Aceasta o vom denumi **formula de calcul rapid a dispersiei** deoarece presupune mai puține operații decât formula teoretică de definiție. Ținând cont că $T_2 = \sum x_i^2$ rezultă că:

dispersia este “**media pătratelor** minus **pătratul mediei** valorilor seriei”.

Formulara se reține ușor deoarece conține un joc de cuvinte.

Exemplul 3.6.1.

Să se calculeze media și dispersia pentru următoarea serie de 6 valori: 2, 3, 2, 4, 3, 3. Rezultatele finale vor fi rotunjite la două zecimale.

Rezolvări rapide

(AȘA DA !)

(a) Plasăm seria în coloana x , calculăm pătratele valorilor în coloana x^2 și calculăm T_1 și T_2 sumând valorile din fiecare coloană:

x	x^2	
2	4	Apoi, conform formulelor de calcul: $M = \frac{T_1}{N} = \frac{17}{6} \approx 2,833$ și $S^2 = \frac{T_2}{N} - M^2 \approx \frac{51}{6} - 2,833^2 \approx 8,5 - 8,026 = 0,474 \approx 0,47.$
3	9	
2	4	
4	16	
3	9	
3	9	
$T_1 = 17$	$T_2 = 51$	Răspuns: $M \approx 2,83$ și $S^2 \approx 0,47.$

Observații:

- ✓ Deoarece în calcul trebuie să avem o zecimală în plus față de rezultatul final, iar media este un rezultat intermediar în calculul dispersiei, am calculat media cu trei zecimale. În final, am rotunjit-o la două zecimale, așa cum se cere în enunț.
- ✓ În total s-au executat 20 operații algebrice.

(a) Dacă seria este grupată în distribuția de frecvențe absolute din primele două coloane ale tabelului următoare, calculăm mai întâi coloanele cu valorile x_j^2 și cu produsele $N_j \cdot x_j$, respectiv, $N_j \cdot x_j^2$ și calculăm T_1 și T_2 sumând valorile din ultimele două coloane:

N_j	x_j	x_j^2	$N_j \cdot x_j$	$N_j \cdot x_j^2$
2	2	4	4	8
3	3	9	9	27
1	4	16	4	16
$N = 6$			$T_1 = 17$	$T_2 = 51$

Apoi, conform formulelor de calcul:

$$M = \frac{T_1}{N} = \frac{17}{6} \approx 2,833 \text{ și}$$

$$S^2 = \frac{T_2}{N} - M^2 \approx \frac{51}{6} - 2,833^2 \approx 0,47.$$

+ Rezolvare greoaie

(AȘA NU !! - folosind formula teoretică a dispersiei)

(a') Din suma elementelor primei coloane de mai jos determinăm mai întâi media $M = 17 / 6 \approx 2,833$. Apoi calculăm coloanele doi și trei și suma celei de-a treia coloane:

x	$x-M$	$(x-M)^2$
2	-0,833	0,694
3	0,167	0,028
2	-0,833	0,694
4	1,166	1,340
3	0,167	0,028
3	0,167	0,028
$N = \Sigma x = 17$		$\Sigma(x-M)^2 = 2,812$

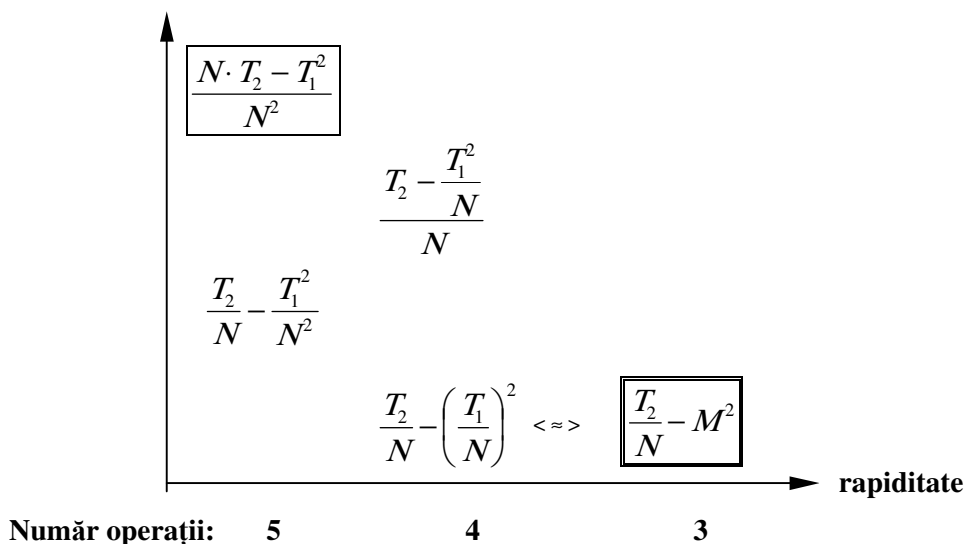
În final

$$S^2 = \frac{\sum (x-M)^2}{N} = \frac{2,812}{6} \approx 0,47.$$

Răspuns: $M \approx 2,83$ și $S^2 \approx 0,47$.

- ✓ În calculul dispersiei, media și celelalte rezultate intermediare au fost rotunjite la trei zecimale, pentru a avea, așa cum se cere, o zecimală în plus față de rezultatul final.
- ✓ În total s-au executat 23 operații algebrice, cu trei mai multe decât la punctul (a).
- ✓ În mod evident, operațiile executate au fost mai dificile decât în varianta (a) de mai sus.

+3.6.2. Formule rapide și precis pentru dispersie



Deoarece variabilitatea biologică este mult mai mare decât variabilitatea produsă de erorile de măsurare și de rotunjire prin calcul, în biostatistică ne va interesa mai mult rapiditatea decât precizia. De aceea, cel puțin în calculul manual, vom prefera formula din chenarul dublu care presupune calculul prealabil al mediei. În programele de calculator va fi de preferat formula din chenarul simplu. Aceasta este utilizată în științele exacte și inginerie, adică oriunde este nevoie de o precizie cât mai bună.

+3.6.3. Metode de calcul manual simplificat prin artificii

Acest subparagraf este introdus nu atât pentru înarmarea cititorului cu metode de calcul manual ci, mai degrabă, pentru a-l familiariza cu un mod de gândire fundamental în matematică și statistică, mod ce va fi necesar pentru înțelegerea paragrafului dedicat distribuției normale.

Pentru aceasta să considerăm următorul exemplu pe care îl vom soluționa, mai întâi direct, după modelul prezentat la punctul 3.6.1.(a).

Exemplul 3.6.3.

Să se calculeze media și dispersia pentru următoarea serie de 5 valori: 1, 7, 10, 16, 19.

Rezolvare:

Plasăm seria în coloana x , calculăm pătratele valorilor în coloana x^2 și calculăm T_1 și T_2 sumând valorile din fiecare coloană:

x	x^2
1	1
7	49
10	100
16	256
19	361
$T_1 = 53$	$T_2 = 767$

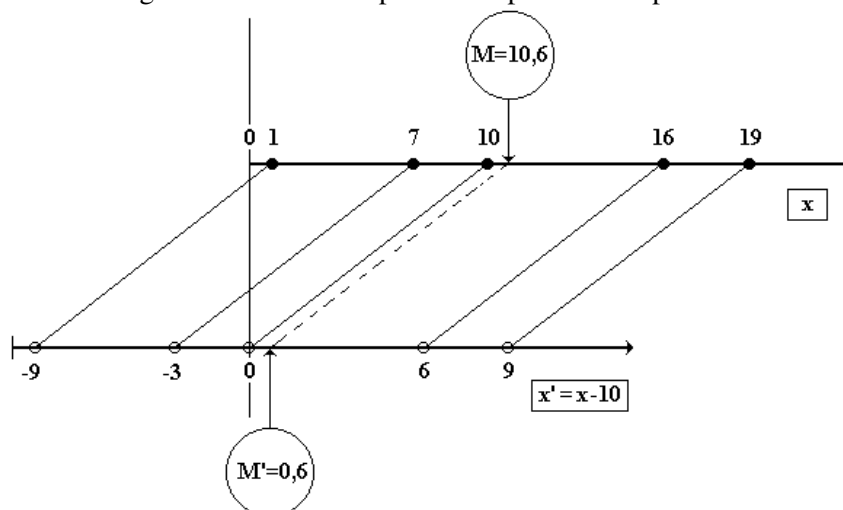
Apoi, conform formulelor de calcul:

$$M = \frac{T_1}{N} = \frac{53}{5} = 10,6 \text{ și}$$

$$S^2 = \frac{T_2}{N} - M^2 = \frac{767}{5} - 10,6^2 = 153,4 - 112,36 = 41,04.$$

1^o Translatarea datelor (sau introducerea unei medii provizorii, x_0)

Observăm că numerele cu care s-a lucrat în seria de 5 valori de mai sus sunt relativ mari și, deci, calculul a decurs relativ greoi. Datele sunt reprezentate prin cercuri pline în desenul următor.



Pentru micșorarea numerelor introduse în calcul putem face o *translație* convenabilă de lungime x_0 , adică vom calcula $x' = x - x_0$. De exemplu, dacă vom lua $x_0 = 10$, noul șir va avea valorile reprezentate mai sus prin ceruțele goale și enumerate în coloana a II-a a tabelului următor:

x	$x' = x - x_0$	$(x')^2$
1	-9	81
7	-3	9
10	0	0
16	6	36
19	9	81
	$T'_1 = 3$	$T'_2 = 207$

De regulă vom alege un x_0 cât mai aproape de tendința centrală a șirului, așa cum o sesizăm intuitiv. De aceea metoda se numește și *introducerea unei medii provizorii*, x_0 . Dacă este posibil, este convenabil ca x_0 să fie chiar o valoare a șirului pentru că aceasta va deveni 0 ușurând calculele ulterioare. Calculăm noua medie și noua dispersie:

$$M' = \frac{T'_1}{N} = \frac{3}{5} = 0,6$$

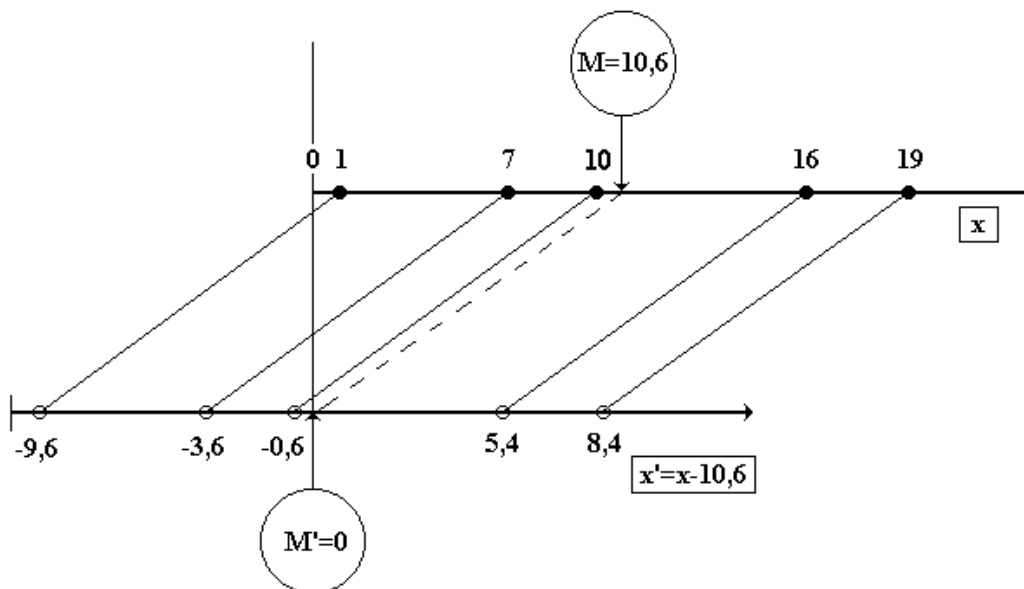
$$(S')^2 = \frac{T'_2}{N} - (M')^2 = \frac{207}{5} - 0,6^2 = 41,4 - 0,36 = 41,04.$$

Deoarece datele au fost translate la stânga cu $x_0 = 10$ unități, atunci și noua medie M' va fi cu $x_0 = 10$ unități mai mică decât media inițială, M . Adică $M' = M - x_0$.

Observația 1:

(importantă pentru paragraful dedicat distribuției normale)

- ✓ Dacă x_0 ar fi chiar media seriei inițiale M , noua medie va fi $M' = M - M = 0$, ca în desenul următor.



Observația 2:

- ✓ Media căutăată, M se poate obține din noua medie, M' astfel:

$$M = M' + x_0 = 0,6 + 10 = 10,6.$$

Împrăștierea, în particular dispersia, nu s-a modificat în nici un fel, deci

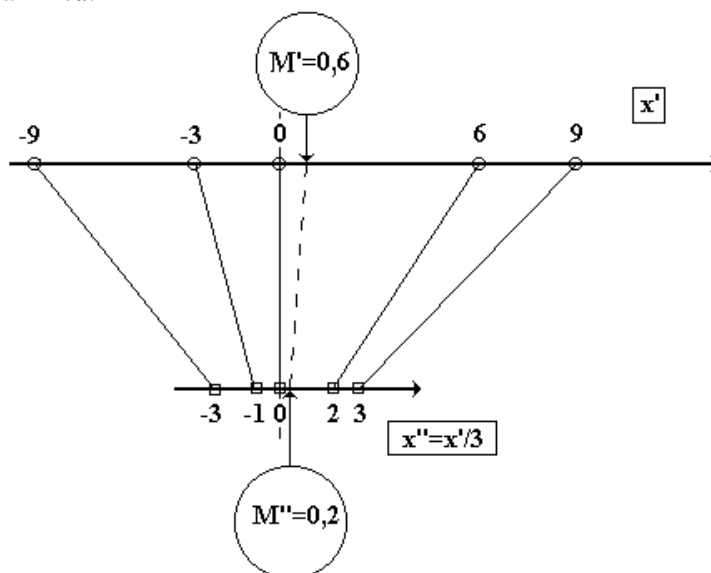
$$S^2 = (S')^2 = 41,04.$$

2° Dilatarea ori contractarea datelor (sau introducerea unui factor de scară h , sau schimbarea unității de măsură)

Să considerăm acum șirul valorilor x' din exemplul anterior. Observăm că aceste valori pot deveni și mai mici (deci și mai ușor de manevrat în calculul manual) dacă le divizăm cu $h = 3$, obținând valorile $x'' = x' / h$ din tabelul următor:

x'	$x'' = x' / h$	$(x'')^2$
-9	-3	9
-3	-1	1
0	0	0
6	2	4
9	3	9
	$T''_1 = 1$	$T''_2 = 23$

Am executat o contractare¹ a axei cu factorul de scară $h = 3$, noile valori fiind pătratele goale din desenul următor. Altfel spus *am schimbat unitatea de măsură* inițială cu o nouă unitate de măsură care este de 3 ori mai mică.



Media și dispersia pentru noul șir x'' vor fi:

$$M'' = \frac{T''_1}{N} = \frac{1}{5} = 0,2$$

$$(S'')^2 = \frac{T''_2}{N} - (M'')^2 = \frac{23}{5} - 0,2^2 = 4,6 - 0,04 = 4,56.$$

Deoarece datele au fost micșorate de $h = 3$ ori, și noua medie M'' , respectiv noua abatere standard S'' vor fi de același număr de ori mai mici decât M' , respectiv S' . Adică $M'' = M' / h$ și $S'' = S' / h$, iar $(S'')^2 = (S')^2 / h^2$.

Observația 3:

(importantă pentru paragraful dedicat distribuției normale)

- ✓ Dacă h ar fi chiar abaterea standard a șirului x' , și anume S' , noua abatere standard va fi $S'' = S' / S' = 1$.

¹ Dacă valorile șirului ar fi fost fracționare este posibil să fi fost avantajoasă o *dilatare* care să le transforme în numere întregi mai ușor de manevrat în calcul.

Observația 4:

✓ Media și dispersia șirului x' vor fi date de formulele:

$$M' = M'' \cdot h = 0,2 \cdot 3 = 0,6 \quad (S')^2 = (S'')^2 \cdot h^2 = 4,56 \cdot 3^2 = 4,56 \cdot 9 = 41,04.$$

Observația 5:

✓ Dacă dorim să aflăm media și dispersia șirului original x , atunci vom ține cont că am efectuat atât:

- o translație de lungime x_0 , cât și
- o comprimare / dilatare de factor de scară h :

$$x'' = x' / h = (x - x_0) / h$$

și formulele de revenire la media M și dispersia S^2 ale șirului inițial se vor obține combinând formulele de mai sus:

$$M = M' + x_0 = M'' \cdot h + x_0$$

$$S^2 = (S')^2 = (S'')^2 \cdot h^2$$

3.6.4. Probleme

- Să se calculeze prin formulele de calcul simultan rapid și precis mediile și dispersiile, precum și abaterea standard și coeficienții de variație pentru seriile $S4''$, $S3$, $S4'$ și $S5$ din subparagraful 3.1.1.
 - Să se compare rezultatele numerice obținute cu rezolvarea prin sinteză grafică prezentată în finalul problemei rezolvate la punctul 5^o de la 3.1.1.

Rezolvare:

a.

Seria	Media (în mm) M	Dispersia (în mm ²) S^2	Abaterea standard (în mm) S	Coefficientul de variație CV%
$S4''$	190,00	0,00	0,00	0,00%
$S3$	190,00	0,67	0,82	0,43%
$S4'$	189,50	3,43	1,85	0,98%
$S5$	172,39	65,90	8,12	4,71%

- Să se execute aceleași calcule și pentru distribuția grupată $S5'$ luând ca valori centrele claselor. Să se observe diferențele între valorile indicatorilor pentru această serie și cea negrupată corespunzătoare ($S5$).

La distribuția grupată față de cea negrupată corespunzătoare:

- media va fi întotdeauna mai mare ?
- dispersia va fi întotdeauna mai mică ?

Rezolvare:

a.

Seria	M	S^2	S	CV%
$S5'$	172,50	48,61	6,97	4,04%

b. Nu. Media la distribuția grupată poate fi mai mare sau mai mică în funcție de modul de plasare a valorilor seriei negrupate față de centrele claselor.

c. Da. Dispersia distribuției grupate va fi întotdeauna mai mică sau egală cu dispersia distribuției negrupate, căci prin grupare valorile împrăștiate în cadrul unei clase vor fi concentrate în centrul clasei respective.

Un răspuns mai elaborat se poate baza pe proprietatea de aditivitate a dispersiei. Seria negrupată are dispersia egală cu dispersia totală (variația totală / volumul seriei), iar seria grupată are dispersia egală cu dispersia intergrupări (variația intergrupări / același volum), grupările fiind dictate de intervalele de grupare. Practic, prin grupare s-a pierdut variația intragrupări, ceea ce intuitiv a constituit explicația de mai sus.