

8. Concepte statistice de bază

Statistica clasică este preponderent uni și bivalentă și se bazează pe teoria probabilităților.

Statistica modernă este (esențialmente) multivariată și se bazează pe geometrie, algebră și logică formală, dar și pe teoria probabilităților și se dezvoltă puternic datorită informaticii (aplicate).

În preocupările noastre se va aborda numai *statistica clasică*. Statistica clasică se bazează pe clasificarea prezentată în continuare. Unitățile statistice pot fi considerate fie *populație statistică*, fie *eșantion*. *Populația statistică* este alcătuită din obiecte, indivizi umani ori dintr-o altă specie, fenomene evenimente, idei, opinii, numere. *Populația statistică* poate fi *finită sau infinită, reală sau ipotetică*.

Cuvinte cheie :

- unitate statistică
- caracteristică variabilă (variabilă
- șir statistic/serie statistică, respectiv distribuție de frecvențe
- populație statistică
- eșantioane (independent prelevate, de observații perechi)

Statistica studiază mulțimi de observații efectuate asupra unor obiecte de aceeași natură, denumite *unități statistice* care prezintă (se încadrează în) anumite caracteristici (variabile). *Unitățile statistice* pot fi clasate, ordonate sau măsurate în raport cu caracteristicile respective. Mulțimile de observații se numesc șiruri sau serii (statistice)

Exemplul 1

Într-o crescătorie de păsări (unitățile statistice), acestea prezintă următoarele caracteristici:

- specia de păsări (poate fi constantă, dacă avem o singură specie, sau variabilă, în caz contrar), aceste date se *clasează*
- notă de frumusețe a exemplarelor, aceste date se *ordonează*
- lungimea / greutatea a pasarilor, se *măsoară*

8.1. Clasificarea variabilelor

Grosieră

- Variabile calitative = variabile ale căror variante pot fi doar clasate, nu ordonate sau măsurate.

Exemplu: variabila sex cu variantele masculin și feminin, variabila culoarea ochilor cu variantele negri, albaștri, verzi,

- Variabile cantitative = variabile ale căror valori pot fi *ordonate* sau chiar măsurate.

Exemple: greutatea, înălțimea, tensiunea arterială

Cele care pot fi ordonate se mai numesc și *semicantitative* (ordinale), iar valorile respective *ranguri*.

Clasificarea duală a mulțimilor (Anderberg)

Această metodă realizează clasificarea după *mulțimile de reprezentare* și după *scalele de reprezentare*.

a). Mulțimile de reprezentare pot fi:

- Discrete / discontinue
 - finite $\{a_1, a_2, \dots, a_n\}$
 - infinite $\{a_1, a_2, \dots, a_n, \dots\}$
- continue [numai finite]

b). Scalele de reprezentare sunt: nominală, ordinală, interval și raport.

Scalele se diferențiază prin proprietățile matematice pe care le exprimă.

Fie A și B două unități statistice, x_A și x_B fiind variantele, rangurile sau valorile unei variabile x pentru cele două obiecte.

- Scala **nominală**, realizează numai o distincție între A și B și anume fie $x_A = x_B$, fie $x_A \neq x_B$ (în acest caz x_A și x_B sunt denumite variante)

Exemplu: rasa, specia, tratamentul

- Scala **ordinală** este o scară nominală cu relație de ordine. În cazul $x_A \neq x_B$, fie $x_A > x_B$, fie $x_A < x_B$ (în acest caz x_A și x_B sunt denumite ranguri).

Exemple: scala durității mineralelor (Mohs), ierarhia militară.

- Scala **interval** sau scala de intervale egale, este o scală ordinală cu o măsură semnificativă a diferenței, a intervalului între două valori. În cazul $x_A > x_B$ spunem în plus că A este mai mare cu $x_A - x_B$ unități față de B.

Scala interval are originea (o) arbitrară și permite valori negative (în acest caz x_A și x_B sunt denumite valori).

Exemple: temperaturi și grade Celcius sau Fahrenheit, axa timpului (i.n.Christos, d.n.Christos)

- Scala **raport** / scala de proporții egale este o scală interval în care originea (o) este un zero absolut, altfel spus nu permite valori negative. În cazul $x_A > x_B$ putem spune și că A este mai mare de x_A/x_B ori față de B.

Exemple: temperaturi în grade Kelvin, greutatea, înălțimea.

Correspondența cu clasificarea grosieră este următoarea:

- variabilele calitative se pot reprezenta pe scala nominală
- variabilele semicantitative se pot reprezenta pe scalele nominală și ordinală
- variabilele cantitative se pot reprezenta pe scalele nominală, ordinală, interval, raport, după caz.

c). **Transformări permise în cadrul fiecărei scale:**

- Permutare și redenumirea

Exemplu: Sex M, F, sau F, M (permutare) sau 1, 2 (redenumire).

- Orice funcție $f(x)$ strict crescătoare

Exemplu: Liga x , cu $a > 1$; reținerea rangurilor în locul valorilor.

Variabile tip rang, pot proveni:

- din variante dispunând de relația de ordine
- din valori, ignorând proprietățile scalei interval

Variabile tip măsurătoare, pot proveni:

- măsurătoare propriu-zisă
- numărătoare.

8.2 Clasificări ale șirurilor statistice

A. Funcție de ordinea elementelor în șir

A₁ – ordinea elementelor nu contează

A₂ – șiruri, serii cu ordinea conformă unei succesiuni

- temporale :
 - serii temporale
 - serii cronologice
- spațiale

Ne vom ocupa numai de prima categorie de șiruri

B. Funcție de numărul de variabile luate simultan în considerație

B₁. șiruri statistice univariate

B₂. șiruri statistice bivariate

B₃. șiruri statistice multivariate

B₁. {crap, caras, somn, nisetru}; {7, 9, 6, 8}; {1,5 kg, 0,5 kg, 2 kg, 5 kg}

B₂. $\left\{ \begin{array}{cccc} \text{crap,} & \text{caras,} & \text{somn,} & \text{nisetru} \\ 7 & 9 & 6 & 8 \end{array} \right\}$

B₃. $\left\{ \begin{array}{cccc} \text{crap,} & \text{caras,} & \text{somn,} & \text{nisetru} \\ 7 & 9 & 6 & 8 \\ 1,5 \text{ kg} & 0,5 \text{ kg} & 2 \text{ kg} & 5 \text{ kg} \end{array} \right\}$

Statistica clasică este preponderent *uni* și *bivalentă* și se bazează pe *teoria probabilităților*. Statistica modernă este esențialmente *multivariată* și se bazează pe geometrie, algebră și logică formală.

8.3. Clasificarea multimilor de unități statistice și o structura a statisticii clasice

Funcție de orizontul analizat (studiat), mulțimea de unități statistice poate fi considerată fie:

- populație statistică
- eșantion

, dintr-o populație statistică

Populație statistică, alcătuită din obiective, indivizi (umani sau dintr-o altă specie), idei, evenimente, opinii, numere. Poate fi: *finită* sau *infinită* și *reală* sau *ipotetică*.

Populațiile statistice reale sunt în majoritatea cazurilor foarte mari. Deoarece este practic imposibil (total neeconomic) să fie studiate exhaustiv toate unitățile statistice ale unei populații statistice foarte mari se recurge la *eșantioane*.

Eșanation, mostră, probă, colectivitate de selecție, lot - o submulțime dintr-o populație statistică considerată cu scopul de a obține informații cu privire la populația respectivă. Populația statistică, din care s-a extras eșantionul se numește *populația mamă*, *populația țintă*.

Rezultatele obținute din analizele (studiile) bazate pe eșantioane cu *gradul de certitudine strict subunitar*. Extrapolarea rezultatelor obținute pe baza eșantioanelor la populația țintă se poate face:

- empiric (fără a putea marca gradul de certitudine)
- științific (exprimând exact gradul de certitudine).

Studiul incomplet al populațiilor statistice prin intermediul eșantioanelor probabilistice este scopul *statisticii inductive*.

Statistica clasică, se bazează pe trei componente:

- statistica descriptivă
- teoria probabilităților (parțial)
- statistica inductivă

8.4. Esantioanele prelevate independent si esantioane de observatii perechi

În marea majoritate a situațiilor reale se studiază *populațiile statistice* prin *eșantioane* provenite din acestea. Eșantioanele pot fi *produse de diverse fenomene naturale*, ori *pot fi selectate/generate* de cel care cercetează.

Astfel, apar *studiile de observație*, respectiv *studiile experimentale*. În toate aceste cazuri două sau mai multe eşantioane se pot produce, sau pot fi prelevate în două moduri: dependent / independent. Situația în care 2 eşantioane pot fi prelevate dependent este cea a observațiilor perechi.

Două eşantioane sunt eşantioane de observații perechi, dacă selectarea unei unități într-un eşantion impune selectarea unei anumite unități, perechi în celălalt eşantion. Cele două eşantioane de observații perechi au același volum.

În eșantioanele independente prelevate volumul eşantioanelor poate fi egal sau diferit ca mărime.

Exemplu: Cuplul de eşantioane utilizate în experimentele clasice de studiu al eficacității unei substanțe medicamentoase. Se ia un lot de subiecți cărora li se măsoară o caracteristică (tensiune arterială) înainte și după tratarea respectivei substanțe medicamentoase.

O greșeală metodologică gravă este amestecarea eşantioanelor de observații perechi, cu cele prelevate independent. Considerații asupra eşantioanelor de observație perechi - unitățile statistice dintr-un eşantion sunt *observate* sau măsurate:

- de două ori
- de doi operatori
- de două aparate
- de două momente de timp diferite
- după aplicarea unui tratament

Exemplu: „Studii longitudinale” antropologice care urmăresc probleme de creștere-dezvoltare prin 2 eşantioane (un eşantion cu copii la o anumită vârstă v , al doilea eşantion cu aceiași n copii la vârste $v + \Delta t$).

8.5. Statistica descriptiva univariată

Introducere în statistica descriptivă

Statistica descriptivă:

Ce face?

- sintetizează grafic și numeric informația culeasă [exhaustiv] dintr-o populație statistică
- descrie, dar NU explică esențialul ce rezultă din datele culese.

Cum face?

- prezintă grupat materialul în două maniere:
 - tabele statistice
 - reprezentări grafice

Paradigma centrală a statisticii (descriptive) este: „renunțarea la o parte din informație pentru câștig în relevanță”

8.5.1. Sinteza grafică univariantă

Se face prin evidențierea intuită și aproximativă a aspectelor esențiale de variabilitate dintr-o serie statistică. Se execută în doi pași:

- tabele statistice, simple sau cu simplă intrare
- reprezentări grafice adecvate timpului de variabile, astfel:
 - pentru variabile calitative și ranguri:
 - diagrame circulare;
 - diagrame prin coloane și prin benzi.
 - pentru ranguri și măsurători:
 - ✓ poligoane de frecvențe;
 - ✓ interograme.

Recomandări pentru variabile

- calitative – diagrame circulare
- tip rang – diagrame de frecvență
- tip măsurătoare – diagramele prin coloane sau prin benzi, poligoane de frecvență sau (mai ales) histogramele.

Sinteza grafică în tabele statistice se poate face prin:

- grupare, fără pierdere de informație - în tabele statistice simple cu frecvențele variabilelor ori valorilor, construind distribuțiile frecvențelor variabilelor/valorilor *denumite distribuții de frecvență negrupate*.
- gruparea, cu pierdere de informație - în tabele statistice simple cu frecvențele claselor sau intervalelor de grupare, construind distribuțiile frecvențelor claselor sau intervalelor de grupare denumite *distribuții de frecvențe grupate*.

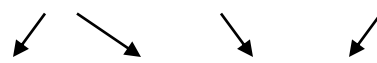
Pierderea de informație provine din comasarea unor variante *în clase* ori gruparea unor valori consecutive în clase, care în acest caz, se numesc și *intervale de grupare*.

8.5.1.1. Șir invariant, tabel statistic simplu distribuții de frecvențe și reprezentări grafice

A. Distribuții negrupate

a) Culoarea ochilor studenților = variabilă calitativă

$S_1 = \{a, v, a, a, n, n, n, c, c, n, a, c\}$



 albaștri verzi negri căprui

b) Notele obținute la biostatistică, de 12 studenți = Var. tip rang

$S_2 = \{6, 7, 8, 8, 7, 6, 9, 10, 7, 7, 8, 7\}$

c) 36 de studenți au măsurat cu precizie $\pm 0,5$ mm lungimea unei cărți \equiv var. tip măsurătoare obținând următoarele valori, ordonate ascendent.

$S_3 = \{188, 189 (8 \text{ ori}), 190 (18), 191 (8), 192\}$ măsurători repetate ale aceleiași mărimi = măsurători replicate

Distribuțiile de frecvență

Pentru S_1

Variabile distincte	Frecvențe absolute	Frecvențe relative	Frecvențe (relative) procentuale	Frecvențe procentuale cumulate
x_j	N_j	$F_j = N_j/N$	$P_j = 100 \cdot F_j \%$	$PC_j = P_1 + P_2 + \dots + P_j$
a	4	4/12	$100 \cdot 4/12 \approx 33\%$	
v	1	1/12	$100 \cdot 1/12 \approx 9\%$	
n	4	4/12	$100 \cdot 4/12 \approx 33\%$	
c	3	3/12	$100 \cdot 3/12 \approx 25\%$	

Totaluri $N = 12$

Pentru S_2

Perechile;

Valori distincte	Frecvențe absolute	$(x_j \cdot N_j)_j = 1 \cdot p =$ distribuții/repartiții de frecvențe absolute
x_j	N_j	$(x_j \cdot F_j)_j = 1 \dots p =$ distribuții/repartiții de frecvențe relative
6	2	$(x_j \cdot P_j)_j = 1 \dots p =$ distribuții/repartiții de frecvențe procentuale
7	5	
8	3	
9	1	$(x_j \cdot PC_j)_j = 1 \dots p =$ distribuții/repartiții de frecvențe absolute
10	1	

Totaluri $N = 12$

Pentru S_3

Valori distincte	Frecvențe absolute
------------------	--------------------

x_j	N_j
188	1
189	8
190	18
191	8
192	1
Totaluri	$N = 36$

B. Reprezentări grafice univariante

Definițiile care urmează sunt formulate pentru distribuțiile negrupate. În cazul distribuțiilor grupate termenii „variante” sau „valoare” trebuie înlocuite cu termenul „clasă”.

- **Diagrama circulară** - Cerc format din sectoare pentru fiecare variant/valoare, x_j astfel încât unghiul, respectiv aria fiecărui sector să fie proporțional(ă) cu frecvența respectivă.

Ex. seria S_1

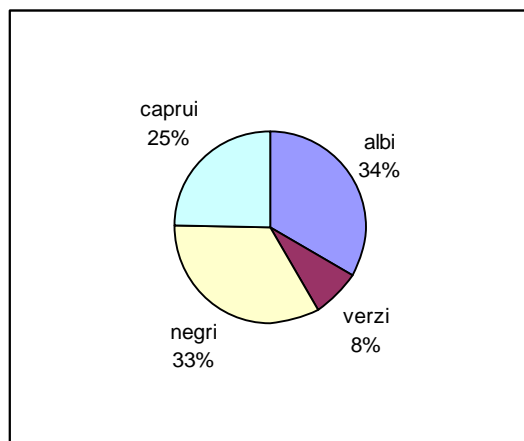


Figura 8.1 – Exemplu de diagramă circulară (pentru seria S_1)

- **Diagrama prin benzi sau bare** – reprezentare caracteristică plană în care pe axa verticală avem marcate variantele/valorile, în fiecare fiind construită o bandă orizontală de

lungime proporțională cu frecvența corespunzătoare; benzile sunt dreptunghiuri nelipite și de aceeași lungime, de regulă mult mai mică decât lungimile lor.

Ex. Seria S2

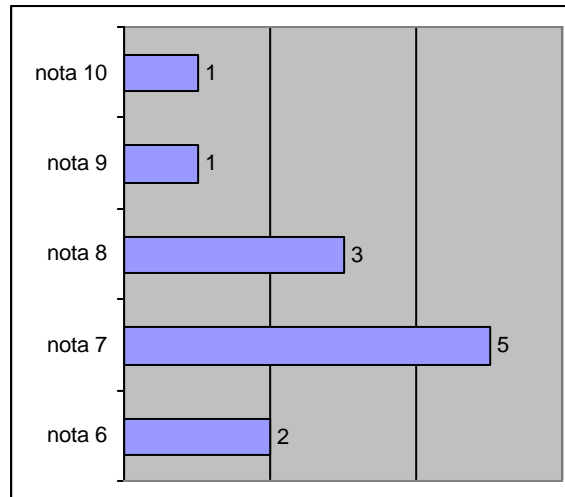


Figura 8.2 – Exemplu de ertica prin benzi sau bare (pentru seria S₂)

- **Diagrama prin coloane sau batoane** – *reprezentare carteziană plană, în care pe axa orizontală avem marcate variantele / variabile în fiecare fiind construită pe ertical o coloană de înălțime proporțională cu frecvențe corespunzătoare; coloanele sunt dreptunghiuri nealipite și de aceeași lățime, de regulă mult mai mică decât înălțimea lor.*

Ex. Seria 3

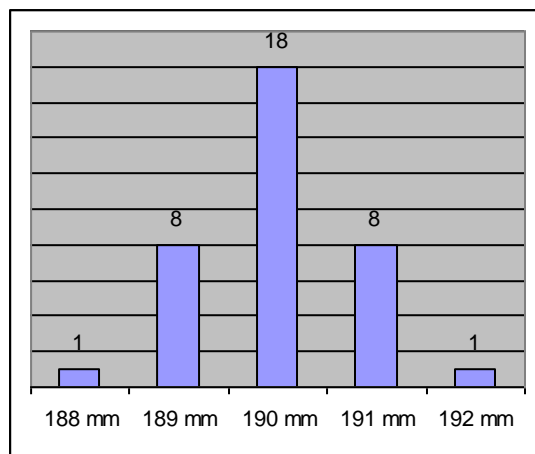


Figura 8.3 – Exemplu de diagramă prin coloane sau batoane (pentru seria S_3)

Valori aberante - valori care contrastează puternic cu marea majoritate a celorlalte valori ale șirului; 36 de studenți au măsurat lungimea palmei unuia dintre ei cu o precizie de $\pm 0,5\text{mm}$, obținând Ex. seria S_4

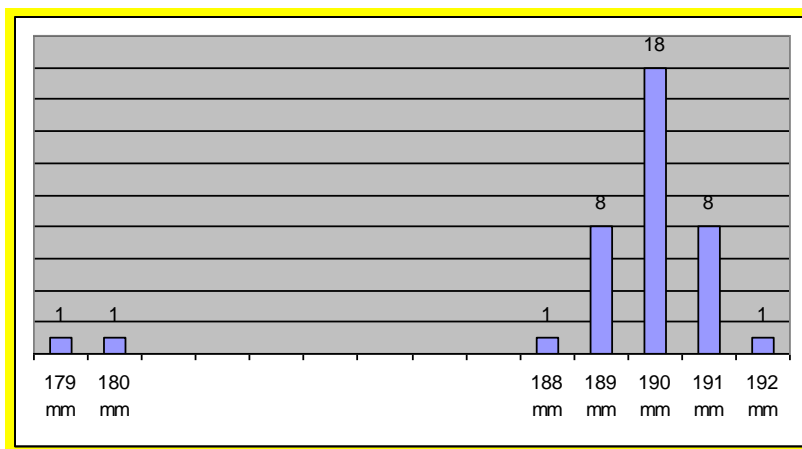


Figura 8.4 – Diagramă ce include valorile aberante

Valorile aberante se elimină. $S'_4 = S_4$, fără valorile aberante și rămâne diagrama din dreapta coform desenului de mai jos.

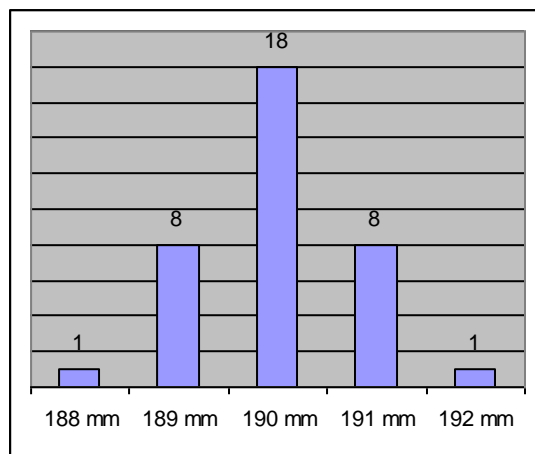


Figura 8.5 – Diagrama obținută după eliminarea valorilor aberante

C. Distribuții grupate pentru măsurători - histograma

Măsurându-se lungimea palmei drepte la 36 de studenți s-a obținut șirul S5, grupat fără pierdere de informație, ca distribuție de frecvențe este figurat în tabelul statistic următor, reprezentat apoi ca diagramă de batoane. Datorită distribuției „rare” de-a lungul intervalului 160 – 190 se recomandă o distribuție grupată, care se poate tabela și reprezenta după cum urmează:

Șirul 5

x_j	160	165	166	167	168	169	170	173	174	175	178	179	184	190
N_j	3	1	2	7	3	1	3	3	2	1	3	1	3	3

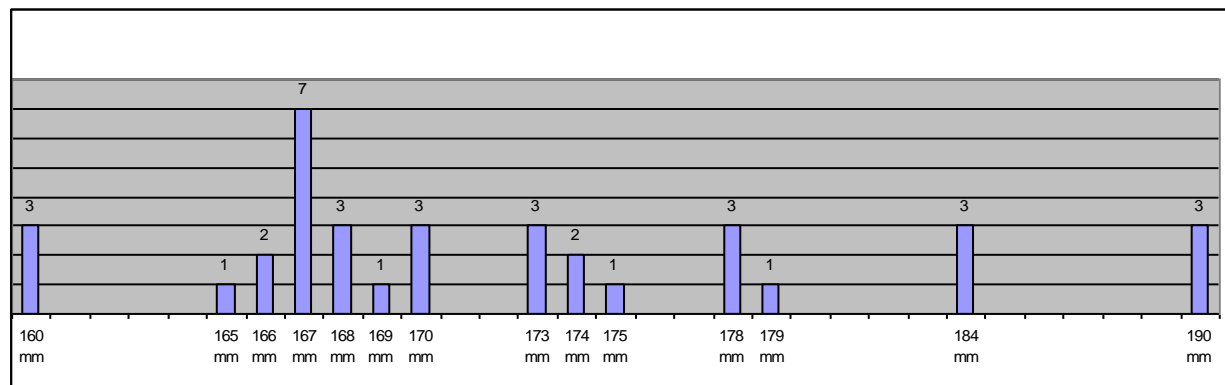


Figura 8.6 – Diagrama batoane fără reprezentare grupată

Datorită distribuției „rare” dealungul intervalului 160 – 190 se recomandă o distribuție grupată care se poate tabela și reprezenta după cum urmează:

Șirul 5'

Interval de clasa	[160,164] mm	[165, 170] mm	[171, 175] mm	[176, 180] mm	[181, 185] mm	[186, 190] mm
N _j	3	14	8	5	3	3

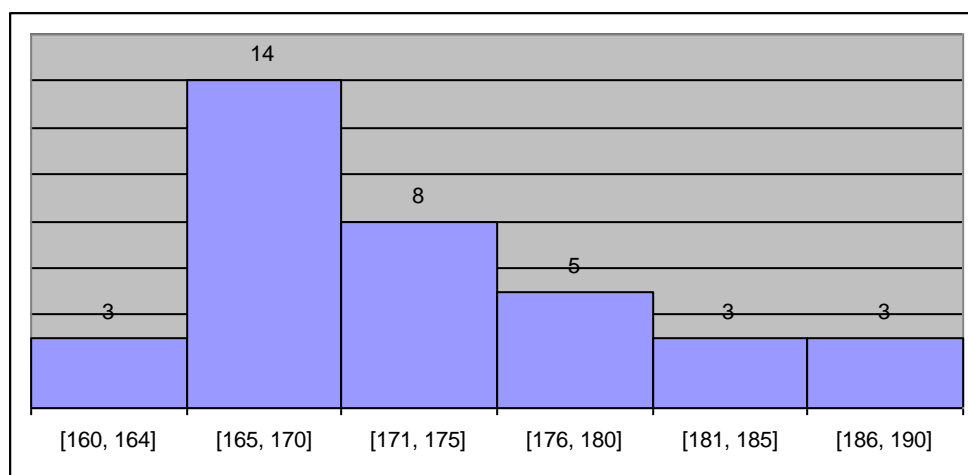


Figura 8.7 – Diagramă batoane cu reprezentare grupată – HISTOGRAMĂ

O astfel de reprezentare se numeste histograma, ea contine dreptunghiuri alipite, deoarece intervalele de grupare sunt intotdeauna alipite.

Histograma = reprezentare carteziana plana a unei distributii grupate, formata din dreptunghiuri alipite, cu bazele plasate pe intervalele de grupare si cu ariile proportionale cu frecventa claselor.

D. Distribuții grupate pe variante [variabile] calitative și ranguri

Cazul variantelor

În cazul șirului S₁ (culoarea ochilor), putem comasa *verde* și *albastru* în clasa culorilor deschise (cd) și culorile *căprui* și *negru* în clasa culorilor închise (ci).

Șirul S₁ (S₁ comasat)

Variante	Variante	Frecvențe relative	Frecvențe (rel.) procentuale
----------	----------	--------------------	------------------------------

distincte	absolute		
x_j	N_j	$F_j = N_j/N$	$P_j = 100 \cdot F_j \%$
(cd)	5	5/12	$100 \cdot 5/12 \cong 42\%$
(ci)	7	7/12	$100 \cdot 7/12 \cong 58\%$

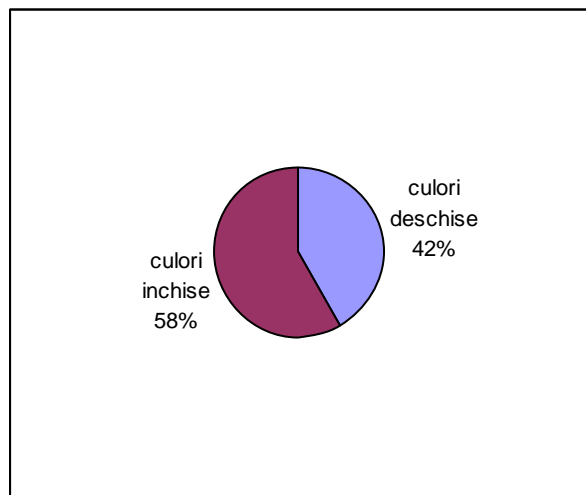


Figura 8.8 – Diagramă circular pentru șirul S_1 (pentru culoarea ochilor împărțită în închis și deschis)

Cazul rangurilor

Gruparea notelor, în cazul S_2 (notele studenților) - notele 5 și 6 formează clasa „Suficient”, 7 și 8 clasa „Bine”, 9 și 10 clasa „Foarte Bine”.

Clasa	Frecvențe absolute	Frecvențe relative	Frecvențe (relativ) procentuale
x_j	N_j	$F_j = N_j/N$	$F_j = 100 \cdot 2/12 \cong 17\%$
Suficient [5, 7]	2	2/12	$100 \cdot 2/12 \cong 17\%$
Bine [7, 9]	8	8/12	$100 \cdot 8/12 \cong 66\%$
Foarte bine [9, 10]	2	2/12	$100 \cdot 2/12 \cong 17\%$

În continuare prezentăm diagrama circulară, diagrama prin coloane și histograma (clasele au fost considerate intervale de grupare)

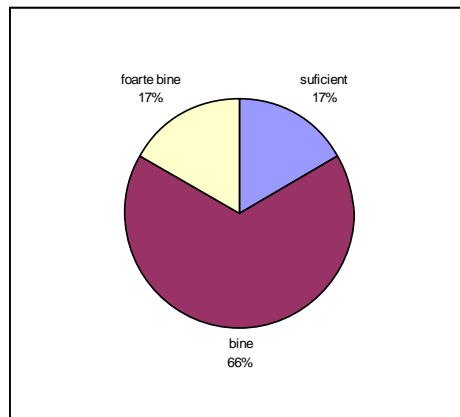


Figura 8.9 – Diagramă circulară după efectuarea grupării notelor pentru șirul S_2

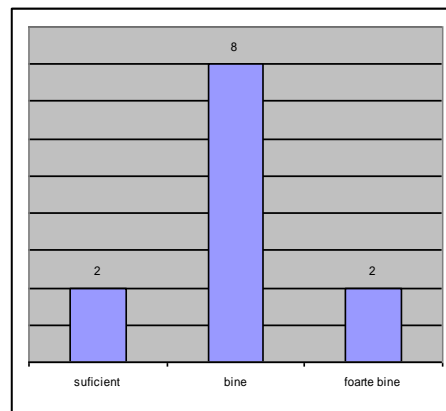


Figura 8.10 - Diagramă prin coloane după efectuarea grupării notelor pentru șirul S_2

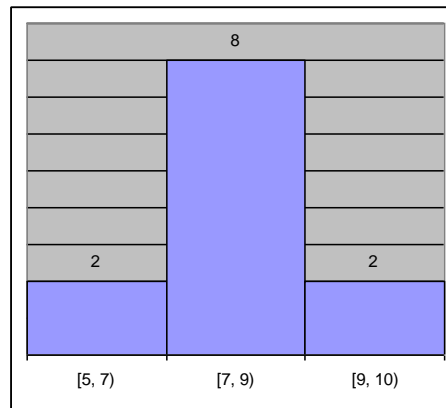
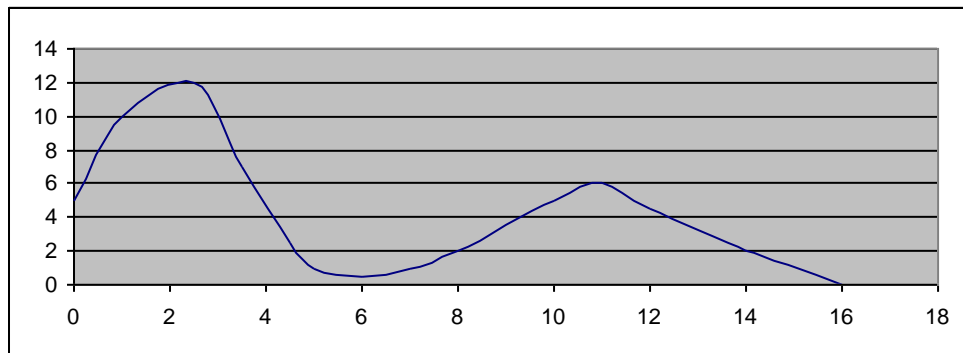
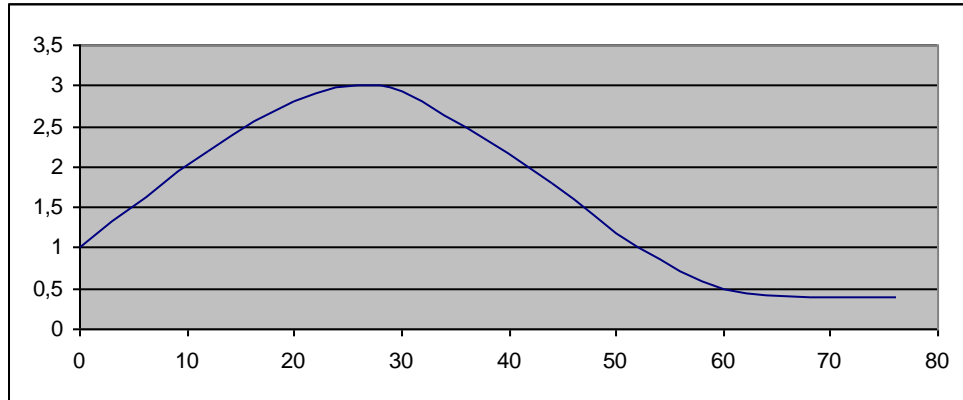


Figura 8.11 - Histograma după efectuarea grupării notelor pentru șirul S_2

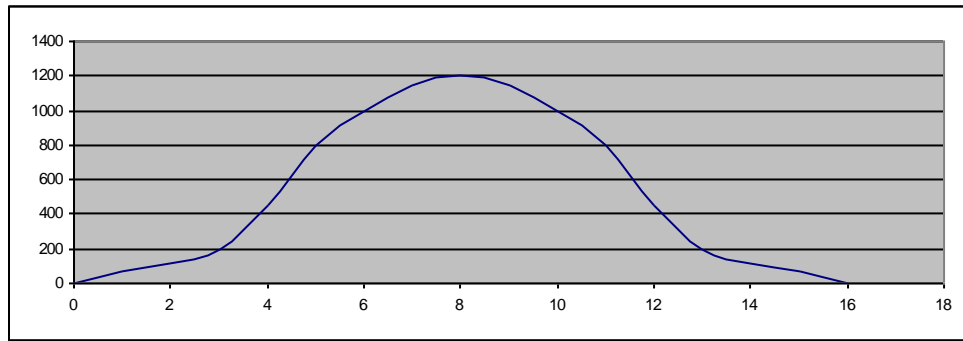
8.5.1.2 Limbajul repartițiilor (modul de grupare a măsurătorilor)

O distribuție se numește unimodală, când are o singură modă, respectiv bimodală atunci când are două mode. Rata fecundității specifică vârstei (*Microtus agrestis*)



O modă este un punct de maxim local. O distribuție bimodală, respectiv o distribuție multimodală pot fi considerate suma a două, respectiv mai multor distribuții unimodale.

O distribuție unimodală și simetrică se consideră a fi o distribuție cvasinormală, deoarece seamănă cu repartiția normală (Clopotul lui Gauss, curba erorilor).



Distribuția de frecvențe a înălțimii a 8500 de bărbați din Anglia (Distribuția unimodală și simetrică)

S-a lăsat intenționat la sfârșit forma de distribuție normală sau cvasinormală, pentru a atrage atenția că este o greșeală răspândită de a presupune această formă de distribuție în spatele oricărui fenomen de masă.

Pornind de la studiul formelor acestor distribuții empirice sau teoretice se poate construi tabelul prezentat în continuare.

Concluzii generale

1. De ce grupăm?

Grupăm (fără sau cu pierdere de informație) pentru a obține un câștig de relevanță.

2. Pentru ce grupăm?


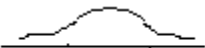

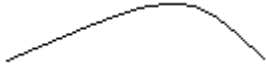
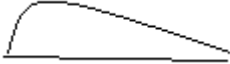
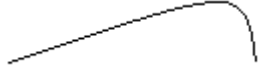

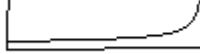
Grupăm ca să sesizăm (să ne încadrăm) în una din formele tip din tabelul prezentat.

Concluzii tehnice

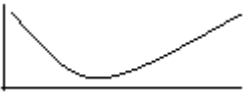
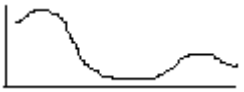
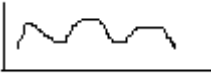
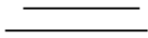
Modul de tratare a fiecărei forme depinde de:

- eterogenitățile vor fi tratate ca un amestec de două sau mai multe omogenități (adică distribuțiile bi sau multimodale, vor fi descompuse eventual prin decupare în două respectiv n distribuții unimodale.
- tendința centrală este cel mai bine exprimată de distribuțiile unimodale simetrice; vom încerca să sintetizăm prin transformări (de simetrie adecvate orice distribuție asimetrică.

Forme tip de distribuții

Unimodală simetrică (1 moda)	concentrată într-un punct (1)		Exprima <u>omogenitate</u> absoluta
	neconcentrată într-un punct (2)		
			Exprima cel mai bine o <u>tendinta centrala</u>
Unimodală asimetrică (1 moda)	slab asimetrica	de stanga (3)	
			
		de dreapta (4)	
			
	puternic asimetrica	de stanga (5)	
			
		de dreapta (6)	
			
	extrem asimetrica	de stanga (7) – in forma de <i>i</i>	
			
		de dreapta (8) – in forma de <i>j</i>	
			

Forme tip de distribuție (continuare)

Bimodala (2 mode)	simetrica (9) - de exemplu <u>în forma de u</u> 	Exprima <u>eterogenitate</u> , ca amestec de <u>2</u> <u>omogenitati diferite</u>
	asimetrica (10) 	
Multimodala (plurimodala)	multimodala propriu-zisa (11) ($n > 2$, mode) 	Exprima <u>eterogenitate</u> cu amestec de <u>n</u> <u>omogenitati diferite</u> ($n > 2$)
	uniforma (12), numai mode - omnimodala 	Exprima <u>eterogenitate</u> <u>absoluta</u>

OBSERVAȚII

1. – descompunerea, în particular decuparea în distribuții unimodale este obligatorie în cadrul statisticii descriptive (atunci când o serie este tratată drept populație statistică).
2. – transformarea pentru simetrizare nu este obligatorie în statistica descriptivă, fiind productivă în statistica inductivă.

8.5.1.3. Gruparea măsurătorilor

Nu poate exista o teorie matematică care să precizeze concret modul de grupare. Modalitățile de grupare pot fi alese de către fiecare specialist (medic, biolog, ecolog, biochimist) care cunoaște specificul material și obiectivele specifice.

Din experiențele anterioare, statistica pune la dispoziție doar reguli empirice de grupare, după cum urmează:

- grupăm doar serii cu volume ≥ 50
- Intervalele de grupare (intervalele de clasă/clasele de grupare) sunt: 20-40; 10-15; 8-20; 15-25; 8-15, ...
- se pot utiliza intervale de grupare egale sau inegale, după particularitățile datelor și interesul urmărit.

Gruparea cu intervale de clasă egale

În cazul intervalelor de grupare egale, există unele formule empirice de calcul al numărului de clase (n_c).

$n_c \approx 1 + 10/3 \cdot \lg N$, unde N = volumul seriei (formula lui Sturges)

Valoarea n_c se rotunjește la un număr întreg convenabil. ungimea intervalului de clasă (i_c) se poate calcula cu relația:

$$i_c = (x_{\max} - x_{\min})/n_c$$

, unde x_{\max} , x_{\min} sunt cea mai mare, respectiv cea mai mică valoare din serie. Valoarea i_c se rotunjește convenabil.

Exemplu

Se consideră următoarea distribuție negrupată de frecvențe, reprezentând adâncimi ale stațiilor pentru prelevare de probe din Delta Dunării, perioada (1978 – 1993). Se cere, gruparea cu intervale de clasă egale

Adancimea (cm) x_j	95	100	105	110	120	125	130	134	135	140	147	148	150	153	155
Frecvența N_j	1	4	1	3	4	4	4	1	2	4	1	1	7	1	3

x_j	157	160	163	167	170	175	180	185	188	190	198	200	208	210	211	220
N_j	1	7	1	1	2	2	3	1	1	4	1	3	1	4	1	2

x_j	240	257	290
N_j	3	1	1

Rezolvare:

Volumul $N = 81$ este mai mare ca 50, deci se poate grupa

Calculăm numărul de clase nc

$$nc = 1 + 10/3 \lg N = 1 + 10/3 \cdot \lg 81 \cong 1 + 10/3 \cdot 1,91 \cong 7,36$$

Rotunjim convenabil valoarea 7,36 și obținem 8, deci $nc = 8$

Lungimea intervalului de clasă:

$$ic = (x_{\max} - x_{\min}) / nc = (290 - 95)/8 = 24,375$$

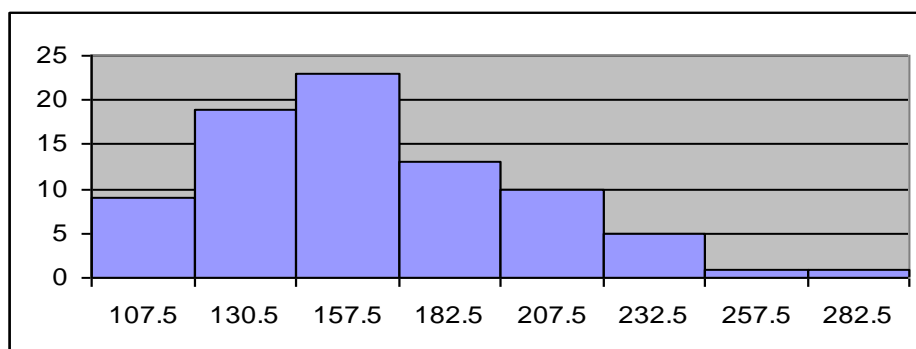
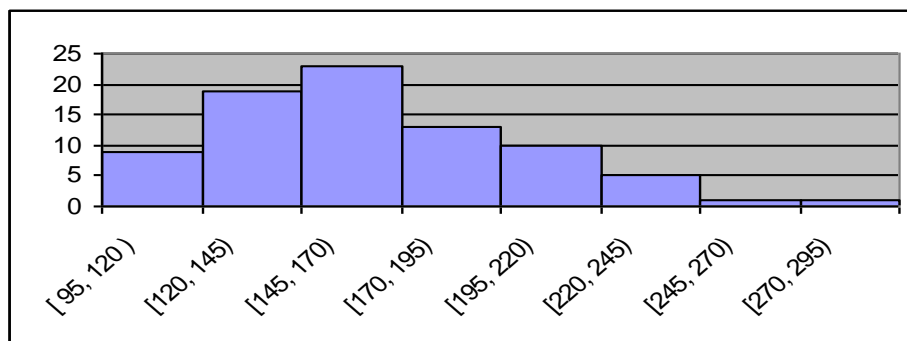
Rotunjim convenabil 24,375 și obținem $ic = 25$, deci $ic = 25$

Prima clasă începe cu valoarea minimă $x_{\min} = 95$

Se obțin astfel clasele distribuției de frecvențe propuse, cu intervale de grupare egale, conform tabelului de mai jos (coloana 1)

Intervalele de clasa (x_j, x_{j4})	Centrele intervalelor c_j	Frecvențele absolute N_j
[95,120)	107,5	9
[120,145)	132,5	19
[145,170)	157,5	23
[170,195)	182,5	13
[195,220)	207,5	10
[220,245)	232,5	5
[245,270)	257,5	1
[270,295)	282,5	1

Pentru construirea histogramei se vor utiliza coloana 1 și coloana 3 din tabelul de mai sus. Pentru constituirea poligonului frecvențelor pentru această distribuție grupată se calculează col. 2 din tabelul de mai sus (centrele intervalelor) și se utilizează coloanele 2 și 3.



Se observă că această distribuție empirică este o distribuție unimodală, asimetrică de stânga.

Concluzii:

În zona din Delta Dunării analizată, predomină adâncimi de cca 160 cm, urmează adâncimile mai mici lângă maluri, dar există și „gropi” de cca 2-3 m.

8.6. Sinteza numerică univariată

Se referă la aspecte de variabilitate și reprezintă un instrument complementar sintezei grafice, care oferă măsuri obiective și exacte (conform tabel din pag. 2/3) Cantitativ variabilitatea este concepută ca o împrăștiere, iar calitativ variabilitatea se poate denumi diversitate.

Modul de gândire cantitativ se aplică variabilelor cantitative, calitative binare sau binarizate și se realizează în indicatori (valori tipice) de:

- localizare, poziționare a tendinței centrale, poziționare a tendințelor extreme, de poziționare a tendințelor intermediare.
- împrăștiere (variabilitate, dispersie) de regulă în jurul tendinței centrale.

Pentru variabile cantitative continue sau compatibile cu variabilele continue se calculează și indicatori de:

- formă (pentru compararea cu o distribuție normală).

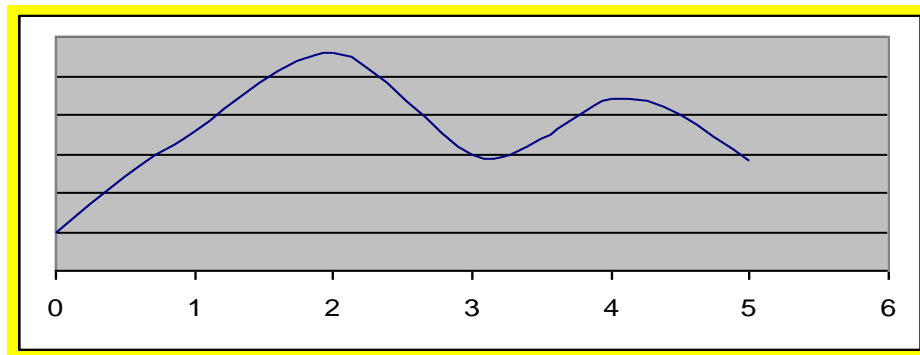
8.7. Tratarea unei variabile cantitative (indicatori de tendință centrală)

Condițiile lui Yule asupra indicatorilor de tendință centrală:

- a. să fie definit în mod obiectiv, independent de aprecierea subiectivă a cercetătorului;
- b. să fie expresia tuturor termenilor repartiției (serie)
- c. să posede proprietăți simple, evidente, făcând posibile înțelegerea sensului său general;
- d. să poate fi calculat cu ușurință și rapiditate;
- e. să se preteze ușor la calcule algebrice ulterioare;
- f. în cazul eșantioanelor, să nu fie afectat de fluctuațiile de selecție (în particular de valorile aberante)

Vom analiza următorii indicatori de tendință centrală: moda, mediana și media aritmetică.

Moda (modul, dominantă, valoare modală, valoare dominantă)



Definiții: În cazul unei curbe de frecvență (distribuția continuă a unei variabile continue) modă = punct de maxim local.

Valorile 2 și 4 sunt mode pentru distribuția continuă, deoarece sunt puncte de maxim local.

În cazul seriilor statistice pentru sesizarea modelor, datele trebuie să fie prezentate în distribuții de frecvențe (negrupate). În cazul utilizării intervalelor de grupare obținându-se distribuții de frecvențe grupate, în loc de mode se vorbește despre intervale modale. În continuare, se vor analiza numai distribuțiile negrupate. Modă = valoarea cu frecvența maximă locală în distribuție de frecvențe.

Pentru observarea modelor, în acest caz, este necesară gruparea datelor seriilor statistice în distribuții de frecvențe grupate sau nu.

Exemplu:

x_j	2	4	6	8	10
N_j	1	3	2	7	5

, unde 4 și 8 sunt mode deoarece 3 și 7 sunt frecvențe maxime locale.

Proprietăți:

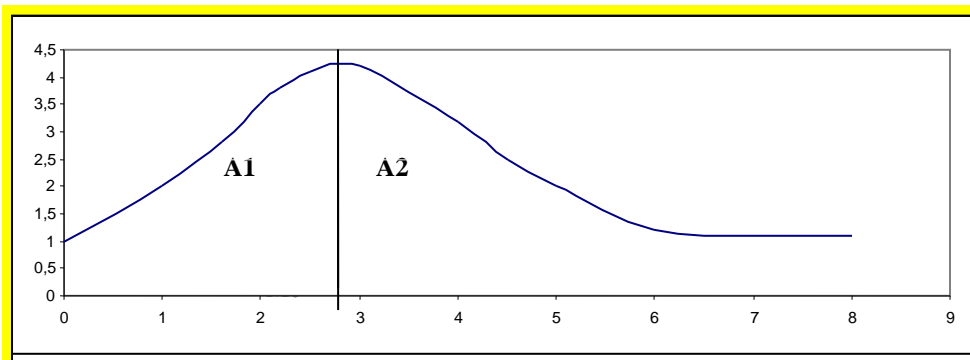
- Modelele induc clasificarea în distribuții unimodale, respectiv multimodale, clasificare esențială în gândirea statisticii clasice.
- Nu se pretează la calcule algebrice.

Mediana

Notatie: Me (pentru populația statistică)

\bar{x} pentru eșantioane

Definiție: În cazul unei curbe de frecvențe (distribuția continuă a unei variabile continue), mediana este valoarea care împarte *aria de sub curba de frecvențe în două arii egale* $A1 = A2$ (fiecare arie reprezentând 50% din întreaga arie de sub curbă).



În cazul seriilor statistice:

mediana = Valoarea care împarte seria statistică ordonată în două subserii de volume egale, volumele fiind măsurate în unități statistice și eventual jumătăți ale acestora.

- a) Dacă seria are număr impar de valori, 2_{k+1} , mediana este unic determinată de definiție și este valoarea x_{k+1} , din seria ordonată.
- b) Dacă seria are un număr par de valori, 2_k , definiția este satisfăcută de orice număr cuprins între x_k și x_{kM} , din seria ordonată.

Pentru unicitatea soluției, se ia prin convenție, drept mediană, semi-suma valorilor x_{kM} , din seria ordonată.

Exemple:

a) Fie seria ordonată 1, 3, 7, 8, 12 (5 termeni – nr. impar)

Me = 7

Considerăm că valoarea 7 se află în mijlocul seriei ordonate de volum impar.

Practic $rg(5/2) = 2,5$ (nr. fracționar care se rotunjește prin adaos la 3, de Me = termenul de rang 3, deci 7.

b) Fie seria ordonată cu 4 termeni, 1, 3, 6, 18

Conform definiției, orice rang între 3 și 6 (3, 7; 4, 5; 5, 2), Me este semisuma termenilor din mijlocul seriei ordonate $= (3+6) / 2 = 4,5$

Practic $rg(4/2) = 2$ (nr. întreg), deci Me = semisuma termenilor de rang 2 și 3 $= 4,5$

Proprietăți

- mediana este relativ ușor de observat și de calculat
- exprimă cel mai bine tendința centrală (în special distribuțiile asimetrice)
- mediana tratează valorile ca pe ranguri
- nu este sensibilă la valori extreme (în particular la valori aberante)
- se poate calcula și pentru serii pentru care nu se poate calcula exact media (valorile extreme nu sunt cunoscute)
- mediana este un element al șirului, când șirul are un număr impar de termeni.

Alte denumiri :

Toxicologie: LD50 = Lethal Dose 50 = Doza letală 50 = Doza care omoară 50% din indivizii care au fost intoxicați cu doza respectivă.

Farmacologie: ED 50 = Effect Dose 50 = Doza care are efect asupra 50% din indivizii tratați cu doza respectivă.

Biologia populațiilor: Media de viață

Mortalitatea populației în funcție de vârstă pe o curbă de frecvențe, are o mediană care reprezintă vârsta până la care au murit 50% din indivizii populației respective.

Media (aritmetică)

Termenul „medie” este folosit, în sens general de indicator de tendința centrală și în sens restrâns de medie aritmetică.

Notății: M – pentru populații statistice în general

μ – pentru populații statistice teoretice

x, m – pentru eșantioane.

Definiții:

- a) În cazul unei serii statistice formate din N valori distincte (sau nu) $x_1, x_2 \dots x_k, \dots x_N$, media M este suma valorilor seriei împărțită la volumul seriei.

$$\sum_{j=1}^N x_j \quad (\text{formula mediei simple})$$

$$M = \frac{\quad}{\quad}$$

$$N$$

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

$$\sum_{j=1}^p N_j \cdot x_j \quad (\text{formula mediei ponderate})$$

$$M = \frac{\quad}{\quad}$$

$$\sum_{j=1}^p N_j$$

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

Exemplu:

Fie seria de 6 valori:

1, 4, 2, 2, 1, 2

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

$M = 2$ este media simplă

x_j	1	2	4
N_j	2	3	1

$$N = 6$$

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

$M = 2$ este media ponderată a seriei de valori distincte

1, 2, 4 cu ponderile 2, 3, 1

$$\text{Media simplă a seriei } (1, 2, 4) \Rightarrow M' = (1+2+4)/3 = 2,33$$

Proprietăți:

- a. se pretează la calcule algebrice ulterioare
- b. media aritmetică ia în considerare toate valorile seriei cu întreaga lor informație
- c. oarecum dificil de calculat manual
- d. este sensibilă la valorile extreme (în particular la cele aberante).

Indicatorii de localizare a tendințelor extreme sau intermediare, valabili pentru orice distribuții

Ex. val. min și val. max dintr-un șir (localizarea extremelor).

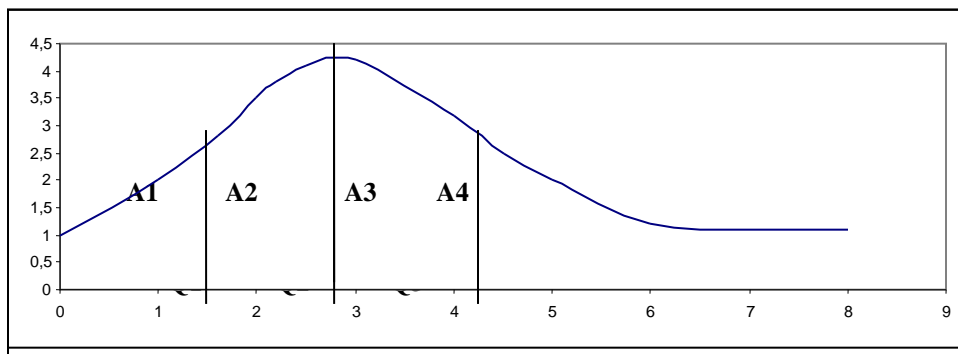
Generalizând modelul „geometric” al medianei vom introduce o gamă frecvent utilizată de indicatori de localizare (cuartilele, decilele, centilele)

Cuartile

Notăție: Q_1 , Q_2 , Q_3

Definiții

În cazul unei curbe de frecvențe (distribuția continuă a unei variabile continue), cuartilele sunt cele 3 puncte care împart *aria de sub curba de frecvențe* în 4 arii egale $A_1 = A_2 = A_3 = A_4$ (fiecare arie reprezentând 25% din întreaga arie de sub curbă).



Q_2 = mediana

În cazul seriilor statistice cuartilele sunt 3 valori care împart seria statistică, ordonată crescător, în 4 subserii de volume egale (volumele fiind măsurate în număr de unități statistice).

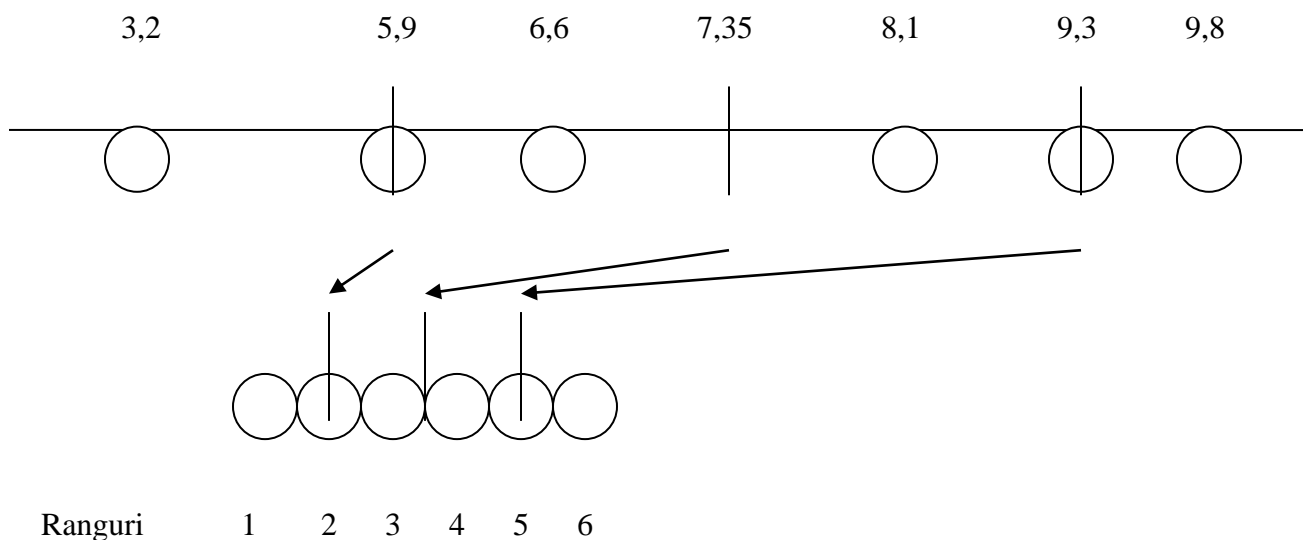
Q_1 = cuartila inferioară, lasă la stânga sa, în seria statistică ordonată crescător, 25% din termeni și eventual pătrimi ale acestora.

Q_2 = mediana

Q_3 = cuartilă superioară, și lasă la stânga sa, în seria statistică ordonată crescător, 75% din și eventual pătrimi ale acestora.

Exemplu:

Fie seria de 6 concentrații de oxigen măsurate în mg/l, în apă din Delta Dunării și ordonate crescător.



Considerăm numerele ordonate ca niște mărgelile înșirate pe o ață, la diverse distanțe. Strângem „mărgelile” unele lângă altele, definind distanțele. În acest fel, numerele devin ranguri:

Tăiem acest nou șirag în 4 părți egale de câte o „mărgea și jumătate”.

Quartila inferioară Q_1 va tăia mijlocul, mărgelii a 2-a, adică va fi 5,9

Mediana = Q_2 , va cădea între cea de-a 3-a și a 4-a „mărgea” (va fi semisuma acestora

$$Me = (6,6 + 8,1)/2 = 7,35$$

Quartila superioară Q_3 va tăia mijlocul „mărgelii” a 5-a, adică va fi 9,3

Practic cuartilele Q_1 , Q_2 , Q_3 se vor face astfel, conform convențiilor introduse, mai sus:

- ordonăm ascendent seria de volum N

- calculăm rangul cuartilei respective $rg(Q_1) = N(1/4)$
- dacă $rg(Q_1)$ este număr fracționar, îl restrângem prin adaos și Q_1 este semisuma dintre termenul cu rangul rg și următorul termen

3,1	5,9	6,6	8,1	9,3	9,8
x_1	x_2	x_3	x_4	x_5	x_6

(rang) • $rg(Q_1) = 6 \cdot (1/4) = 1 \frac{1}{2}$, rotunjit prin adaos = **2** $\rightarrow Q_1 = x_2$ (**5,9**)

• $rg(Q_2) = 6 \cdot (2/4) = 3$, $\rightarrow Q_2 = \mathbf{Me} = (x_3 + x_4) / 2$, (x_3, x_4 din serie ordonate crescător)

$$(6,6 + 8,1) / 2 = \mathbf{7,35}$$

• $rg(Q_3) = 6 \cdot (3/4) = 4 \frac{1}{2}$, rotunjit prin adaos = **5** $\rightarrow Q_3 = x_5$ (**9,3**)

Ex. : Seria este de volum 4 ordonată ascendent

1, 2, 8, 8

1,5 5 8

Q_1 Q_2 Q_3

Ex. : Seria de volum 5 8, 7, 3, 1, 2 ; ordonăm ascendent:

1 2 3 7 8

2 3 7

Q_1 Q_2 Q_3

Decile și centile

Analog, se întrunesc noțiunile de decile ($D_1, D_2, \dots D_9$) și de (per)centile ($C_1, C_2, \dots C_{99}$), respectiv de decilă inferioară (D_1), decila superioară (D_9), centila inferioară (C_1) și centila superioară (C_{99}).

Algoritmul de calcul al acestora se obține înlocuind în algoritmul de calcul al cuartilelor, expresia $N(1/4)$ cu $N(1/10)$, respectiv cu $N(1/100)$.

Metodă de calcul rapid al centilelor

Etapa 1

Se pornește de la distribuția de frecvențe relative procentuale (conform primele 2 coloane din tabelul următor). În col. 1 sunt trecute distinct și ordonat ascendent valorile seriei, în coloana 2 sunt înscrise frecvențele relative procentuale ale valorilor din prima coloană (în procente).

Etapa 2

Se calculează coloana 3, care cuprinde frecvențele relative procentuale cumulate (procentele cumulate) prin cumularea frecvențelor relative procentuale.

Exemplu: S-a măsurat greutatea (kg) pt . 103 băieți de cca 17 ani calculându-se procentele valorilor distincte și procentele cumulate. S-a obținut tabelul următor:

Etapa 3

Determinarea centilei dorită

Kg. Greut .	col .1	44	46	47	49	51	52	53	54	55	56	57	58	59	60	61	62
% distin ct	col .2	1,0	1,9	1,9	2,9	1,0	1,9	1,0	6,8	3,9	7,8	2,9	1,0	4,9	6,8	7,8	5,8
% cumu l	col .3	1,0	2,9	4,8	7,7	8,7	10,6	11,6	18,4	22,3	30,1	33,0	34,0	38,9	45,7	53,5	59,3

Kg. Greut.	col. 1	63	63,5	64	65	66	67	68	69	70	71	72	75	77	80
% distinct	col. 2	1,0	1,0	7,8	6,8	2,9	1,9	1,0	6,8	2,9	3,9	1,0	1,9	1,0	1,8
% cumul	col. 3	60,3	61,3	39,1	75,9	78,8	80,7	81,7	88,5	91,4	95,3	93,3	98,2	99,2	100

Se caută în coloana 3, cel mai apropiat procent mai mare sau egal cu indicele centilei respective.

Dacă procentul cumulat, astfel determinat, este mai mare strict decât indicele centilei, valoarea din coloana 1 de pe aceeași linie va fi centila căutată.

În caz de egalitate, centila va fi semisuma dintre valoarea din coloana 1 de pe aceeași linie și valoarea de pe linia următoare.

Pentru centila C_3 , găsim procentul cumulat 4,8 care este pe linia valorii 47. Deoarece $4,8 > 3$, rezultă că $C_3 = 47$

În mod analog, pentru centila C_{33} , găsim procentul cumulat 33, care este pe linia valorii 57.

Procentul cumulat este egal cu indicele centilei $C_{33} = (57+58)/2 = 57,5$

Indicatori de împrăștiere

Indicatorii de împrăștiere se raportează la indicatorii de localizare, existând asemenea indicatori, bazați pe :

- indicatori de tendință extremă (amplitudine)
- indicatori de tendință intermediară (intercuartila)
- indicatori de tendință centrală (dispersia, abaterea standard, coeficientul de variație)

Amplitudinea

Notății: A, ω

Definiție: Amplitudinea este diferența dintre valoarea maximă și valoarea minimă din serie:

$$A = x_{\max} - x_{\min}$$

Exemplu: să se calculeze amplitudinea seriei: 30, 30, 26, 32, 30

$$A = 32 - 26 = 6$$

Proprietăți:

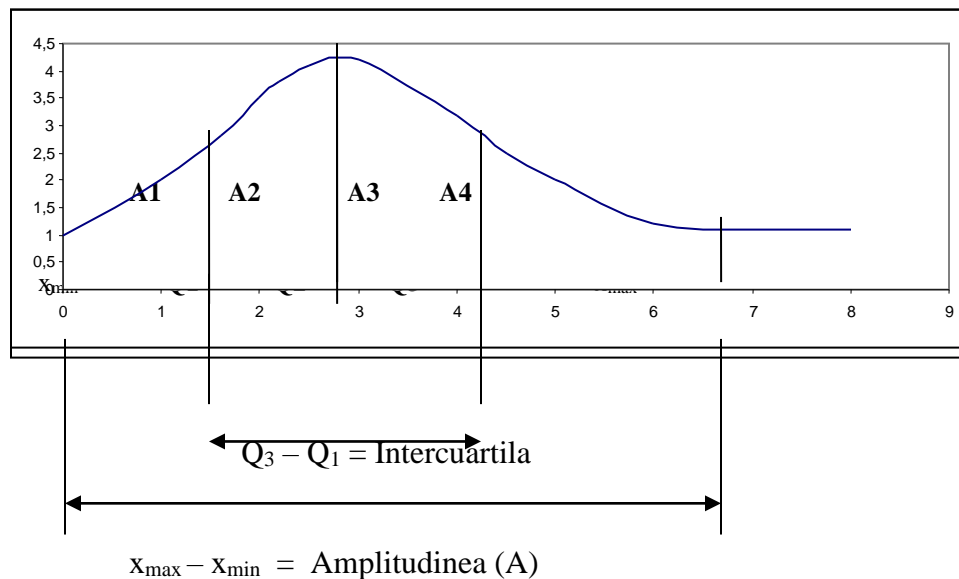
- a) oferă o imagine generală asupra împrăștierii
- b) consideră doar valorile extreme
- c) sensibilă la valorile extreme (în particular la valorile aberante)
- d) nu se pretează la calcule algebrice

Intercuartilă

Notăție: IQ

Definiție: Intercuartila reprezintă intervalul intercuartil (abaterea cuartilă este diferența între cuartila superioară și cuartila inferioară ($Q_3 - Q_1$))

Curba de frecvență



Proprietăți

- a. Intercuartila exprimă abaterea față de mediană a aproximativ 40% dintre valori.
- b. Nu consideră valorile extreme (în particular valorile aberante)
- c. Oferă o indicație despre împrăștierea celor 50% din valorile grupate în centrul repartiției, astfel:

dacă $IQ \leq A/2$, distribuția este intens concentrată

dacă $IQ > A/2$, distribuția este intens dispersată.

d. Nu se pretează la calcule algebrice.

Dispersia (Variația/fluctuația/sigma pătrat σ^2)

Notăție: S^2 (pentru populații în general) σ^2 pentru populații teoretice) s^2 (pentru eșantioane).

Definiții:

- a) În cazul unei serii statistice formate din N valori distincte sau nu $x_1, x_2, x_3, \dots, x_j, \dots, x_N$ dispersia este media pătratelor abaterilor (valorilor seriei) față de media seriei :

$$(1) S^2 = \frac{\sum_{j=1}^N (x_j - M)^2}{N}$$

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

$$(2) M = \frac{\sum_{j=1}^p N_j \cdot (x_j - M)^2}{\sum_{j=1}^p N_j}$$

,unde $\sum_{j=1}^p N_j = N$ (volumul seriei)

Numaratorul din expresiile (1) si (2) $\sum_{j=1}^N (x_j - M)^2$; $\sum_{j=1}^p N_j \cdot (x_j - M)^2$ se noteaza cu V si se numeste variatia seriei.

Proprietățile dispersiei:

- a) Este o valoare pozitivă sau nulă, fiind o sumă de pătrate (este nulă dacă șirul este constant);
- b) Se utilizează pentru:
 - b1. Compararea variabilității unui caracter în două sau mai multe populații pentru care datele au același ordin de mărime
 - b2. compararea a două sau mai multe caractere ale aceleiași populații, dacă acestea sunt exprimate în aceeași unitate de măsură și valorile au același ordin de mărime (medii apropiate),
- c) Ține cont de toate valorile din cadrul seriei;
- d) Numărătorul expresiei sale, variația, îndeplinește o proprietate de aditivitate.
- e) Este sensibilă la valorile extreme (în particular, la cele aberante)
- f) Are alt ordin de mărime față de datele inițiale și medie (se exprimă în unitatea de măsură a datelor ridicată la pătrat).

Abaterea standard (abaterea medie pătratică / derivația standard / σ -ul seriei / abaterea tip SD serie - Standard Derivation).

Notatii: **S** – pentru populații statistice în general,
 σ – pentru populații statistice teoretice
 s – pentru eșantioane

Definiție: Rădăcina pătrată din dispersie,

$$S = \frac{\sum_{j=1}^N (x_j - M)^2}{N}, \quad N = \text{volumul seriei}$$

Serii statistice grupate în distribuția de frecvențe absolute (x_j, N_j), a celor $p \leq N$ valori distincte, x_j

$$S = \frac{\sum_{j=1}^p N_j \cdot (x_j - M)^2}{\sum_{j=1}^p N_j}$$

Proprietăți

a) Variante abatere standard :

- este un număr pozitiv sau nul, fiind rezultatul extragerii unui radical de ordin par;
- este nulă dacă și numai dacă șirul este constant

b) Se utilizează pentru:

- Compararea variabilității unui caracter în două sau mai multe populații pentru care datele au același ordin de mărime (medii apropiate);
- Compararea a două sau mai multe caractere ale aceleiași populații, dacă acestea sunt exprimate în aceeași unitate de mărime (medii apropiate)

c) Ține cont de toate valorile din cadrul seriei

d) Au alt ordin de mărime față de datele inițiale și medie

Coeficientul de variație

Notății: CV%, CV, Cv, V

Definiție: Fie o serie de valori pe o scală raport. Coeficient de variație = proporția reprezentată de abaterea standard (S) din medie (M):

$$CV = S / M = S \cdot 100 / M \% = CV\%$$

Se utilizează des, în exprimarea procentuala notată $CV\%$ (coeficient procentual de variație) = procentul reprezentat de abaterea standard (S) din medie (M).

Proprietăți:

a) $CV\% \geq 0$, deoarece $S \geq 0$ și $M > 0$, fiindcă orice șir pe o scală raport nu are valori negative și nici medie negativă.

b) $CV\% = 0$, dacă $S = 0$, adică dacă șirul de date este constant.

c) Se utilizează în special atunci când nu pot fi utilizate dispersia sau abaterea standard, în scopul comparării variabilității:

- unui caracter în doua sau mai multe populații dacă valorile măsurate au ordine de mărime diferite;
- doua sau mai multe caractere în aceeași populație, dacă acestea sunt exprimate, fie în unități de măsură diferite, fie în aceeași unități de măsură, dar diferite.

d) Se poate utiliza și în cazurile recomandate pentru folosirea dispersiei sau abaterii standard; coeficientul de variație este indicatorul universal de comparare a variabilității, pe scala raport.

e) Ține cont de toate valorile din cadrul seriei

f) $CV\%$ este independent de unitatea de măsură folosită pentru valorile seriei, este adimensional și se exprimă procentual.

g) Este sensibil la valorile extreme (inclusiv la valori aberante).

h) Valabil numai pentru măsurătorile pe scale raport.

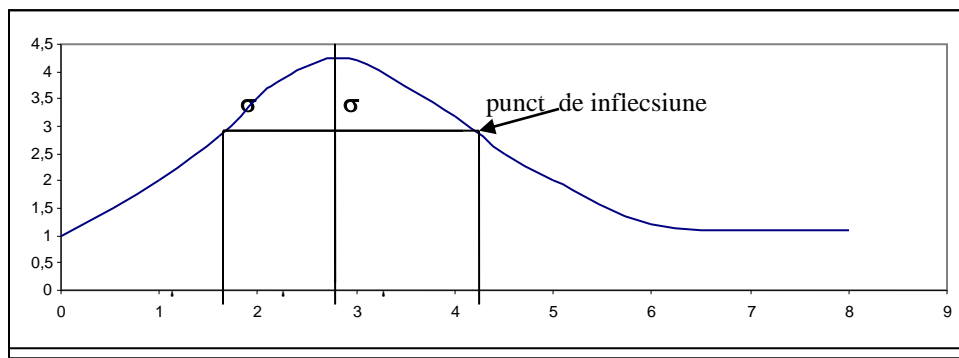
Distribuția normală (curbă a erorilor - de măsurare întâmplătoare / clopot a lui Gauss / distribuție Laplace)

Descriere:

- Distribuție continuă în formă de clopot (unimodală și simetrică)
- Este caracterizată de doi parametri specifici pentru μ și σ

μ – media aritmetică

σ – abatere standard



- Are două puncte de inflexiune situate simetric față de verticală $x = \mu$, la distanța σ

Distribuție normală și consultarea tabeli corespunzătoare

Dintre distribuțiile normale se distinge distribuția cu $\mu = 0$ și $\sigma = 1$, care se numește distribuția normală standard și se notează $N(0,1)$.

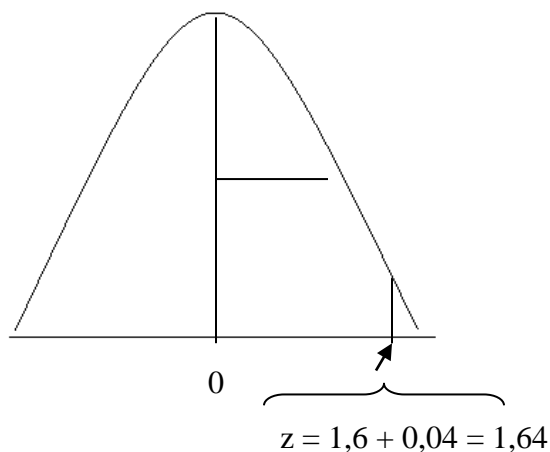
Determinarea ariilor la dreapta punctelor și a cuartilelor superioare

Se poate realiza direct prin consultarea tabeli de cuartile superioare din anexa 1 la acest material. Utilizarea tabeli:

- a) pentru determinarea proporției de arie α (aria relativă α) aflată sub distribuția normală standard la dreapta unui punct dat, z .
- b) pentru determinarea punctului z care lasă la dreapta sa, sub distribuția normală standard, aria relativă α

Exemplu

- a) Aria relativă α se află la dreapta punctului $z = 1,64$ se obține citind în tabela a doua din anexa 1, valoarea înscrisă la intersecția liniei 1,6 cu coloana 0,4 (care însumate dau valoarea 1,64). Se obține $\alpha = 0,0505 = 0,05 = 5\%$.



- b) Valoarea z care lasă la dreapta sa aria relativă $\alpha = 0,05$ se află căutând în aceeași tabelă o valoare cât mai apropiată de valoarea α căutată. În acest caz, aceasta poate fi 0,050 sau 0,495 (ambele la aceeași distanță de $\alpha = 0,05$). Alegem una dintre acestea de exemplu 0,0505 și citim pe linie valoarea 1,6, iar pe coloana corespunzătoare, 0,04. Valoarea z va fi suma dintre ultimele două numere: $z = 1,6 + 0,04 = 1,64$.

Reținem că aria relativă aflată la dreapta unui punct sub distribuția normală standard este tabelată (anexa 1) iar aria din stânga este complementul față de 1 al ariei tabelate.

Tratarea unei variabile calitative

Tratarea calitativă a unei variabile calitative. O variabilă calitativă se manifestă printr-o serie statistică univariată, calitativă (x_i) $i = 1, 2, \dots, N$ unde x_i sunt variante distincte ale variabilei.

Exemplu:

Se dă seria de culori ale unor flori:

(alb, roșu, galben, alb, verde, alb, roșu, galben, alb, alb)

Seria prezentată grupat ca o distribuție de frecvențe absolute ale variantelor distincte x_j , arată astfel:

$$\left\{ \begin{array}{c} x_j \\ N_j \end{array} \right\}_{j=1, \dots, p} \quad (x_j, N_j)_{j=1, \dots, p}$$

unde $\sum_{j=1}^p N_j = N$

Seria din exemplu devine: $\left\{ \begin{array}{cccc} \text{alb} & \text{roșu} & \text{galben} & \text{verde} \\ 5 & 2 & 2 & 1 \end{array} \right\}$

Distribuția de frecvențe relative al variabilelor distincte x_j , notată

$$\left\{ \begin{array}{c} x_j \\ F_j \end{array} \right\}_{j=1, \dots, p} \quad (x_j, F_j)_{j=1, \dots, p}$$

unde $\sum_{j=1}^p F_j = 1$

în cazul nostru: $\left\{ \begin{array}{cccc} \text{alb} & \text{roșu} & \text{galben} & \text{verde} \\ 5/10 & 2/10 & 2/10 & 1/10 \end{array} \right\}$

Binarizarea unei variabile calitative

Tratarea cantitativă a unei variabile calitative presupune studierea unei singure variante în opoziție cu ceea ce rămâne în afara ei = binarizarea variabilei calitative.

În exemplul de mai sus, dacă ne interesează doar culoarea alb, în opoziție cu celelalte culori, sintetizăm distribuția binară

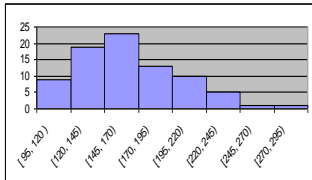
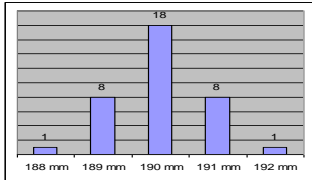
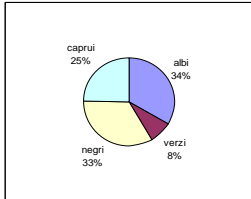
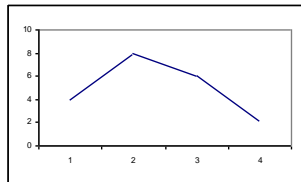
$$\left\{ \begin{array}{cc} \text{alb} & \text{non-alb} \\ 5/10 & 5/10 \end{array} \right\}$$

În general , pentru o distribuție de frecvențe relative a unei variabile calitative:

$$\left\{ \begin{array}{c} x_1, \ x_2, \dots, \ x_p \\ F_1, \ F_2, \dots, \ F_p \end{array} \right\}$$

dacă ne interesează variația x_j în opoziție cu restul, sintetizăm distribuția binară

$$\left\{ \begin{array}{cc} x & \text{non } x \\ F & 1 - F \end{array} \right\}$$

		Variabila				
		cantitativa		calitativa		
		tip masuratoare	tip rang			
S i n f e c z a d a t e l o r	g r a f i c e	Grupare in	tabel statistic simplu			
		Reprezentari grafice tip	histograma	diagrama cu batoane	diagrama circulara	
						
	poligon de frecvente					
						
	n u m e r i c a	In valori tipice de :				
		Tendinta centrala	M (media)	Me (mediana)	Mo (moda)	Pentru variabile binarizate : proportiile p, q (= 1-p)
		Variabilitate ca imprastiere	S (abaterea standard) S² (dispersia) CV% (coeficientul de variatie)	IQ (intercartila) A (amplitudinea)		Pentru variabile binarizate : S² si S specifice S² = p*q ; S = √ p*q
		Variabilitate ca diversitate				p (numar de variante), impreuna cu H_{rel} (entropia relativa)

Bibliografie

1. A. Indrayan, Medical biostatistics, Second Edition, Chapman and Hall Publishing, 2008
2. Andrew Tanenbaum, Organizarea structurata a calculatoarelor, 2005
3. Dragomirescu L., Drane J.W., - Biostatistica pentru începători, Ed. Ars Docenti, București, 2001