

Capitolul 3

STATISTICĂ DESCRIPTIVĂ UNIVARIATĂ

În continuare vom trata sinteza grafică univariată și sinteza numerică univariată.

§ 3.1. Sinteza grafică univariată

Sinteza grafică se face pentru evidențierea *intuitivă* și *aproximativă* a aspectelor esențiale de variabilitate dintr-o serie statistică.

Sinteza grafică se execută în doi pași și anume construirea de:

- ◆ *tabele statistice* denumite *simple* sau *cu simplă intrare* și
- ◆ reprezentări grafice adecvate tipului de variabilă și anume:
 - pentru variabile calitative și ordinale:
 - *diagrame circulare*,
 - *diagrame prin coloane* și *diagrame prin benzi*;
 - pentru variabile ordinale și cantitative:
 - *poligoane de frecvențe*,
 - *histograme*.

Recomandăm pentru variabile:

- calitative - diagramele circulare,
- ordinale - diagramele prin coloane sau, uneori, poligoanele de frecvențe,
- cantitative – diagramele prin coloane sau prin benzi, poligoanele de frecvențe și, mai ales, histogramele.

Sinteza grafică în tabele statistice se poate face prin:

- ◆ grupare, fără pierdere de informație
 - în tabele statistice simple cu frecvențele variantelor, rangurilor ori valorilor, altfel spus,
 - construind distribuțiile frecvențelor variantelor, rangurilor ori valorilor, denumite, pe scurt, **distribuții de frecvențe (negrupate)**;
- ◆ grupare, cu pierdere de informație
 - în tabele statistice simple cu frecvențele *claselor* sau *intervalelor de grupare*, altfel spus,
 - construind distribuțiile frecvențelor claselor sau intervalelor de grupare, denumite, pe scurt, **distribuții de frecvențe grupate**.

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

- ✓ O *distribuție de frecvențe* conține aceeași informație ca și seria din care provine, dar este mai intuitivă, fiind mai apropiată de reprezentarea grafică care urmează a se construi.
- ✓ O *distribuție de frecvențe grupate* conține mai puțină informație decât seria din care provine dar poate oferi un câștig în relevanță ca în cazul unei caricaturi, care este mai relevantă decât o fotografie, pentru esențialul fizionomiei unei persoane.

Variante distincte x_j	Frecvențe absolute N_j	Frecvențe relative $F_j = N_j / N$	Frecvențe (relative) procentuale $P_j = 100 \cdot F_j \%$	Frecvențe procentuale cumulate $PC_j = P_1 + P_2 + \dots + P_j$
a	4	4 / 12	$100 \cdot 4 / 12\% \approx 34\%$	34%
v	1	1 / 12	$100 \cdot 1 / 12\% \approx 8\%$	42%
n	4	4 / 12	$100 \cdot 4 / 12\% \approx 33\%$	75%
c	3	3 / 12	$100 \cdot 3 / 12\% = 25\%$	100%
Totaluri:	$N = 12$	1	100%	

Valori distincte x_j	Frecvențe absolute N_j	Valori distincte x_j	Frecvențe absolute N_j
6	2	188	1
7	5	189	8
8	3	190	18
9	1	191	8
10	1	192	1
Totaluri:	$N = 12$	Totaluri:	$N = 36$

Pentru S2, respectiv, pentru S3.

- Perechile $(x_j, N_j)_{j=1,2,\dots,p}$ se numesc **distribuții** sau **repartiții de frecvențe absolute**,
- perechile $(x_j, F_j)_{j=1,2,\dots,p}$ se numesc **distribuții** sau **repartiții de frecvențe relative**, și
- perechile $(x_j, P_j)_{j=1,2,\dots,p}$ se numesc **distribuții** sau **repartiții de frecvențe (relative) procentuale**
- perechile $(x_j, PC_j)_{j=1,2,\dots,p}$ se numesc **distribuții** sau **repartiții de frecvențe procentuale cumulate** și putem adăuga, **ale variantelor / rangurilor / valorilor șirului**, pentru a le deosebi de *clasele* sau *intervalele* de la distribuțiile grupate.

✓ Adeseori, distribuțiile de frecvențe se scriu sub forma $\begin{pmatrix} x_j \\ N_j \end{pmatrix}$. Exemple în 8°. Altfel spus,

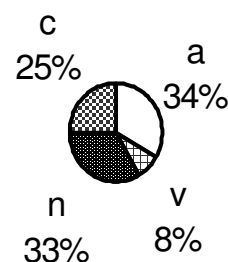
primele două coloane ale unei astfel de tabele formează o *distribuție* sau *repartiție de frecvențe absolute*, iar prima și a treia coloană o *distribuție* sau *repartiție de frecvențe relative*. Pentru a se deosebi de distribuțiile grupate prezentate mai jos la punctele 3° și 4°, acestea sunt denumite **distribuții negrupate**, deși reprezintă rezultatul grupării și eventual al ordonării unităților șirului, dar fără pierdere de informație.

✓ Observăm că am notat $\{x_j\}_{j=1,2,\dots,p}$ șirul *variantelor / rangurilor / valorilor distincte* pentru a se deosebi de șirul *variantelor / rangurilor / valorilor distincte sau nu*, $\{x_i\}_{i=1,2,\dots,N}$. Evident

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

2° Reprezentări grafice univariate pentru distribuții negrupate

Diagramă circulară = cerc format din sectoare pentru fiecare variantă / rang / valoare, x_j astfel încât unghiul, respectiv *aria* fiecărui sector să fie proporțional(ă) cu frecvența respectivă.

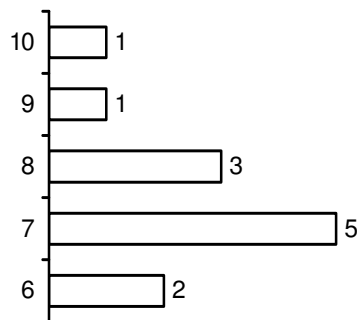


Exemplu pentru seria S1 →

Diagramă prin (în) benzi sau bare = reprezentare carteziană plană, în care pe axa verticală avem marcate variantele / rangurile / valorile, în fiecare fiind construită o *bandă* orizontală de lungime proporțională cu frecvența corespunzătoare.

Benzile sunt dreptunghiuri nealipite și de aceeași lățime, de regulă mult mai mică decât lungimile lor.

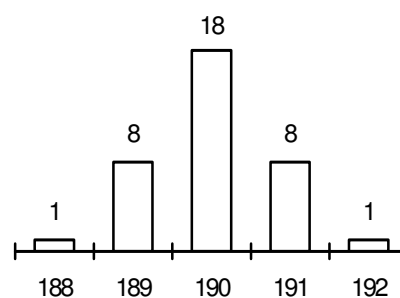
Exemplu pentru seria S2 →



Diagramă prin (în) coloane sau batoane = reprezentare carteziană plană, în care pe axa orizontală avem marcate variantele / rangurile / valorile, în fiecare fiind construită pe verticală o *coloană* de înălțime proporțională cu frecvența corespunzătoare.

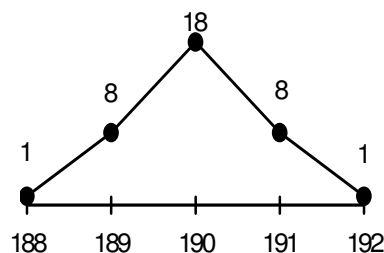
Coloanele sunt dreptunghiuri nealipite și de aceeași lățime, de regulă mult mai mică decât înălțimile lor.

Exemplu, pentru seria S3 →



Poligon de frecvențe = linia frântă formată din segmentele care unesc mijloacele laturilor din vârfurile coloanelor consecutive figurate în diagrama prin coloane, fără a mai reprezenta și coloanele.

Exemplu pentru seria S3 →



- ✓ Toate distribuțiile provenite din serii statistice empirice, chiar dacă provin din variabile continue, sunt **distribuții discrete** în sensul că mulțimea valorilor este discretă. Aceasta se întâmplă deoarece orice serie statistică empirică este, prin construcție, finită deci discretă.

3^o Reprezentare grafică univariată pentru distribuții teoretice ale variabilelor cantitative

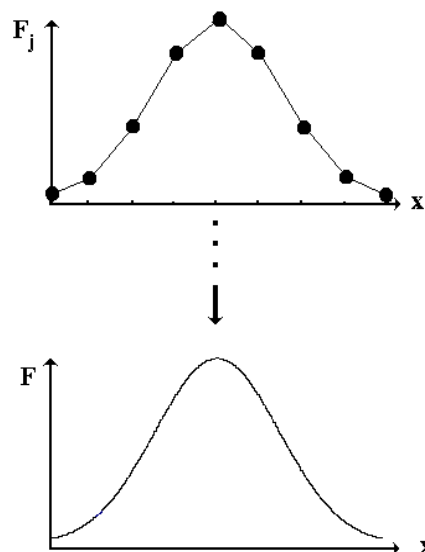
Putem imagina și **distribuții continue**, pornind de la variabile continue (măsurători) și construindu-le teoretic, altfel spus, considerând populații statistice infinite nenumărabile (ca \mathbb{R} , mulțimea numerelor reale).

Curbă (teoretică) de frecvențe pentru o variabilă cantitativă continuă (măsurătoare).

Să presupunem acum că vom crește din ce în ce mai mult precizia și numărul de măsurători ale lungimii cărții și vom reprezenta ca poligoane de frecvențe fiecare distribuție cu număr din ce în ce mai mare de măsurători.

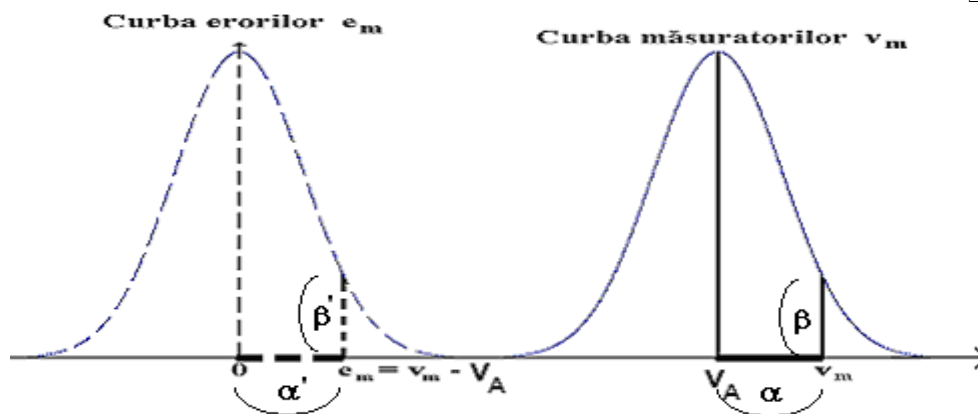
De exemplu, într-o primă fază putem obține o distribuție cu 5 valori distincte ca mai sus, într-o a doua, 9 etc.

Se observă intuitiv că *poligoanele de frecvențe* (x_j, F_j) tind, odată cu creșterea preciziei și a numărului de măsurători, către o curbă teoretică numită **curbă de frecvențe** și notată (x, F). În acest caz curba teoretică este clopotul lui Gauss.



Exemplu:

Să notăm cu v_A valoarea adevărată a lungimii cărții (pe care nu o cunoaștem și nu o putem determina exact niciodată) și să presupunem că am efectuat un număr foarte mare de măsurători cât mai precise, fără să facem erori grosolane de măsurare. Atunci măsurătorile vor tinde să se grupeze într-o curbă de frecvențe, cu atât mai evident cu cât precizia și numărul măsurătorilor sunt mai mari. Dacă notăm cu v_m o *valoare măsurată* și cu $e_m = v_m - v_A$ *eroarea de măsurare întâmplătoare* corespunzătoare, *curba de frecvențe ale măsurătorilor*, v_m și *curba erorilor de măsurare întâmplătoare*, e_m vor avea forme identice dar prima curbă se va centra în jurul lui v_A , iar a doua în jurul lui 0:



Curba erorilor de măsurare întâmplătoare sau *aleatoare*, denumită pe scurt **curba erorilor** având forma unui clopot este cunoscută și sub denumirea de **clopot al lui Gauss**. Forma de clopot simetric a *curbei măsurătorilor*, respectiv a *curbei erorilor aleatoare* inerente oricărui proces de măsurare exprimă următoarele fapte experimentale:

În cazul curbei erorilor:	În cazul curbei măsurătorilor:
(1) Marea majoritate a erorilor de măsurare (α' în figura anterioară) au valori apropiate de centrul distribuției, de zero.	1. Marea majoritate a măsurătorilor (α) au valori apropiate de centrul distribuției, de valoarea adevărată v_A .
	2. Numărul măsurătorilor (β) care se abat de

- | | |
|---|---|
| <p>(2) Numărul erorilor (β) care se abat de la centru scade o dată cu creșterea abaterii de la centru și</p> <p>(3) numărul erorilor cu semn pozitiv este relativ egal cu numărul erorilor cu aceeași valoare absolută, dar cu semn negativ.</p> | <p>la centru scade o dată cu creșterea abaterii de la centru și</p> <p>3. numărul măsurătorilor cu o anumită abatere pozitivă este relativ egal cu numărul măsurătorilor cu aceeași abatere, dar cu semn negativ.</p> |
|---|---|

Punctele 1 și 2 explică forma cu o singură "cocoasă", iar punctul 3 explică simetria curbei erorilor.

Intuitiv, valoarea centrală din curba măsurătorilor trebuie să fie, cel mai probabil, valoarea adevărată. (De aici rezultă importanța sintezei numerice care va urma, mai precis sensul parametrilor de tendință centrală.)

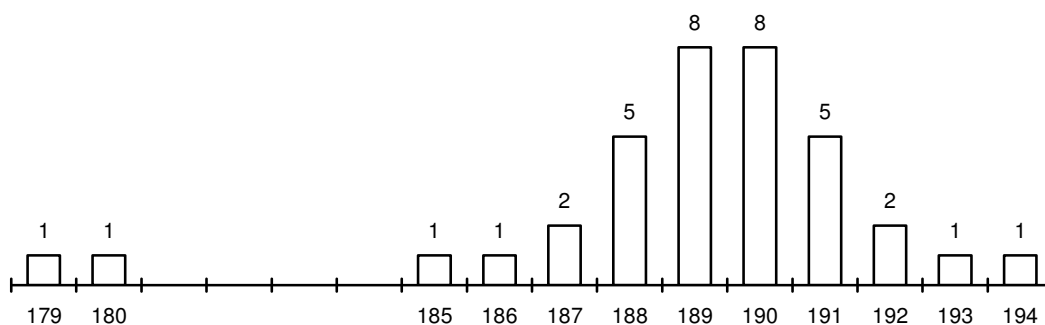
Graficul pentru distribuția corespunzătoare șirului S3 "ascunde", deci, în spatele său clopotul lui Gauss, ceea ce confirmă corectitudinea măsurătorilor executate în acest caz.

Valori aberante

Aceiași 36 de studenți care au măsurat o carte, producând seria S3, au măsurat cu aceeași precizie (sau eroare) de $\pm 0,5 \text{ mm}$ și lungimea palmei unui voluntar (distanța între prima brățară și vârful degetului mijlociu – dimensiune chirometrică). S-a obținut seria S4, scrisă în continuare ca distribuție de frecvențe și reprezentată ca diagramă în batoane. Valorile x_j sunt exprimate în *mm*.

S4:

x_j	N_j	x_j	N_j	x_j	N_j
179	1	187	2	191	5
180	1	188	5	192	2
185	1	189	8	193	1
186	1	190	8	194	1
Total: 36					



Constatăm că procesul de măsurare a fost afectat nu numai de inerentele erori întâmplătoare, care sunt relativ mici și se produc în ambele sensuri și deci se compensează reciproc, ci și de unele erori grosolane. Acestea sunt denumite **erori sistematice**, deoarece sunt în mod sistematic în același sens și deci produc o deplasare sistematică a valorii căutate. Procesul de măsurare nu s-a desfășurat, deci, corect.

Studiind această diagramă, bunul simț ne îndeamnă să considerăm că primele două măsurători - 179 și 180 - sunt afectate de erori sistematice. Astfel de măsurători se numesc, în

statistică, *valori aberante*. Foarte probabil, cele două măsurători au fost făcute asupra palmei insuficient întinse.

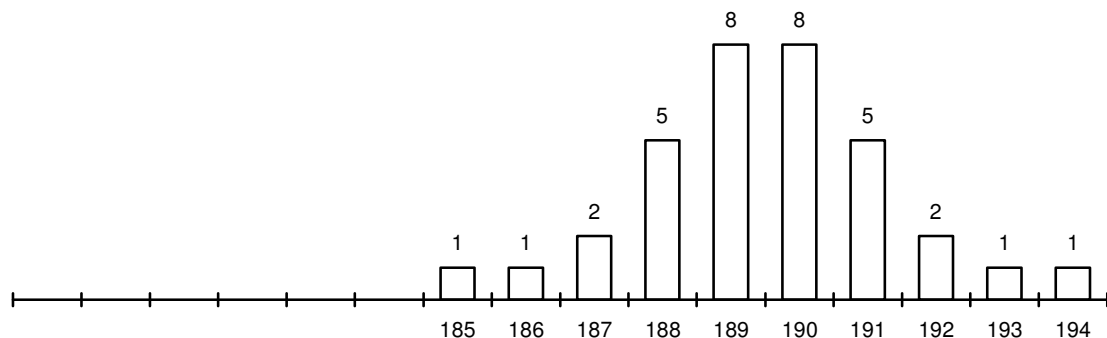
O valoare care contrastează puternic cu marea majoritate a celorlalte valori ale șirului, altfel spus, “iese din regula” șirului se numește **valoare aberantă**.

Asumându-ne riscul de a considera aberante aceste două măsurători, le putem elimina obținând seria **S4'** (= S4 fără valorile aberante), tabelată și reprezentată ca diagramă în batoane în continuare. (În paragraful dedicat distribuției normale vom vedea că există criterii statistice pentru detectarea valorilor aberante, dacă acceptăm anumite ipoteze asupra distribuției datelor. De fapt detectarea acestor valori altfel decât intuitiv, așa cum am făcut aici, este o problemă de statistică inductivă.)

S4':

x_j	N_j	x_j	N_j
185	1	190	8
186	1	191	5
187	2	192	2
188	5	193	1
189	8	194	1

Total: 34

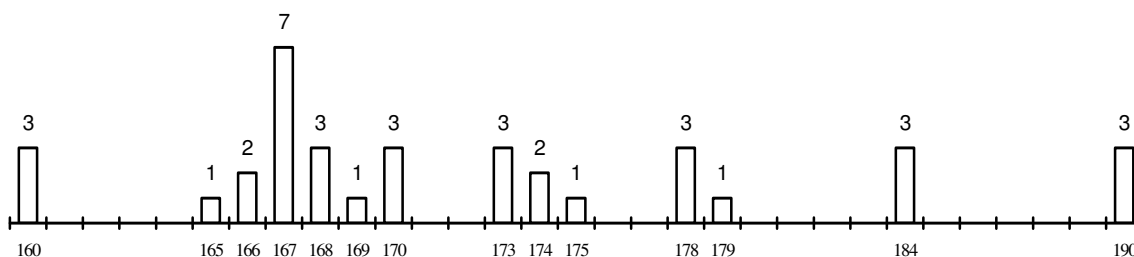


4° Distribuții grupate pentru variabile cantitative și histograma

Măsurându-se lungimea palmei drepte la 36 studenți s-a obținut șirul **S5**, care, grupat fără pierdere de informație, ca distribuție de frecvențe, este figurat în următorul tabel statistic simplu, reprezentat apoi ca diagramă în batoane:

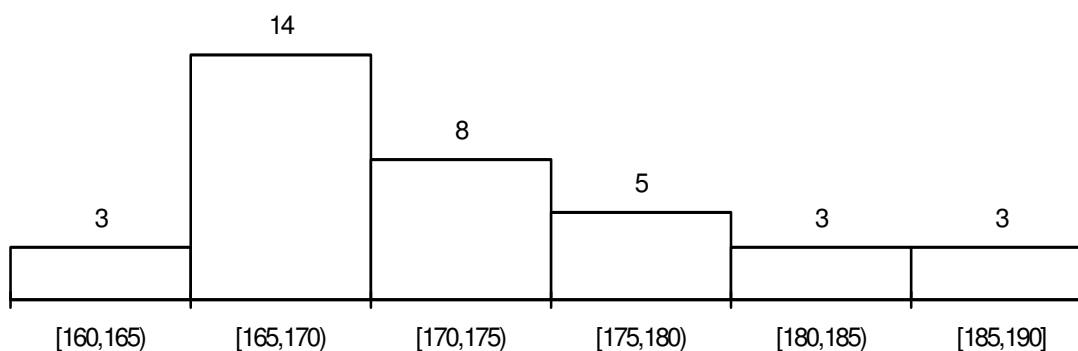
S5:

x_j	N_j	x_j	N_j	x_j	N_j
160	3	169	1	178	3
165	1	170	3	179	1
166	2	173	3	184	3
167	7	174	2	190	3
168	3	175	1		
				Total:	36



Deoarece reprezentarea anterioară nu ne sugerează nici o formă "frumoasă", relevantă, vom grupa datele folosind intervale consecutive de grupare denumite **intervale de grupare**, **intervale de clasă** sau, mai general, **clase (de grupare)**. Le vom lua de lungimi egale, de exemplu, de lungime 5. *Distribuția de frecvențe ale intervalelor de grupare sau, ale claselor*, denumită, mai scurt, **distribuție grupată**, și notată **S5'** (= S5 grupată) se poate tabela și reprezenta după cum urmează.

S5'		intervalul de clasă	N_j	intervalul de clasă	N_j
		[160, 165)	3	[175, 180)	5
		[165, 170)	14	[180, 185)	3
		[170, 175)	8	[185, 190] ¹	3
					Total: 36



O astfel de reprezentare se numește *histogramă*. Se observă că, spre deosebire de diagrama în batoane, histograma conține dreptunghiuri alipite, deoarece intervalele de grupare, în comparație cu valorile seriei, sunt întotdeauna alipite. Pentru că intervalele de grupare pot avea și lungimi diferite, se convine ca ariile dreptunghiurilor să fie proporționale cu frecvențele intervalelor de grupare. Aceasta este a doua deosebire față de diagrama în batoane la care înălțimile sunt proporționale cu frecvențele.

Pentru o distribuție grupată, se poate da, prin urmare, următoarea definiție.

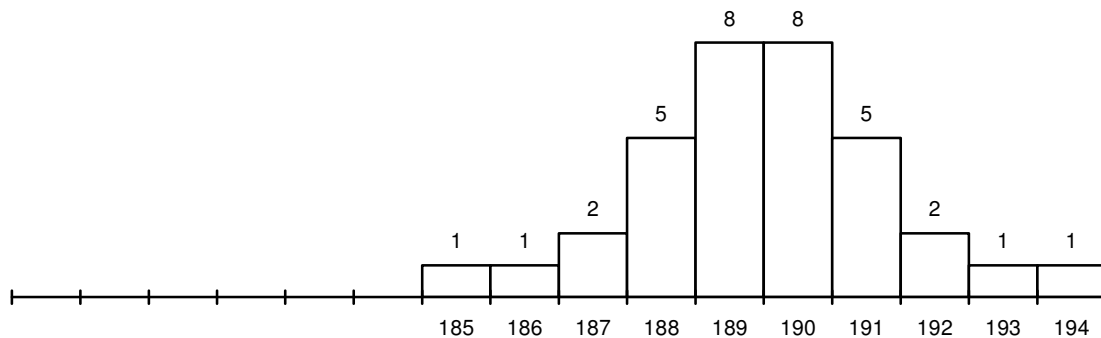
Histogramă = reprezentare carteziană plană a unei distribuții grupate, formată din dreptunghiuri alipite, cu bazele plasate pe intervalele de grupare și cu *ariile* proporționale cu frecvențele intervalelor de grupare, claselor.

- ✓ Dacă intervalele de grupare (de clasă) sunt egale, atunci vor fi proporționale cu frecvențele și *înălțimile*.

¹ Ultimul interval este închis și la dreapta pentru a nu se pierde valoarea maximă din șir.

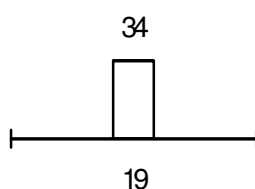
- ✓ Prin acest mod de grupare s-a pierdut o parte din informație. De exemplu, această diagramă nu ne poate spune câte unități au valoarea 165, ci doar câte au valori cuprinse între 165 și 170.
- ✓ Renunțând, însă, la o parte din informație am câștigat în relevanță, deoarece "în spatele" acestei histogramme putem "întrezări" o formă relevantă, cea a unui clopot asimetric. Histograma distribuției grupate S5' este, deci, mai relevantă decât diagrama în batoane a distribuției negrupate S5, ca urmare a pierderii de informație.
- ✓ Modul de alcătuire a unei histogramme pentru o distribuție grupată este un bun model intuitiv al paradigmei centrale a statisticii, enunțate mai sus.
- ✓ Deseori se reprezintă ca histogramă distribuții de frecvențe (negrupate) de valori întregi, considerându-se, drept intervale de grupare, intervalele unitare centrate în valorile respective.

De exemplu, seria S4' poate fi prezentată ca histogramă în acest mod, în care 185 înseamnă de fapt intervalul [184,5; 185,5), 186 înseamnă intervalul [185,5; 186,5) etc.



O imagine și mai sugestivă a ideii că prin gruparea cu pierdere de informație obținem câștig în relevanță, o putem obține dacă rotunjim – rotunjirea fiind o formă de grupare - la cifra zecilor, valorile din seria S4'. Altfel spus, dacă vom lucra doar cu număr întreg de centimetri vom obține seria notată **S4''** (= S4' cu valorile rotunjite la cifra zecilor) care, prezentată în tabel ca distribuție grupată, respectiv reprezentată sub formă de histogramă, va fi o **distribuție concentrată într-un punct**:

S4'' :	Valori distincte x_j (în cm)	Frecvențe absolute N_j
	19	34



Câștigul în relevanță este evident aici și pentru un necunoscător al statisticii: aflarea lungimii palmei voluntarului, ce-i drept cu o eroare mai mare decât eroarea de măsurare de $\pm 0,5$ mm, și anume eroarea de $\pm 0,5$ cm, provenită din rotunjire.

Vom putea afirma cu certitudine că palma măsurată prin mai multe replicare și grupată în S4'', are lungimea de $19\text{ cm} \pm 0,5\text{ cm}$. Deoarece am specificat mărimea erorii – de grupare, ca formă de aproximare, în acest caz – exprimarea este științifică. Exprimându-ne mai tehnic spunem că aproximarea este exactă. Este maximum posibil, deoarece în cazul măsurătorilor propriu-zise este imposibil un rezultat exact.

- ✓ Prin statistică obținem "*aproximări exacte*, nu *exactități aproximative*", [24]. Exprimarea (semidocă), aplicată exemplului de mai sus, "palma are (exact) 19 cm" este un exemplu

de exactitate aproximativă. În general, specificarea unui singur număr în cazul unei măsurători a unei variabile continue este o formă de aproximare neștiințifică.

- ✓ Observăm că și numărătorile pot fi grupate, la fel ca măsurătorile, cu pierdere de informație. Altfel spus, „rețeta” de mai sus se poate aplica oricărei variabile cantitative.

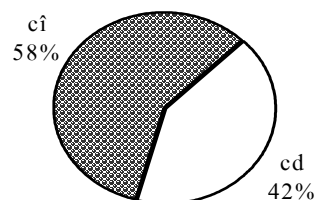
5° Distribuții grupate pentru variabile calitative și variabile ordinale

Distribuții grupate se pot construi și în cazul variabilelor calitative și în cel al variabilelor ordinale. Pentru acestea se vor utiliza reprezentările grafice adecvate prezentate pentru distribuții negrupate. Definițiile acestora se vor modifica înlocuindu-se termenii variantă, respectiv rang, cu cel de clasă.

Cazul variabilelor calitative

De exemplu, în cazul șirului S1 putem grupa culorile verde și albastru în *clasa* “culoare deschisă (cd)” și culorile căprui și negru în *clasa* “culoare închisă (cî)”. Se va obține șirul notat **S1'** (= S1 grupat) descris mai jos ca distribuții de frecvențe absolute, relative și (relative) procentuale și reprezentat ca diagramă circulară.

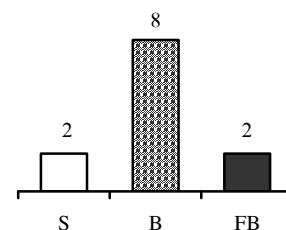
Variante x_j	Frecvențe absolute N_j	Frecvențe rel. F_j $= N_j / N$	Frecv. (rel.) procentuale P_j
(cd)	5	5 / 12	5*100/12% \approx 42%
(cî)	7	7 / 12	7*100/12% \approx 58%
Totaluri:	$N=12$	1	100%



Cazul variabilelor ordinale

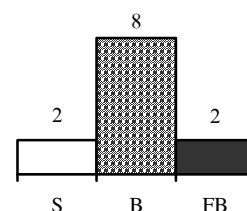
În cazul șirului S2 - șir de ranguri - putem grupa notele după regula tradițională: notele 5 și 6 formează *clasa* “Suficient”, 7 și 8, *clasa* “Bine”, iar 9 și 10, *clasa* “Foarte Bine”. Se va obține astfel seria **S2'** (= S2 grupat) prezentată în continuare ca distribuții de frecvențe absolute, relative și (relative) procentuale și reprezentată adecvat ca diagramă în batoane.

Clase x_j	Frecvențe absolute N_j	Frecvențe rel. $F_j = N_j / N$	Frecv. rel. procentuale P_j
Suficient {5, 6}	2	2 / 12	2*100/12 % \approx 17%
Bine {7, 8}	8	8 / 12	8*100/12 % \approx 66% #
Foarte Bine {9, 10}	2	2 / 12	2*100/12 % \approx 17%
Totaluri:	$N=12$	1	100%



Valoarea marcată cu “#” este rotunjită prin trunchiere pentru ca suma procentelor să fie 100 %.

Dacă notele sunt obținute printr-un sistem de evaluare asemănător celebrului *IQ* („*Intelligence Quotient*” în engleză), coeficient de inteligență, atunci se poate accepta licența că sunt mai mult decât ranguri. Astfel, putem forma cele trei clase din intervalele de grupare [5, 7), [7, 9), respectiv, [9, 10] și putem reprezenta seria S2' printr-o histogramă.



6° Probleme rezolvate

1. Care dintre seriile S3-S5' măsoară entități constante și care variabile ?
R: S5 și S5' - entități variabile.
2. Care dintre seriile S3-S5' este o serie constantă și care este variabilă ?
R: S4'' - serie constantă.
3. De cine depinde constanța sau variabilitatea unei serii ?
 - a. de constanța sau variabilitatea entității măsurate ?
 - b. de precizia măsurării ?
 - c. de ambele ?R: b.
4. Examinând graficele corespunzătoare, aranjați în ordinea crescătoare a variabilității seriile S3, S4', S4'', S5 și explicați rezultatul.
R: S4'', S3, S4', S5.
 - S4'' este șir constant fiind format din replicare măsurate cu precizia redusă de $\pm 0,5 \text{ cm}$.
 - S3 are variabilitate mică fiind format din replicare, măsurate cu precizia de $\pm 0,5 \text{ mm}$, ale unei entități fizice.
 - S4' are variabilitate mai mare decât S3, fiind format din replicare, măsurate cu aceeași precizie de $\pm 0,5 \text{ mm}$, ale unei entități biologice mai dificil de măsurat și deci producând erori de măsurare mai mari.
 - S5 are cea mai mare variabilitate, fiind format din măsurători de aceeași precizie, de $\pm 0,5 \text{ mm}$, dar care exprimă variabilitatea biologică a lotului, care este mult mai mare decât precizia de măsurare.

7° Rezumat și rolul variabilității biologice

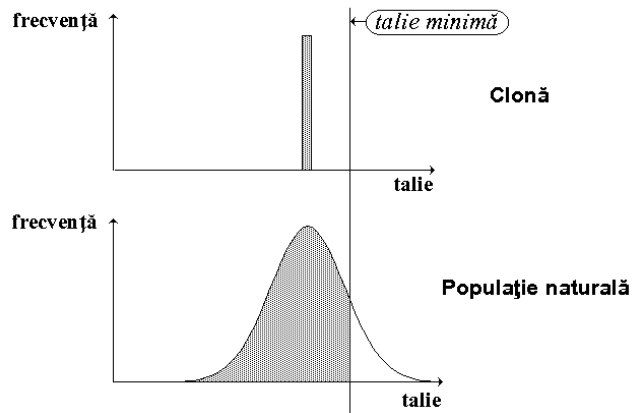
- ✓ Șirurile, seriile statistice se grupează - pentru creșterea relevanței - în tabele statistice simple fără (respectiv, cu) pierdere de informație, obținându-se distribuții negrupate (respectiv, grupate) de frecvențe absolute sau relative.
- ✓ Distribuțiile obținute sunt reprezentate grafic sub formele indicate pentru fiecare tip de variabilă. Acestea pun în evidență gradul de variabilitate al seriilor și, eventual, un centru de grupare.
- ✓ În cazul măsurătorilor replicare, se poate aprecia calitatea procesului de măsurare prin compararea distribuției cu clopotul lui Gauss. Uneori se pot evidenția și elimina valori aberante.
- ✓ În cazul seriilor de măsurători biologice nereplicare, variabilitatea este mult mai mare decât precizia de măsurare. Ca atare, variabilitatea produsă de erorile de măsurare poate fi ignorată.

Rolul variabilității biologice

Dacă variabilitatea care apare în mod inerent în orice proces de măsurare este un "zgomot" care ne împiedică să cunoaștem exact valoarea măsurată, variabilitatea biologică are, dimpotrivă, un aspect pozitiv. Ea constituie o modalitate de asigurare a supraviețuirii populațiilor biologice supuse fluctuațiilor factorilor de mediu. În consecință, variabilitatea biologică asigură conservarea speciilor. Într-adevăr, să presupunem că la un moment dat temperatura scade puternic. Conform regulii lui Bergmann [5] termoreglarea la animalele homeoterme de talie mare este mai eficace. Drept urmare se poate presupune intuitiv că există

o limită minimă a taliei care permite unui organism homeoterm să reziste la o temperatură scăzută dată.

În desenul alăturat, se observă modul diferit de răspuns la o astfel de situație al unei clone (care are variabilitatea nulă) respectiv al unei populații naturale cu variabilitate semnificativă: clona dispare, în timp ce populația naturală se salvează prin indivizii a căror talie depășește limita respectivă. (Ariile hașurate din figură reprezintă indivizii care dispar.)



+ 8^o Aplicație a poligonului de frecvențe în ecologie - distribuția de abundențe

În ecologie, pentru caracterizarea unei biocenoze, se îmbogățește artificial variabila calitativă specie (sau orice alt nivel taxonomic fixat), cu o relație de ordine produsă de abundențele nivelului taxonomic respectiv, în biocenoza dată. Dacă se operează doar cu nivelul "specie" se vorbește de **distribuție de abundențe specifice**.

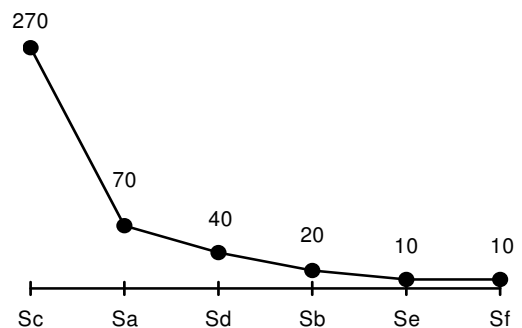
Fie următoarea distribuție de abundențe specifice, ale speciilor $S_a - S_f$, dintr-o biocenoză care conține în total 420 de indivizi provenind din 6 specii.

$$\begin{pmatrix} S_a & S_b & S_c & S_d & S_e & S_f \\ 70 & 20 & 270 & 40 & 10 & 10 \end{pmatrix}$$

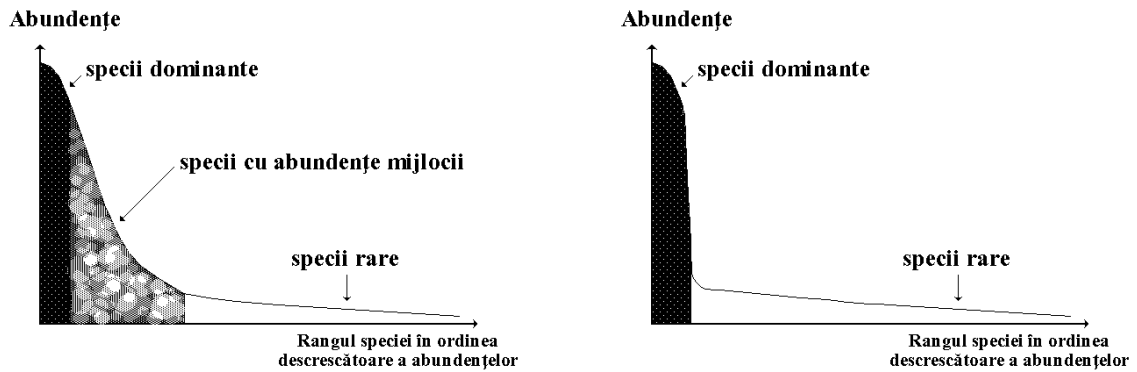
Prin convenție trebuie ca șirul statistic al speciilor prezente să fie ordonat în sensul descrescător al frecvențelor lor în biocenoză. În acest caz vom obține distribuția:

$$\begin{pmatrix} S_c & S_a & S_d & S_b & S_e & S_f \\ 270 & 70 & 40 & 20 & 10 & 10 \end{pmatrix}$$

pe care o reprezentăm sub formă de poligon de frecvențe.

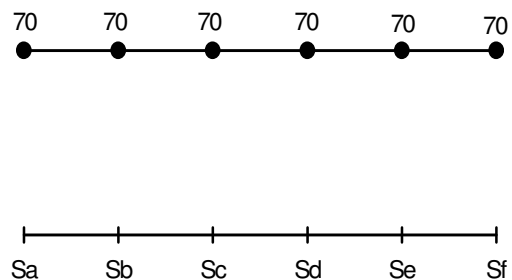


În general o distribuție de abundențe de taxoni de un nivel fixat, în particular de specii, are o formă de grafic de funcție descrescătoare de tipul următor, în care există, respectiv nu există specii cu abundențe mijlocii:



Această formă provine din convenția de reprezentare și din faptul că, de regulă, într-o biocenoză numărul speciilor rare este mult mai mare decât cel al speciilor dominante prin abundențe.

Dacă, *in extremis*, numărul total de exemplare - 420 în exemplul nostru - s-ar distribui "echitabil" între toate speciile din biocenoză - 6 aici - s-ar obține distribuția următoare care, în ecologie, se numește *distribuție echitabilă* sau *distribuție regulată*. În statistică este denumită **distribuție uniformă**:



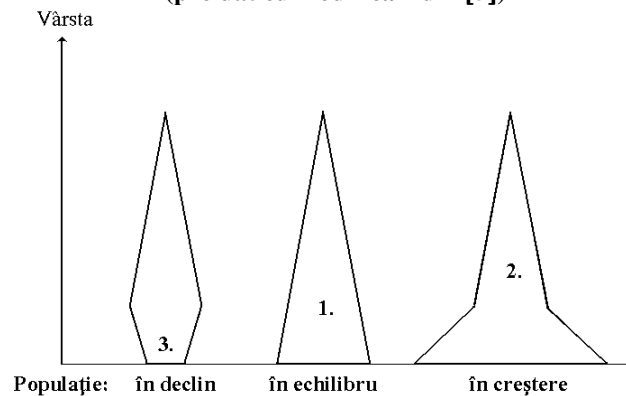
$$\begin{pmatrix} S_a & S_b & S_c & S_d & S_e & S_f \\ 70 & 70 & 70 & 70 & 70 & 70 \end{pmatrix}$$

+ 9^o Aplicație a histogramelor în biologia populațiilor – piramida vârstelor

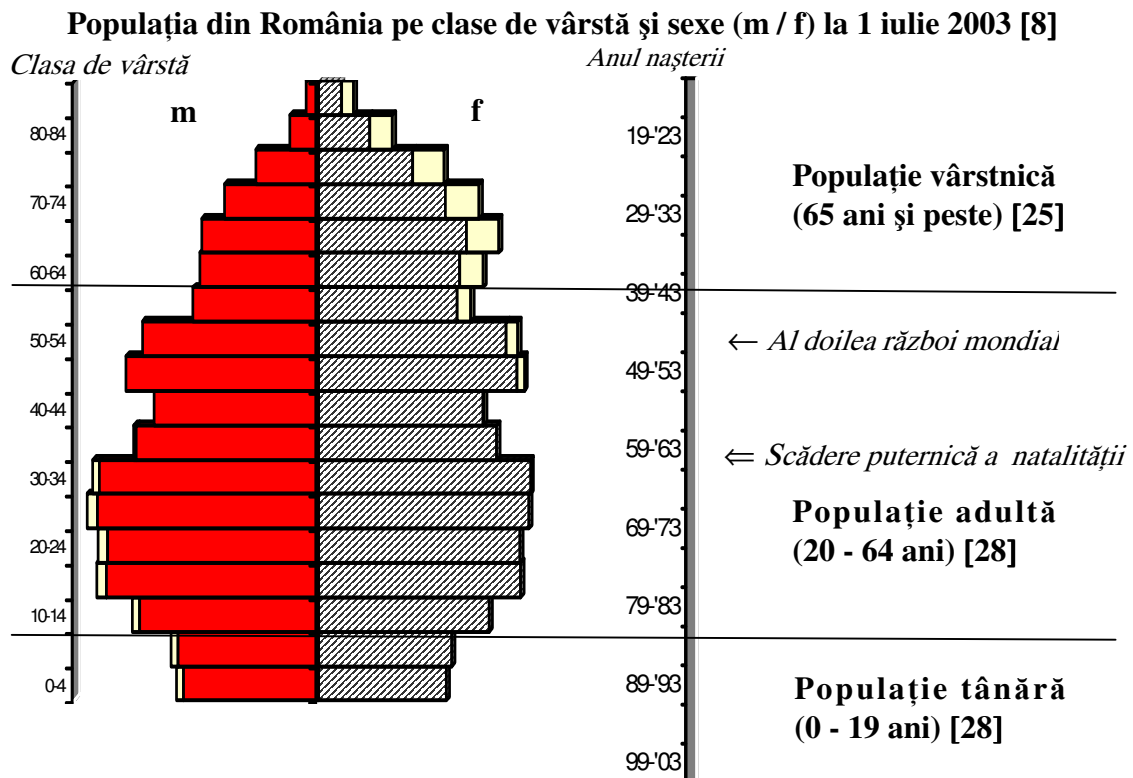
În biologia populațiilor și demografie volumele celor două sexe (*m / f*) pe vârste sau clase (grupe) de vârstă, dintr-o populație biologică animală, respectiv umană sunt reprezentate prin două histograme cu bazele – reprezentând vârsta - alipite pe verticală. Pentru că odată cu înaintarea în vârstă, din cauza mortalității, generațiile scad ca volum, reprezentarea are aspectul unei piramide, de unde și denumirea de **piramidă a vârstelor**.

1. Forma ideală de piramidă indică o populație în *echilibru staționar* ca volum total.
2. Lărgirea bazei piramidei indică creșterea volumului total prin mărirea proporției indivizilor tineri. Este o populație *în creștere* prin "întinerire".
3. Îngustarea bazei semnalează o populație *în declin*, în sensul scăderii volumului total – fenomen denumit și "creștere negativă" – prin reducerea proporției tinerilor și deci "îmbătrânirea populației".

Tipurile principale de piramide ale vârstelor
(preluat cu modificări din [5])



Într-o piramidă a vârstelor reală se pot evidenția și alte fenomene. Pentru a se facilita vizualizarea acestora putem adăuga la dreapta piramidei anii de naștere ai indivizilor din grupele de vârstă înscrise în stânga piramidei.

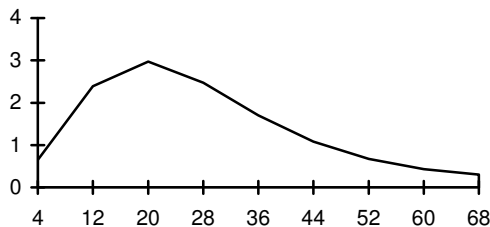


De exemplu, pentru populația din România se pot observa:

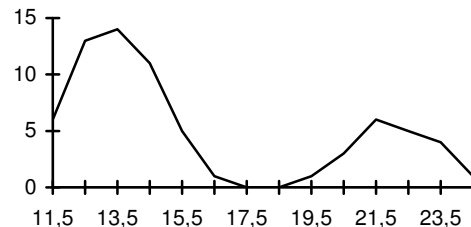
- *Excedentul* de bărbați sau de femei în cadrul fiecărei clase de vârstă, excedent evidențiat prin nuanța deschisă. Acesta este în favoarea sexului masculin la tineri și a celui feminin la vârstnici.
- Declinul numeric început aproximativ din anii '60 și accentuarea sa, în ultimul timp, prin *scăderea dramatică a ponderii populației tinere*, principalele cauze fiind "scăderea natalității și amplificarea migrației externe, îndeosebi în 1990-1992", la care se adaugă "creșterea mortalității" și "reculul nupțialității" din cadrul "crizei pe care o traversează țara" în această perioadă de tranziție [15].
- Efectul celui de-*al doilea război mondial* asupra generațiilor născute în jurul anilor '43-'47 (vezi săgeata simplă de mai sus), mai precis, generațiile '41-'45, ceea ce se observă exact pe o piramidă construită pe vârste [19]. Rezultatul de aici este ușor deformat din cauza grupării vârstelor în clase.
- *Scăderea puternică a natalității* în anii '58-'66 [19] (vezi dubla săgeată de mai sus). Clasele de vârstă indicate cu săgeți se numesc **intrânduri**.

3.1.2. Pentru ce grupăm măsurători sau "limbajul repartițiilor"²

Să urmărim formele următoarelor repartiții (distribuții) bazate pe date biologice reale în volum mare și să le asociem denumiri de caracterizare.



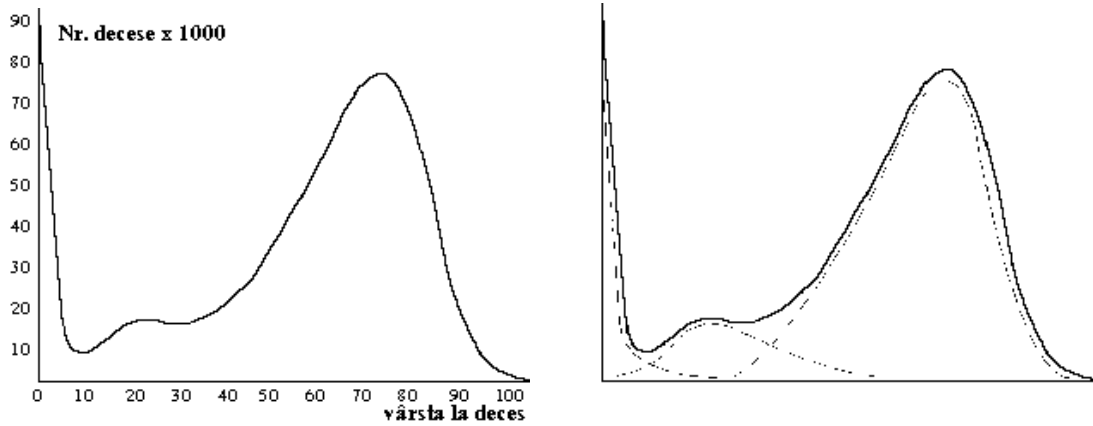
D1 – Rata fecundității specifică vârstei, în săptămâni, la *Microtus agrestis* [30]. **Distribuție unimodală** (slab asimetrică de stânga).



D2 - Talia indivizilor de *Nectophrynoides occidentalis* (clasa *Amphibia*) în luna septembrie [21]. **Distribuție bimodală**.

O distribuție se numește **unimodală** atunci când are o singură modă, respectiv **bimodală** atunci când are două mode, o modă fiind un punct de maxim local (detalii la 3.3.2.). O distribuție unimodală se numește **asimetrică de** (sau **la**) **stânga** atunci când are "capul" la stânga ("coada" fiind la dreapta).

² Expresia aparține lui V. Săhleanu [27].

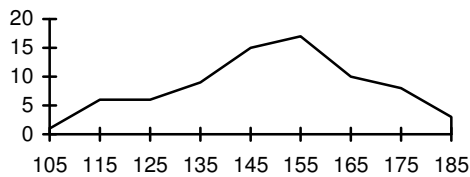


D2' – Distribuția numărului de decese pe vârste [32].

Distribuție multimodală.

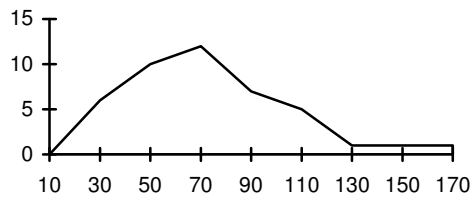
- ✓ O distribuție *bimodală*, respectiv o distribuție **multimodală** - adică o distribuție cu mai mult de două mode - pot fi considerate suma a două, respectiv mai multor distribuții unimodale. Spre exemplu, distribuția din stânga figurii de mai sus poate fi obținută prin suma a trei distribuții unimodale, ca în desenul din dreapta aceleiași figuri.

Continuăm seria exemplurilor de distribuții întâlnite în practică.



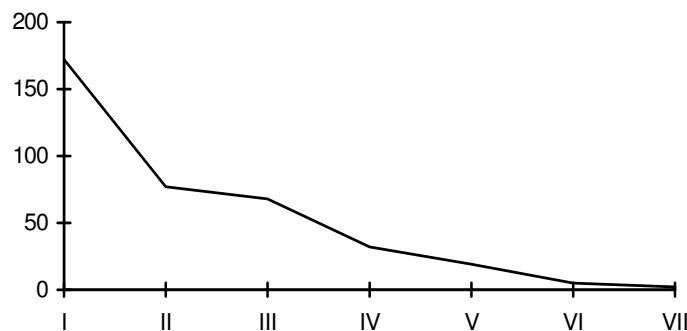
D3 - Frecvența indivizilor de *Cepaea nemoralis* cu diametre ale cochiliilor cuprinse între 104 și 185 mm [21].

Distribuție unimodală, slab asimetrică de dreapta.



D4 – Frecvența plantelor având între 10 și 170 flori per plantă [4].

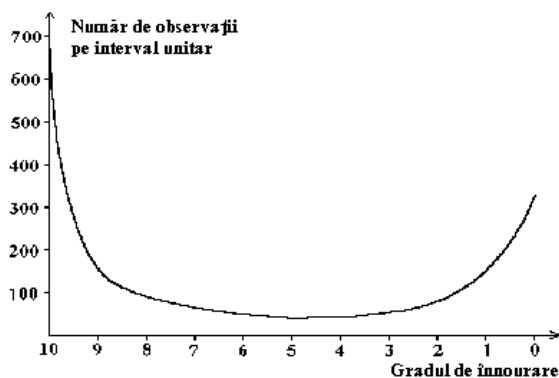
Distribuție puternic asimetrică de stânga.



D5 - Reprezentarea prin poligon al frecvențelor a distribuției de abundențe a grupelor sistematice ale fitoplanctonului românesc al Mării Negre (1972-1977) [5].

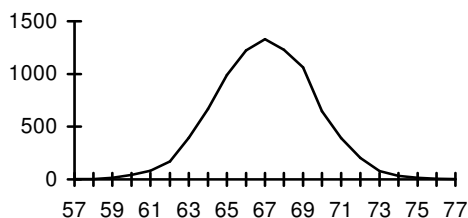
Distribuție extrem asimetrică de stânga (în formă de "l").

De regulă, distribuțiile de abundențe sunt distribuții în formă de "l".

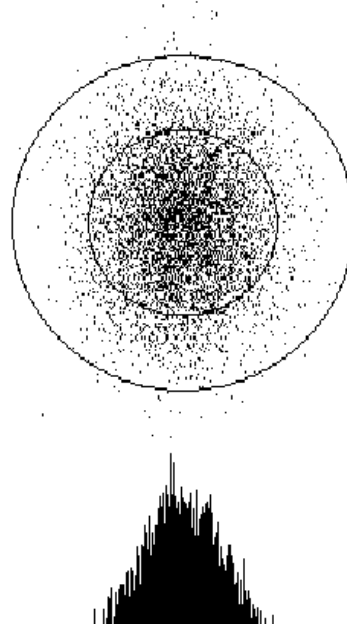


D6 - Distribuția gradului de înnoirare a cerului la Greenwich în iulie [32].
Distribuție bimodală în formă de "u".

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



D7 - Distribuția de frecvențe a înălțimii a 8585 bărbați adulți născuți în Insulele Britanice [32].
Distribuție unimodală și simetrică.



Prin cumulara gloanțelor "trase" la o țintă, pe abscisa la care au lovit ținta, se obține o **distribuție cvasinormală**.
(Simulare pe calculator.)



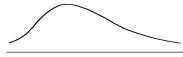
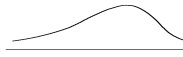
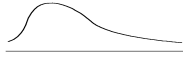
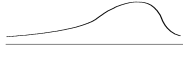


- ✓ Am lăsat intenționat la urmă forma de distribuție normală sau cvasinormală, pentru a atrage atenția asupra greșelii foarte răspândite de a "vedea" sau presupune această formă în spatele oricărui fenomen de masă. În paragraful 3.7. dedicat distribuției normale vom prezenta motivul secund pentru care distribuția normală este considerată o adevărată "stea polară" a statisticii bazate pe teoria probabilităților, iar în **volumul** de statistică inductivă vom pune în evidență motivul cel mai important.

Pornind de la studiul formelor acestor distribuții empirice se poate construi pentru distribuții (repartiții) empirice sau teoretice tipologia prezentată, în continuare, la itemul 2°. Tabelul prezintă sistematic această tipologie, precum și patru observații foarte importante, adăugate în ultima coloană, la începutul și sfârșitul tabelului.

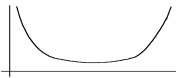
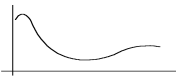
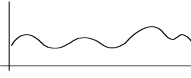

1° Concluzii generale

- ◆ Răspunsul la întrebarea generală "de ce grupăm" este:
"grupăm (fără sau cu pierdere de informație) pentru câștig în relevanță".
- ◆ Răspunsul la întrebarea mai tehnică "pentru ce grupăm" este:
"grupăm ca să sesizăm una din formele tip de mai sus".

2° Forme tip de distribuții

Unimodală (1 modă) Exprimă <u>omogenitate</u> .	simetrică	concentrată într-un punct (1) 	Exprimă <u>omogenitate</u> absolută.
		neconcentrată într-un punct (2) 	Exprimă cel mai bine o <u>tendință centrală</u> .
	[asimetrică]	slab asimetrică	de stânga (3) 
			de dreapta (4) 
		puternic asimetrică	de stânga (5) 
			de dreapta (6) 
		extrem asimetrică	de stânga (7) (în formă de i) 
			de dreapta (8) (în formă de j) 

Forme tip de distribuții (continuare)

Bimodală (2 mode)	simetrică (9) (de exemplu <u>în formă de u</u>)	Exprimă <u>eterogenitate</u> ca amestec de 2 <u>omogenități</u> diferite.
		
Multimodală (plurimodală)	asimetrică (10)	Exprimă <u>eterogenitate</u> ca amestec de n <u>omogenități</u> diferite ($n > 2$).
		
	multimodală propriu-zisă (11) ($n > 2$, mode)	
		
	uniformă (12) (numai mode – omnimodală)	Exprimă <u>eterogenitate</u> absolută.
		

3° Concluzii tehnice

- ◆ Modul în care tratăm fiecare formă tip derivă din două observații fundamentale, deja puse în discuție, drept comentarii ale tabelului anterior:
 - deoarece "nu putem alerga în același timp după doi sau mai mulți iepuri" vom trata eterogenitățile care apar ca un amestec de două sau mai multe omogenități, adică distribuțiile bimodale sau multimodale, vor fi *descompuse* eventual prin *decupare* în două, respectiv n distribuții unimodale;
 - deoarece ideea de tendință centrală este cel mai bine exprimată de distribuțiile unimodale simetrice, vom încerca să simetrizăm - prin *transformări (de simetrizare)* adecvate - orice distribuție asimetrică. Ne apropiem astfel de o distribuție normală. De aceea putem utiliza și sintagmele *transformări de cvasinormalizare* sau *de cvasigaussianizare*.
- ✓ Primul demers - *descompunerea*, în particular *decuparea* în distribuții unimodale - este absolut obligatoriu în cadrul statisticii descriptive, adică atunci când o serie este tratată drept populație statistică. Neaplicarea sa este, probabil, cea mai grosolană eroare statistică.
- ✓ Al doilea demers - *transformarea pentru simetrizare* - nu este strict obligatoriu în statistica descriptivă, însă este deosebit de productiv în statistica inductivă, după cum va rezulta la momentul potrivit.

4° Decuparea unei distribuții bimodale

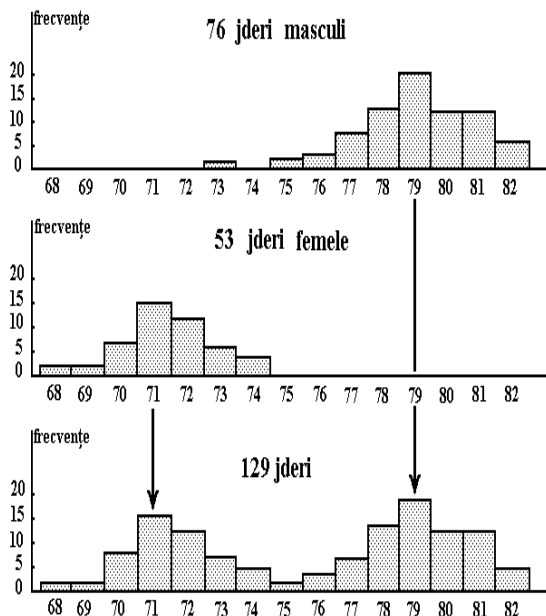
O primă imagine a faptului că o distribuție bimodală sau multimodală trebuie privită ca un cumul de distribuții unimodale o avem din exemplul distribuției D2' de mai sus. Detectarea distribuțiilor unimodale componente este însă o problemă statistică dificilă cu multe soluții și rezultate posibile. În continuare vom prezenta numai o variantă foarte simplă de decupare a unei distribuții bimodale. Presentăm această tehnică drept pregătire pentru înțelegerea viitoare

a trei probleme statistice majore: analiza de discriminare, construcția testelor (aplicabile în științele vieții) și filozofia testelor statistice (obligatorii pentru testarea ipotezelor științifice de specialitate).

Exemplul 3.1.2.

Să figurăm în același desen histogrammele corespunzătoare celor trei distribuții din tabelul alăturat. Acestea reprezintă frecvențele absolute ale lungimii craniilor de jderi masculi (M), femele (F) și total ambele sexe (T), animalele fiind capturate în 1955 în Montana [20].

Lungime (în mm)	Frecvență Masculi (M)	Frecvență Femele (F)	Frecvență total ambele sexe (T)
68	0	2	2
69	0	2	2
70	0	9	9
71	0	16	16
72	0	13	13
73	1	7	8
74	0	4	4
75	2	0	2
76	3	0	3
77	7	0	7
78	13	0	13
79	20	0	20
80	12	0	12
81	12	0	12
82	6	0	6
Totaluri :	76	53	129



"Se observă că distribuțiile de frecvențe ale celor două sexe sunt puternic decalate, moda distribuției masculilor fiind 79 mm (cu frecvența 20), iar moda distribuției femelelor fiind 71 mm (cu o frecvență de 16). Acest decalaj reflectă dimorfismul sexual marcat printr-o talie mai mare (și deci o lungime a craniului mai mare) a masculilor decât cea a femelelor în familia Mustelidae din care face parte jderul. Observăm că dacă se ignoră sexul și reprezentăm datele comasate se obține o distribuție bimodală, bimodalitatea (eterogenitatea) provenind tocmai din juxtapunerea celor două distribuții omogene." [20].

- ✓ În biologie este necesar ca datele morfologice să fie prelucrate separat pe cele două sexe, pentru a evita eterogenitatea provenită din dimorfism sexual.
- ✓ Bazat pe datele de mai sus putem decupa repartiția de frecvențe comasate pe valoarea 75 mm, aceasta convenind cel mai bine descompunerii acesteia în repartițiile pe cele două sexe. Obținem astfel un instrument, numit de unii autori *limită de discriminare* (sau *de identificare*, după alți autori) a sexului unui animal pentru care nu dispunem decât de craniul său după deces la maturitate.
- ✓ Acesta este un exemplu simplu de "*analiză de discriminare (analiză discriminantă)*" care este un tip aparte de prelucrare statistică.
- ✓ Se observă că discriminarea nu este perfectă, deoarece chiar pe datele care au folosit la construcția instrumentului de discriminare, masculul cu lungimea craniului de 73 mm este considerat, de către instrument, femelă.
- ✓ Este de așteptat ca pe volume mai mari de date să apară și erori inverse, adică femele considerate, de către instrument, masculi.
- ✓ Numărul erorilor de identificare de ambele tipuri (mascul considerat femelă și invers) va fi însă mult mai mic decât numărul cazurilor corect identificate, dacă cele două distribuții

unimodale pentru fiecare sex în parte, care formează distribuția bimodală comasată, nu se suprapun prea mult.

+ 5^o Logica de construire a unei transformări de simetrizare

Pentru a înțelege modul de determinare a transformărilor de simetrizare pentru fiecare formă tip de distribuție unimodală, să considerăm doar următorul exemplu.

Exemplul 3.1.2'.

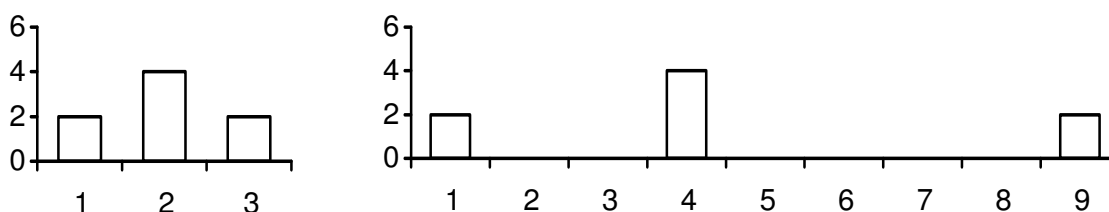
Se consideră mai multe parcele de pământ de formă pătrată, frecvențele acestora distribuindu-se simetric în raport cu lungimea laturii x_j conform primelor două coloane ale tabelului următor:

Tabele statistice simple pentru laturile, respectiv ariile unor parcele de pământ.

latura x_j	N_j	aria x_j^2
1	2	1
2	4	4
3	2	9

În ultima coloană a tabelului am adăugat ariile corespunzătoare parcelelor date.

Să figurăm cele două distribuții (a laturilor, respectiv a ariilor) sub formă de diagrame în batoane:



Se observă că distribuția ariilor nu mai este simetrică, ci are o asimetrie de stânga. Concluzia este imediată: dacă aplicăm distribuției asimetrice de stânga a ariilor transformarea inversă ridicării la pătrat, adică extragerea de rădăcină pătrată, vom obține distribuția laturilor care este simetrică.

Problemă propusă

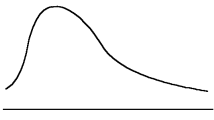
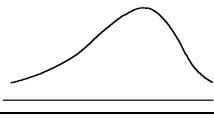
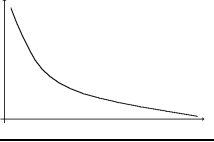
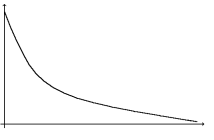
Să se considere în locul parcelelor pătrate din exemplul anterior volume cubice cu laturile în progresie aritmetică și frecvențele acestora distribuite simetric și să se figureze diagrama în batoane pentru distribuția volumelor cuburilor date. Să se observe gradul de asimetrie al distribuției volumelor.

O consecință practică a acestui exercițiu este următoarea observație: pentru populații omogene de oameni s-a constatat că talia are o distribuție aproape gaussiană (ca atare unimodală și simetrică). Deoarece greutatea depinde de talie ridicată la o putere cuprinsă între 2 și 3³, conform celor de mai sus, greutatea la aceeași populație se va distribui unimodal, dar cu asimetrie de stânga.

³ Greutatea corpului omenesc depinde, evident, de volumul acestuia. Volumul se poate calcula aproximând corpul printr-un cilindru, deci va fi proporțional cu produsul dintre înălțimea cilindrului (înălțimea persoanei) și pătratul diametrului de bază (adică un fel de lățime a persoanei, care este însă puternic dependentă de înălțime). Dacă dependența între lățime și înălțime ar fi totală, de exemplu fiind egale, volumul ar fi proporțional cu cubul înălțimii, ca în cazul unui cub sau cel al unei sfere. Dependența nefiind însă totală, se obține proporționalitatea volumului, respectiv a greutății, cu o putere a înălțimii cuprinsă între 2 și 3.

+6° Principalele transformări de simetrizare

Următorul tabel conține patru tipuri de distribuții, caracterizarea lor statistică, precum și transformările indicate pentru simetrizare.

Forma tip a distribuției grupate	Caracterizarea	Transformarea indicată
	Puternic asimetrică de stânga.	$\sqrt[n]{X}$
	Puternic asimetrică de dreapta.	X^n
	Extrem asimetrică de stânga care nu conține valoarea 0.	$\log_a X$ cu $a > 1$
	Extrem asimetrică de stânga conținând și valoarea 0.	$\log_a (X+b)$ cu $a > 1$ și $b > 0$.

3.1.3. Cum grupăm măsurători

Nu există și nu poate exista, în mod principial, o teorie matematică din care să rezulte modul de grupare.

Modalitățile de grupare nu pot fi alese decât de către biolog, ecolog, biochimist etc., care are o cunoaștere cu sens a materialului și un obiectiv specific. De aici rezultă obligativitatea cunoașterii de către aceștia a celor ce urmează, plus necesitatea unei practici statistice îndelungate cu date de specialitate.

Statistica pune la dispoziție doar unele **reguli empirice de grupare**:

- Grupăm doar serii cu volume ≥ 50 .
- Diverși autori indică diverse valori pentru numărul de *intervale de grupare* (denumite și **intervale de clasă**, sau *clase de grupare*, sau, cel mai general, *clase*): 20-40, 10-15, 8-20, 15-25, 8-15 etc.
- Se pot utiliza *intervale de grupare egale* sau *inegale*, după particularitățile datelor și interesul urmărit.

1° Grupare cu intervale de clasă egale

În cazul *intervalor de grupare egale* există unele formule empirice de calcul al numărului de clase (nc). Un exemplu este **formula lui Sturges**:

$$nc \approx 1 + 10 / 3 \cdot \lg N,$$
 unde $N =$ volumul seriei.

Valoarea nc se rotunjește la un număr întreg convenabil. **Lungimea intervalului de clasă** $ic = (x_{max} - x_{min}) / nc$

în care x_{max} , respectiv x_{min} sunt cea mai mare, respectiv cea mai mică valoare din serie.

Valoarea ic se rotunjește, de asemenea, convenabil.

Exemplul 3.1.3.

Fie următoarea distribuție negrupată de frecvențe reprezentând adâncimi ale stațiilor pentru prelevare de probe din lacul Babina - Delta Dunării. (Date ale Colectivului de Ecologie din perioada 1987-93.). Să se grupeze cu intervale de clasă egale.

Adâncimea în cm, x_j	Frecvența absolută N_j	Adâncimea în cm, x_j	Frecvența absolută N_j	Adâncimea în cm, x_j	Frecvența absolută N_j
95	1	150	7	190	4
100	4	153	1	198	1
105	1	155	3	200	3
110	3	157	1	208	1
120	4	160	7	210	4
125	4	163	1	211	1
130	4	167	1	220	2
134	1	170	2	240	3
135	2	175	2	257	1
140	4	180	3	290	1
147	1	185	1		
148	1	188	1		
				Total	$N = 81$

Rezolvare:

Volumul, $N = 81$ este mai mare ca 50, deci grupăm. Calculăm numărul de clase, nc , după formula lui Sturges:

$$nc = 1 + 10 / 3 \cdot \lg N = 1 + 10 / 3 \cdot \lg 81 \approx 1 + 10 / 3 \cdot 1,91 \approx 1 + 6,36 = 7,36.$$

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

Lungimea intervalului de clasă:

$$ic = (x_{max} - x_{min}) / nc = (290 - 95) / 8 = 195 / 8 = 24,375.$$

Rotunjind convenabil obținem valoarea $ic = 25$.

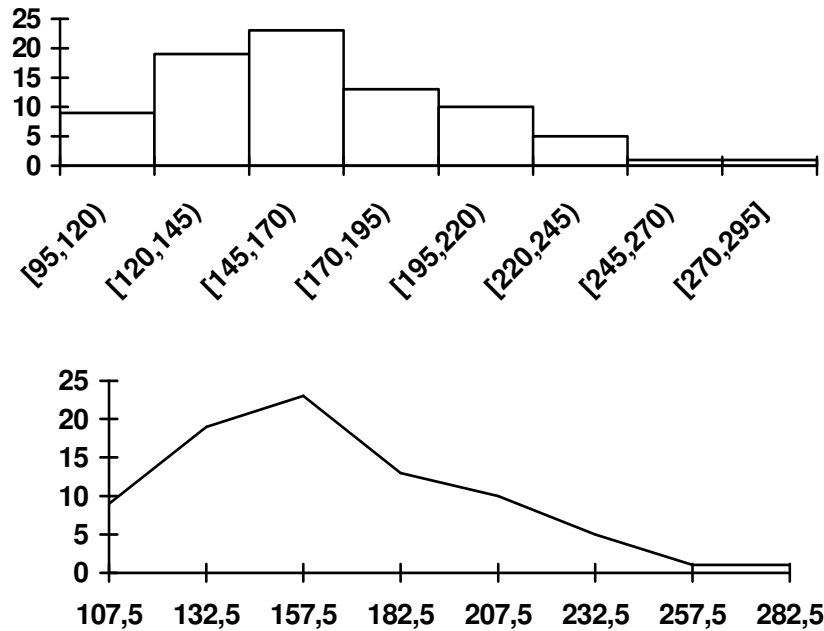
Pornim prima clasă din valoarea minimă $x_{min} = 95$. Se obțin astfel clasele din următoarea distribuție de frecvențe cu intervale de grupare egale (vezi prima coloană):

Intervalele de clasă $[x_j, x_{j+1})$	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
		Total $N = 81$

- ✓ Se observă că ultimul interval se consideră închis și la dreapta (pentru a nu pierde, niciodată, cea mai mare valoare din șir, deși aici nu este cazul).

Pentru *histogramă* utilizăm prima și ultima coloană. Dacă dorim însă *poligonul frecvențelor* pentru această distribuție grupată, se calculează coloana a II-a cu centrele intervalelor și se utilizează ultimele două coloane. *Centrele intervalelor* s-au plasat la mijlocul fiecărui interval de lungime 25, deci la distanță de 12,5 față de ambele extreme ale intervalului respectiv și, evident, la distanță de 25 față de centrele alăturate.

Histograma, respectiv poligonul frecvențelor se prezintă astfel:



Observăm "în spatele" acestei distribuții empirice o *distribuție unimodală, asimetrică de stânga*, ceea ce caracterizează corect distribuția tuturor adâncimilor lacului Babina: predomină adâncimi de circa 160 cm, urmează adâncimile mai mici din apropierea malurilor și există, mai rar, unele "gropi" de circa 2-3 m.

2° Reguli de rotunjire a datelor în calcule

Un mod de grupare cu intervale egale este și rotunjirea datelor. Prezentăm în continuare câteva *observații* asupra rotunjirii datelor în calculul manual î29ș.

a. Reguli de rotunjire în calculul manual:

- Cifrele 0,1,2,3,4 se șterg . De exemplu : 2,64 devine 2,6 .
- Cifrele 6,7,8,9 se rotunjesc prin adaos de o unitate la zecimala superioară. De exemplu 2,68 devine 2,7 .
- Cifra 5 se rotundește prin adaos, respectiv lipsă, la valoarea pară cea mai apropiată. De exemplu 2,65 devine 2,6 iar 2,55 va deveni de asemenea 2,6 . (Observăm că, în calculator, **cifra 5 este tratată întotdeauna prin adaos. Și noi vom proceda în continuare în acest mod.**)

b. Rotunjirea prin afectarea mai multor zecimale.

Numărul 1,959964 rotunjit:

la 5 zecimale = 1,95996

la 4 zecimale = 1,9600

la 3 zecimale = 1,960

la 2 zecimale = 1,96

la 0 zecimală = 2,0.

c. Numărul de zecimale păstrate în calculul manual:

- Se recomandă păstrarea a 2, 3, maximum 4 zecimale,
- În calculele intermediare putem păstra o zecimală în plus care dispare în rezultatul final.
- Precizia rezultatelor depinde doar de precizia măsurătorilor și nu de mărirea numărului de zecimale în cadrul calculelor.

✓ Gruparea cu intervale de clasă inegale este o problemă care depășește cadrul de față.