1.1. NOȚIUNILE DE BAZĂ ALE STATISTICII

- **Definiție**: știința colectării, organizării, prezentării, analizării și interpretării datelor
- **Importanță practică**: asistarea în luarea unor decizii eficiente
- **Scurt istoric** al statisticii românești
- **Tipuri de analiză statistică**: descriptivă și inferențială
- **Tipuri de variabile**: calitative și cantitative (discrete și continue)
- **Nivelurile de măsurare a datelor**: nominal, ordinal, interval și raport
- **Variabilitatea datelor medicale**

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O1-1: explice de ce cunoașterea noțiunilor statistice deține un rol important

O1-2: definească statistica și să ofere un exemplu practic de aplicare a acesteia

O1-3: facă diferența între statistica descriptivă și cea inferențială

O1-4: clasifice variabilele în calitative sau cantitative / discrete sau continue

O1-5: distingă între nivelurile de măsurare nominal, ordinal, interval și raport

O1-6: explice diferențele dintre tipurile de erori statistice

INTRODUCERE

Instrumentele statisticii sunt utilizate în diferite domenii de activitate printre care medicină, biologie, economie, educație, psihologie, agricultură, pentru a menționa doar câteva. Când datele analizate aparțin domeniilor Biologiei și Medicinei este utilizat termenul de BIOSTATISTICĂ, pentru a face o distincție clară a zonei de aplicabilitate a conceptelor, metodelor și instrumentelor statistice.

Biostatistica este știința care analizează și sintetizează datele statistice medicale și biologice, convertindu-le în informații valoroase, capabile să ghideze deciziile din domeniul sănătății. Fie că vorbim despre înțelegerea mecanismelor bolilor, evaluarea eficienței unui tratament sau anticiparea evoluției unor epidemii, biostatistica stă la baza acestor analize complexe. În acest curs, vom explora cum putem utiliza metodele statistice pentru a descifra regulile și tendințele existente în structura profundă a datelor biologice și medicale, oferind suportul necesar unor concluzii temeinic fundamentate, bazate pe seturi de date reale.

Biostatistica este un element fundamental în aplicarea metodei științifice în domeniul biologiei și al sănătății, oferind instrumentele necesare pentru a transforma informațiile și datele brute în cunoștințe utile pentru specialiștii din aceste zone de cercetare.

METODA ȘTIINȚIFICĂ este un proces sistematic de investigare a realității utilizat pentru a înțelege fenomenele naturale, a testa diferite ipoteze și a genera cunoștințe noi. Ea presupune o serie de PAȘI STRUCTURAȚI, care includ observarea unui fenomen, formularea unei ipoteze, efectuarea experimentelor sau colectarea datelor, analiza rezultatelor și, în final, formularea unor concluzii care pot confirma sau infirma ipoteza inițială de cercetare.

Principalii pași ai metodei științifice sunt:

1.OBSERVAREA: identificarea unui fenomen sau a grup de fenomene a căror observație generează o serie de întrebări la care se va putea răspunde într-un mod riguros științific (*de ex. este ușor de observat faptul că, în cadrul unei diete, persoanele care efectuează regulat și exerciții fizice înregistrează o scădere mai mare în greutate comparativ cu cele care sunt sedentare*).

2.FORMULAREA IPOTEZEI: în acest al doilea pas, se formulează o ipoteză în scopul explicării fenomenului studiat și pentru a realiza predicții asupra unui nou set de observații. Se pot formula astfel 2 tipuri de ipoteze:

- **IPOTEZA DE CERCETARE: „Exercițiile fizice par să reducă greutatea corporală”**
- **IPOTEZA STATISTICĂ: Reducerea medie a greutății corporale a persoanelor care efectuează exerciții fizice este mai mare comparativ cu reducerea medie a greutății corporale a persoanelor care nu fac exerciții fizice”**

3. **PROIECTAREA EXPERIMENTULUI**: se referă la planificarea și organizarea unui experiment într-o manieră care să asigure obținerea unor date valide și fiabile, permițând testarea corectă a ipotezelor formulate. Designul experimentului implică luarea unor decizii cheie legate de:
   - **Definirea variabilelor**: stabilirea variabilelor **independente** (factorii controlați sau manipulați) și a variabilelor **dependente** (cele care sunt măsurate ca răspuns la modificările variabilelor independente);
   - **Alocarea în mod aleator a participanților** sau unităților de studiu în grupuri care primesc tratamentul sau intervenția (**grup experimental**) și în grupuri care nu primesc tratamentul (**grup de control**), pentru a compara efectele. Aceasta ne permite, în situația când **toți ceilalți factori potențiali sunt controlați** (nu influențează rezultatele), să testăm o **relație cauzală**: în acest fel, vom putea concluziona sau nu că scăderea mai mare a greutății corporale în grupul experimental a fost rezultatul practicării exercițiilor fizice;
   - **Repetabilitate și validitate**: planificarea experimentului astfel încât acesta să poată fi **replicat de un număr mare de ori** până în momentul când va fi credibil din punct de vedere științific.

4. **ANALIZA DATELOR: Utilizarea tehnicilor statistice pentru a evalua rezultatele obținute (ulterior analizei statistice descriptive a variabilei: greutatea corporală a persoanelor din cele 2 grupuri, vom apela la analiza statistică inferențială, respectiv testarea ipotezelor privind diferența mediilor celor două populații, din care provin eșantioanele de lucru, prin utilizarea testului Student, în cazul exemplului de față).**

5. **CONCLUZIILE: Interpretarea rezultatelor și confirmarea sau infirmarea ipotezei de cercetare. În analiza statistică lucrăm cu două tipuri de ipoteze: H_0 „ipoteza nulă sau ipoteza absenței diferențelor", care va fi testată în scopul expres de a fi discreditată și H_A „ipoteza alternativă", care reprezintă o afirmație „diametral opusă ipotezei nule", pe care o vom considera adevărată doar în situația când datele analizate la nivel de eșantion aduc suficiente dovezi în sprijinul respingerii ipotezei nule H_0.**

6. **REPRODUCTIBILITATEA: Indiferent de concluziile procesului științific, foarte rar rezultatele unui singur studiu sunt considerate a fi concludente. Din acest motiv, este necesară testarea repetată de către alți cercetători, prin studii similare, pentru a confirma rezultatele și a asigura în acest fel respectivului subiect de cercetare o credibilitate mai mare din punct de vedere științific.**

DEFINIȚIE:

STATISTICA este o știință metodologică care studiază fenomenele și procesele colective din natură și din societate pe baza observării lor științifice, a culegerii și prelucrării detaliate a informațiilor statistice cu scopul de a formula și explica legile, legitățile și regularitățile care le guvernează.

Ceea ce caracterizează în mod specific obiectul statisticii este faptul că ea studiază fenomenele și procesele de masă sub formă de mulțimi sau colectivități, din natură și din societate, care s-au constituit ca rezultat al acțiunii unui mare număr de cauze cu caracter esențial asociate cu factori întâmplători, ale căror influențe depind de condițiile concrete de spațiu și timp în care se produc.

UNITĂȚILE STATISTICE

sunt ființe, lucruri, fapte sau evenimente elementare care fac obiectul observației statistice, indiferent de natura lor, cu condiția de a răspunde cerințelor unei definiții precise (totalitatea unităților statistice → VOLUMUL COLECTIVITĂȚII).

CARACTERISTICA UNITĂȚILOR STATISTICE

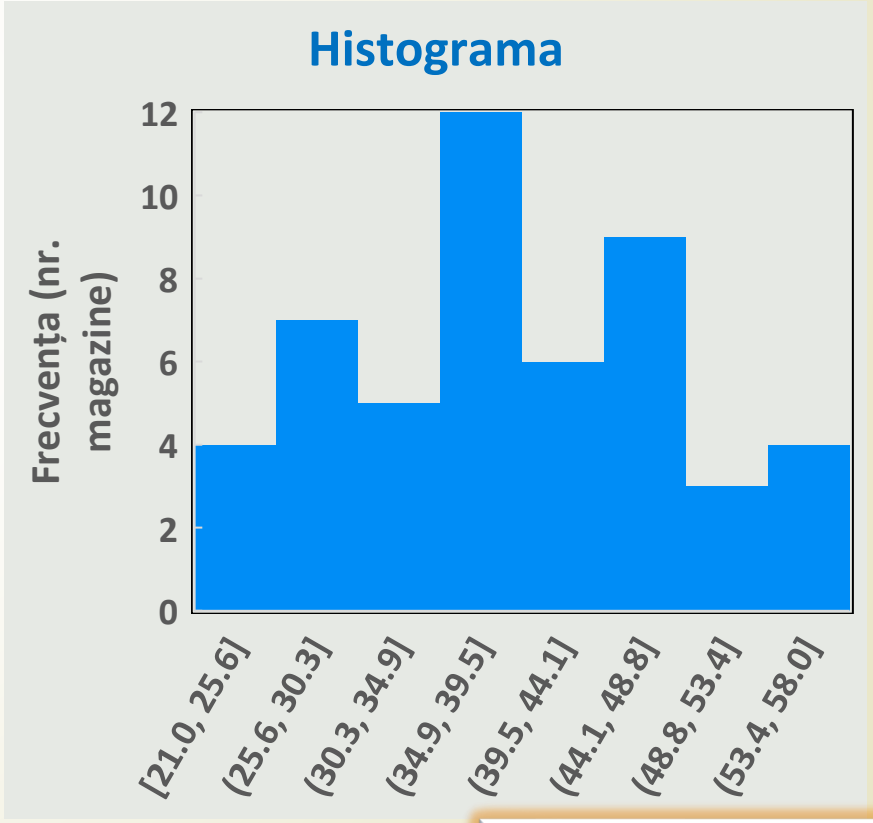
reprezintă însușirea, trăsătura sau proprietatea principală, comună tuturor unităților statistice ale unei colectivități, ale cărei valori diferă, în general, de la o unitate statistică la alta sau de la un grup de unități la altul (VARIABILA STATISTICĂ).

TIPURI DE ANALIZĂ STATISTICĂ DESCRIPTIVĂ ȘI INFERENȚIALĂ

STATISTICA DESCRIPTIVĂ reprezintă metodele de organizare, centralizare și prezentarea a datelor într-o formă cât mai sugestivă, ușor de înțeles și interpretat.

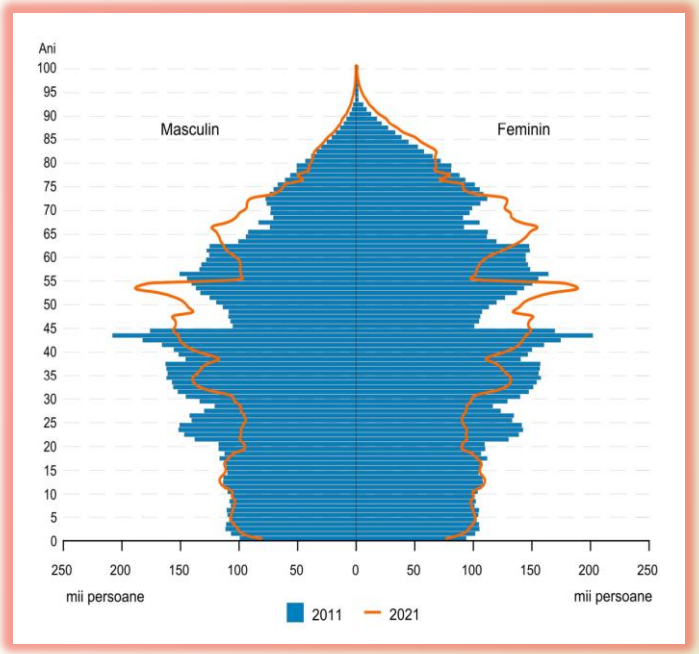
Farmacie	Cutii vândute
1	21
2	50
3	28
4	39
5	41
6	42
7	54
8	35
9	22
10	36
.....	
48	39
49	42
50	34

Cutii vândute	
Mean	38.660
Standard Error	1.330
Median	39.000
Mode	39.000
Standard Deviation	9.406
Sample Variance	88.474
Kurtosis	-0.575
Skewness	0.023
Range	37.000
Minimum	21
Maximum	58
Sum	1933
Count	50
Confidence Level(95,0%)	2.673



Un alt exemplu sugestiv de aplicare a metodelor statisticii descriptive pentru a centraliza volume foarte mari de date și a prezenta informațiile într-o formă accesibilă și interesantă pentru toate categoriile de utilizatori ai informației statistice se referă la prezentarea rezultatelor, RECENSĂMÂNTULUI POPULAȚIEI ȘI LOCUINȚELOR realizat în anul 2022:

Rezultatele RPL din decembrie 2021 arată o populație rezidentă a României de 19.053.815 persoane, în scădere cu 1,1 milioane locuitori față de recensământul precedent (octombrie 2011). Majoritatea populației rezidente este de sex feminin (9.808,3 mii, reprezentând 51,5%) și trăiește în mediul urban (9.941,2 mii, reprezentând 52,2%). Fenomenul de îmbătrânire demografică s-a accentuat, vârsta medie a populației rezidente crescând la 42,4 ani (față de 40,8 ani la RPL 2011). La recensământ, vârsta medie a populației de sex feminin este de 44,1 ani comparativ cu 40,6 ani pentru bărbați.



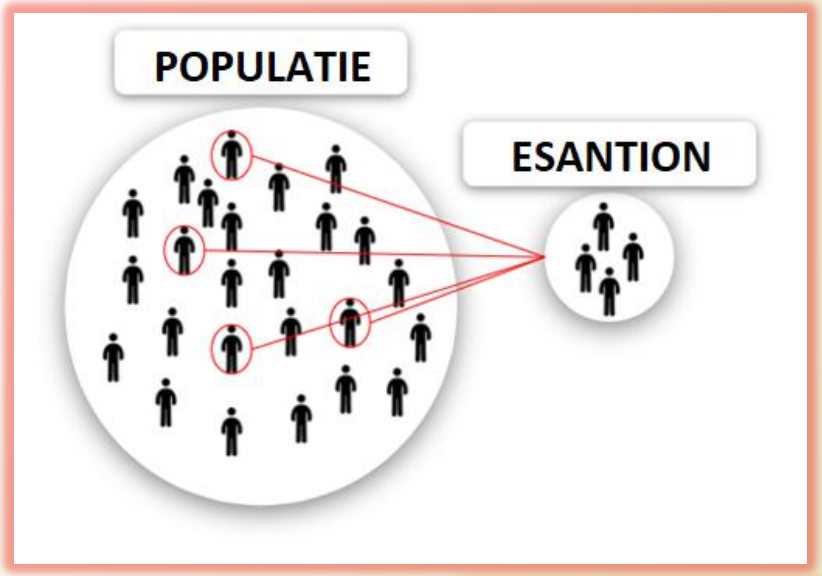
Și acum zece ani și în 2022, cel mai mic municipiu a fost și este Orșova din județul Mehedinți cu o populație de numai 8506 persoane, în scădere cu 1935 persoane față de RPL2011. Exceptând municipiul București, cel mai mare municipiu din România este Cluj-Napoca din județul Cluj, care și-a păstrat această poziție în ultimii zece ani deși și-a redus populația cu 37978 de persoane, coborând la o populație de 286598 locuitori.

STATISTICA INFERENȚIALĂ reprezintă metodele utilizate în vederea estimării unei proprietăți sau caracteristici a unei populații pe baza datelor obținute la nivelul unui eșantion reprezentativ extras din aceasta.

Există numeroase situații când deciziile trebuie luate pe baza unui set limitat de date. De exemplu, în situația în care am dori să măsurăm eficiența consumului de combustibil (l motorină / 100km) a tuturor autovehiculelor SUV utilizate în prezent. Această activitate de colectare a informațiilor de la întreaga „populație" a proprietarilor de SUV-uri ar necesita un consum extrem de mare de resurse de timp și financiare, fiind aproape imposibil de realizat în practică. Cu toate acestea, utilizând metode ale statisticii inferențiale, putem culege informațiile doar de la un număr limitat de deținători de vehicule SUV, utilizând un „eșantion reprezentativ " extras din populația generală.

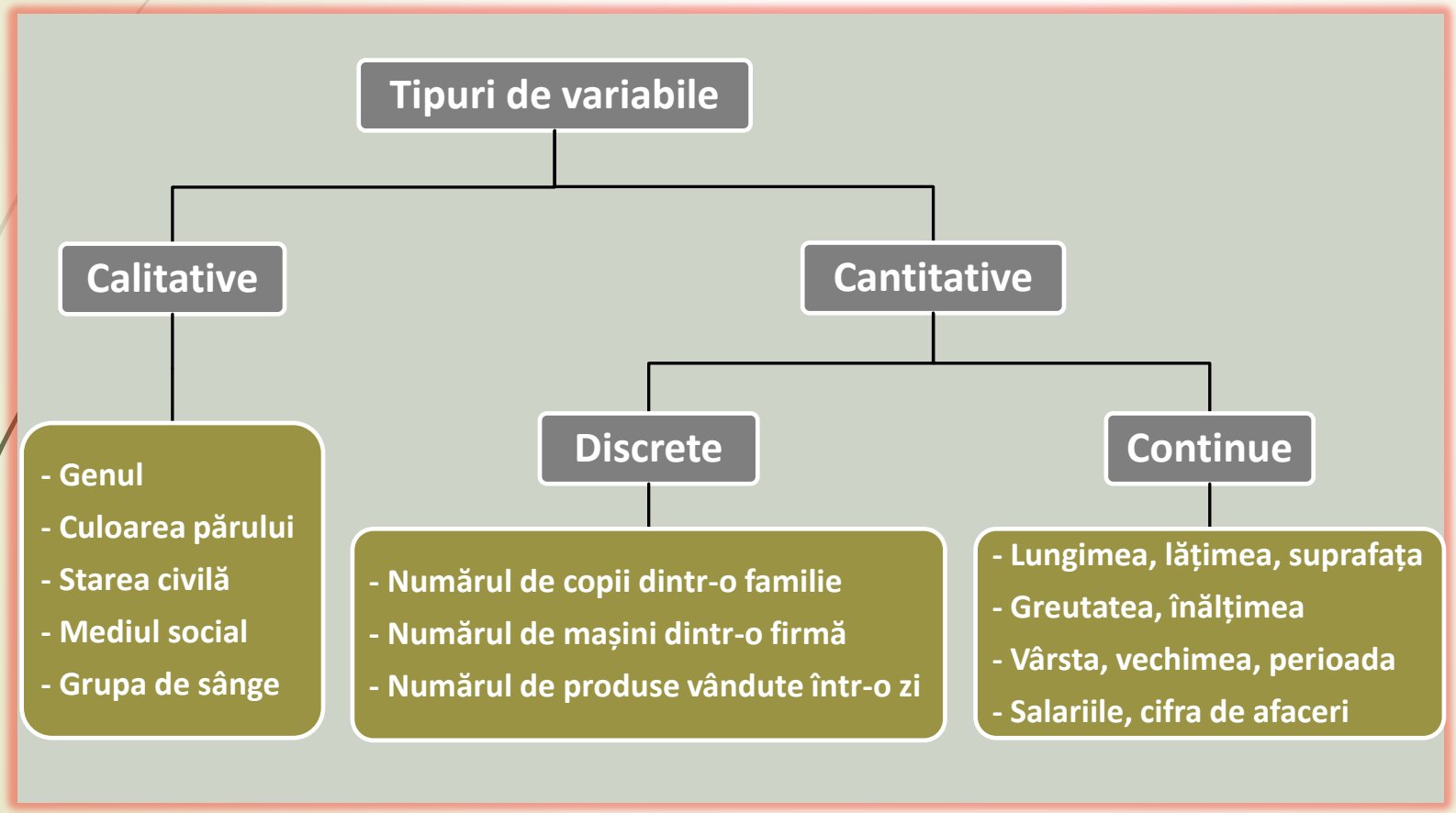
POPULAȚIA reprezintă întregul set de indivizi, obiecte sau fenomene de interes împreună cu caracteristicile lor care fac obiectul unei cercetări statistice. Un indicator statistic de sinteză calculat la nivelul unei populații → **PARAMETRU**

EȘANTIONUL reprezintă doar o mică parte a populației generale de interes, care reproduce într-un mod cât mai fidel trăsăturile acesteia *(o reproducere miniaturală a populației)*. Un indicator statistic de sinteză calculat la nivelul unui eșantion → **STATISTICĂ**



TIPURI DE VARIABILE STATISTICE
CALITATIVE ȘI CANTITATIVE

VARIABILA STATISTICĂ este acea caracteristică sau proprietate a unui individ, obiect sau fenomen, de interes pentru o cercetare statistică, care este măsurabilă și poate lua diferite valori.

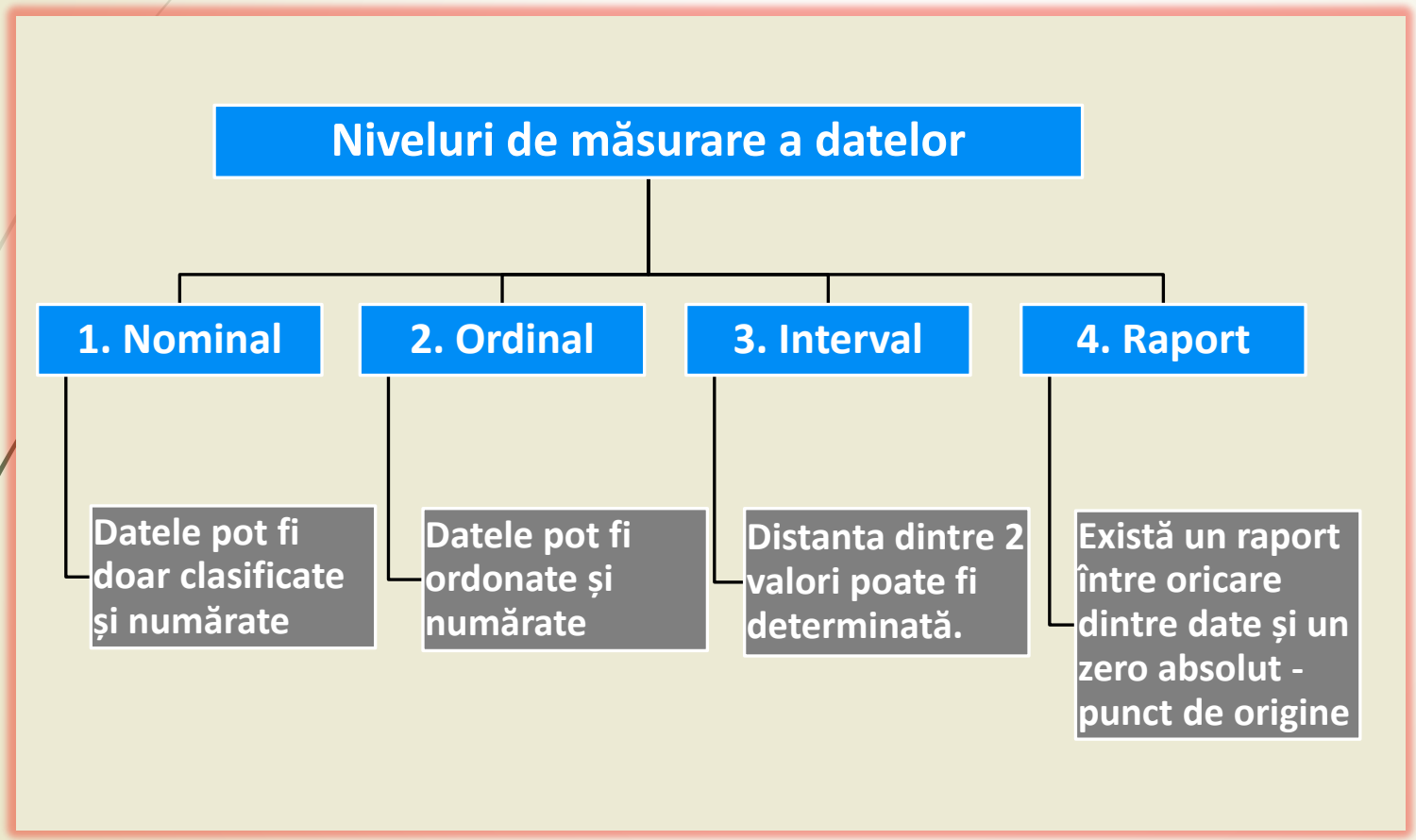


În cazul variabilelor calitative (categoriale), de exemplu starea civilă, vom determina ponderile (%) populației căsătorite, necăsătorite, divorțate și văduve.

Variabilele cantitative sunt cele care se pot raporta numeric și pot fi discrete (iau doar anumite valori, între care există spații goale în care nu există nicio altă valoare posibilă) sau continue (pot lua absolut orice valoare în cadrul unui anumit interval).

NIVELURI DE MĂSURARE A DATELOR

Determină modalitatea în care datele pot fi organizate, centralizate și prezentate și în același timp și tipul de analiză statistică care va fi utilizat în prelucrarea ulterioară a lor.



1. Datele nominale: specifice observațiilor referitoare la variabilele calitative, datele statistice nominale se pot împărți în diferite categorii, **fără să existe o ordine pentru acestea**. Categoriile pot fi doar clasificate și numărate, fiind prezentate ca valori absolute pentru fiecare categorie sau procente din total.

Putem exemplifica nivelul nominal al datelor statistice prin distribuția populației rezidente a României după variabila „stare civilă” la recensământul populației și locuințelor din 2021. Vom realiza o clasificare a nivelurilor variabilei „stare civilă” , fără ca între acestea să putem defini o ordine evidentă (ordonarea celor patru categorii este nerelevantă în acest caz).

Starea civilă	Persoane	% din total
Căsătorit	9.125.111	47,9%
Necăsătorit	7.496.849	39,3%
Divorțat	1.439.139	7,6%
Văduv	992.716	5,2%
Total	19.053.815	100,0%

2. Datele de tip ordinal sunt specifice tot variabilelor statistice calitative, cu precizarea suplimentară că acestea se pot împărți în diferite categorii, care însă pot fi ordonate pe o scală ordinală (de exemplu: insuficient, suficient, mediu, bine, foarte bine).

De ex., utilizăm distribuția cu date fictive a calificativelor celor 100 de studenți din anul I la disciplina Statistică. Datele sunt centralizate în 5 categorii, ordonate în funcție de calificativul obținut pentru prezentarea proiectului. Categoriile reflectă calitatea nivelului de pregătire a studenților, însă nu putem afirma că diferența între studenții care au obținut „foarte bine” și „bine” ar fi la fel comparativ cu cea între cei care au obținut „mediu” și „slab”.

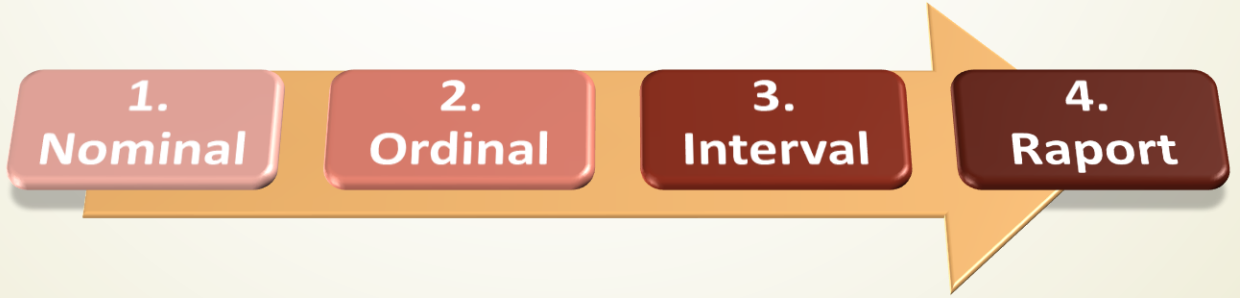
Calificativ	Nr. studenți	% din total
Foarte bine	8	8,0%
Bine	26	26,0%
Mediu	42	42,0%
Slab	19	19,0%
Foarte slab	5	5,0%
Total	100	100.0%

3. Datele de tip interval reprezintă al treilea nivel ierarhic de măsurare, care include toate caracteristicile nivelului anterior (ordinal), iar în plus intervalul sau diferența dintre oricare 2 valori alăturate este aceeași. De asemenea, **prezența lui zero este arbitrară**, reprezentând doar un punct pe scară, care nu semnifică absența condiției. Ne putem gândi la scara de temperatură Celsius, în care diferitele niveluri ale temperaturii pot fi ordonate crescător sau descrescător și putem calcula orice interval între două niveluri ale temperaturii (1 grad Celsius reprezintă o unitate de măsură constantă). Astfel distanța dintre 5 și 10 grade Celsius este egală cu cea dintre 40 și 45 de grade, respectiv 5 grade Celsius. Valoarea zero reprezintă doar un punct pe scara de temperatură, care nu arată absența oricărui nivel al acesteia (căldură sau frig).

4. Datele de tip raport reprezintă cel mai complex nivel de măsurare a variabilelor statistice , având toate proprietățile datelor de tip interval și în plus un „zero absolut”, tratat ca punct de origine. **Banii** sunt un exemplu foarte potrivit. În situația când o persoană are zero lei, înseamnă că nu are niciun ban asupra sa și în același timp, atunci când un salariat câștigă 10000 lei / lună, acest nivel reprezintă dublu comparativ cu alt salariat care câștigă doar 5000 lei / lună. De asemenea, **greutatea** sau **înălțimea** sunt măsurate pe o scară de tip raport, nivelul zero exprimând absența oricărei valori a greutății sau înălțimii. În același timp, o persoană de 150 kg cântărește triplu față de una de 50kg.

Sintetizând, principalele caracteristici ale nivelurilor de măsurare a unei variabile statistice sunt evidențiate după cum urmează:

Nivelul de măsurare a datelor statistice	Nominal	Ordinal	Interval	Raport
Este stabilită ordinea variantelor variabilei?	-	DA	DA	DA
Diferența dintre variante poate fi evaluată?	-	-	DA	DA
Adunarea și scăderea variantelor variabilei?	-	-	DA	DA
Înmulțirea și împărțirea variantelor variabilei?	-	-	-	DA
Există un zero absolut (punct de origine)?	-	-	-	DA



VARIABILITATEA DATELOR MEDICALE

Cercetarea biologică se bazează pe rezultate obținute pe un număr limitat de observații (eșantion) din multiplele posibile (populația), este deci o cercetare prin sondaj. Scopul final al acestei cercetări este reprezentat de generalizarea observațiilor realizate pe un număr limitat de cazuri, la întreaga colectivitate studiată, obținând astfel legi cu aplicabilitate generală.

Acest proces de generalizare depinde de două caracteristici ale datelor statistice:

- ✓ **numărul observațiilor studiate (volumul eșantionului sau lotului analizat);**
- ✓ **împrăștierea sau dispersia acestora (varianța).**

Prin natura lor, datele medicale prezintă o variație intrinsecă, biologică care implică un studiu specific având la bază teoria probabilităților. Pe lângă această variație, există și modificări ale valorilor reale măsurate, datorate erorilor generate de metoda metrologică aplicată cât și impreciziei observatorilor.

Variația biologică specifică datelor medicale prezintă interes în procesul de cercetare științifică, în timp ce restul variațiilor reprezintă erori, care trebuie minimizate.

Variațiile datorate erorilor de măsură sunt caracterizate de inexactitate și de precizie.

INEXACTITATEA se datorează **incapacității de a măsura perfect** o anumită mărime. Aceasta nu se datorează erorii aparatului de măsură ci depinde de factori perturbatori (modificări de temperatură, ale câmpului electric, magnetic, etc.).

EXACTITATEA reprezintă apropierea valorii numerice determinate experimental de valoarea adevărată (este de fapt **eroarea absolută**).

PRECIZIA se referă la **fidelitatea măsurătorii**. Ea depinde de sensibilitatea aparatului de măsură. Se spune despre o metodă că este precisă când rezultatele determinărilor sunt **reproductibile**, adică sunt apropiate ca valoare în contextul unor experimente repetate. Precizia se măsoară în **numărul de zecimale corecte** pe care le obținem printr-o anumită măsurătoare.

EROARE MICĂ

- **PRECIZIE RIDICATĂ**: împrăștierea rezultatelor individuale față de medie este mică
- **EXACTITATE BUNĂ**: media rezultatelor este apropiată de valoarea reală

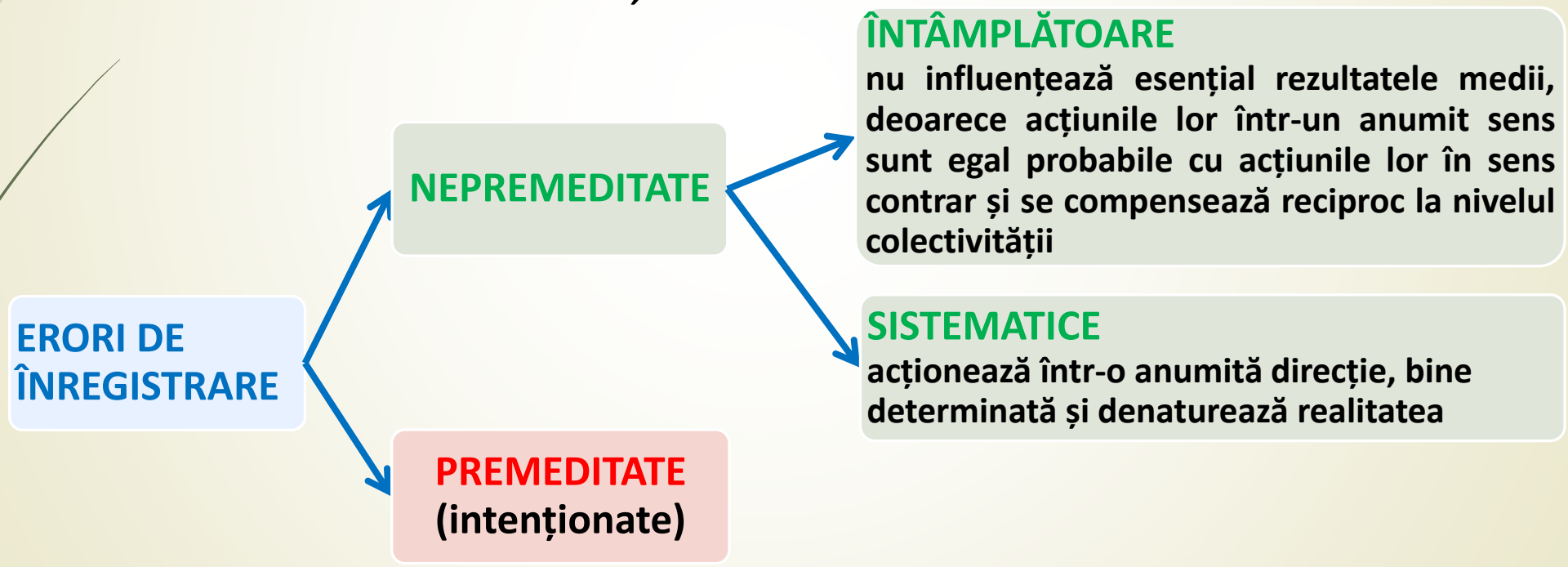
EROARE MARE

- **PRECIZIE SLABĂ**: rezultatele sunt mult dispersate față de valoarea medie
- **EXACTITATE SCĂZUTĂ**: media rezultatelor se îndepărtează mult față de valoarea reală

În urma înregistrării informațiilor statistice cu privire la fenomenele colective pot să apară *diferențe* între nivelurile înregistrate ale unor anumite caracteristici și nivelurile lor reale, *diferențe* care poartă denumirea de ERORI DE ÎNREGISTRARE.

Cauzele producerii erorilor de înregistrare sunt numeroase:

- ***lipsa de experiență a personalului care înregistrează;***
- ***ținerea inexactă a evidențelor care stau la baza înregistrărilor;***
- ***erorile de măsurare, etc.***



În plus față de noțiunile de statistică, sunt necesare și abilități de utilizare a software-ului de analiză statistică în scopul de a analiza, organiza, centraliza și prezenta într-un mod corespunzător rezultatele analizei statistice.

Printre cele mai populare aplicații software de analiză statistică a datelor se numără Microsoft EXCEL, IBM SPSS, Statistica (TIBCO Software Hamburg, Germany), JASP (Jeffreys's Amazing Statistics Program – dezvoltat de un grup de cercetători de la Universitatea din Amsterdam), jamovi (dezvoltat de Jonathon Love, Damian Dropmann si Ravi Selker, Sydney, Australia) .

Un PROGRAM DE ANALIZĂ STATISTICĂ a datelor cuprinde mai multe module, printre care:

- **Statistica descriptivă**
- **Verificarea integrității datelor**
- **Verificarea presupunerilor de normalitate a distribuțiilor și de omogenitate a variațelor**
- **Intervalele de încredere pentru estimarea parametrilor populației**
- **Compararea a 2 grupuri independente (testul t, Mann-Witney U) / dependente (testul t, Wilcoxon)**
- **Compararea mai multor grupuri independente (ANOVA, Kruskal-Wallis) / dependente (RMANOVA, Friedman)**
- **Analiza de regresie și corelație (simplă / multiplă)**
- **Testul HI pătrat de asociere a două variabile calitative**

1.2. PREZENTAREA INFORMAȚIILOR STATISTICE

- Tabelele de frecvențe și reprezentarea grafică a **VARIABILELOR CALITATIVE** – diagramele prin bare și graficele de structură
- Construirea distribuțiilor de frecvențe și a graficelor statistice specifice **VARIABILELOR CANTITATIVE** – histograma și poligonul frecvențelor
- Descrierea **RELAȚIEI SAU ASOCIERII** dintre **două variabile cantitative** (diagrama tip nor de puncte), **calitative** (tabelele de contingență și graficul tip stivă de bare) sau **o variabilă cantitativă și una calitativă** (piramida vârstelor)

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O2-1: centralizeze variabilele calitative prin realizarea tabelelor de frecvențe (absolute și relative)

O2-2: să reprezinte grafic tabelele de frecvențe (graficele de structură sau diagramele prin bare)

O2-3: centralizeze variabilele cantitative prin construirea distribuțiilor de frecvențe (absolute și relative)

O2-4: reprezinte grafic o distribuție de frecvențe (histograma și poligonul frecvențelor)

O2-5: să creeze și să explice diagramele tip nor de puncte și tabelele de contingență

INTRODUCERE

În acest capitol vor fi introduse o serie de proceduri de bază utilizate în analiza descriptivă a datelor statistice. În general, cu ocazia diferitelor studii și cercetări se colectează un volum însemnat de informații privind caracteristicile unităților statistice de interes, care nu pot fi înțelese în profunzime doar prin simpla lor observație.

Informațiile obținute în urma înregistrării fiecărei unități a colectivității cercetate cu caracteristicile ei constituie materialul statistic inițial, "materia primă" a statisticii, pe baza căreia se înfăptuiește cunoașterea științifică a realității.

După obținerea informațiilor statistice și verificarea autenticității lor (prevenirea erorilor de înregistrare și controlul cantitativ și calitativ al datelor statistice), gruparea statistică sub forma unor tabele sau reprezentări grafice adecvate facilitează într-o primă etapă conturarea unei imagini mai clare și înțelegerea mai bună a esenței fenomenelor medicale, sociale sau economice studiate.

Nr.crt	VARSTA	SEX	GRUPA DE SANGE	RH	CELULE ALBE	HTA	GLICEMIE	COLESTEROL	DIABET
1	62	0	2	1	14720	0	159	180	1
2	62	0	2	1	4710	0	85	145	0
3	54	0	3	1	6440	1	154	238	1
4	38	0	1	1	10370	1	153	200	1
5	57	1	1	1	6530	1	156	257	1
6	40	0	1	1	11380	0	136	170	1
7	23	1	1	1	8690	0	154	195	1
8	48	1	4	1	14320	0	101	182	0
9	59	1	2	0	7840	1	105	184	0
10	37	1	3	0	6430	0	80	172	0
...	...	...	...	...	...	...	...	...	...
99	56	1	3	0	9395	0	117	221	1
100	52	1	3	0	11760	1	94	177	0

REALIZAREA TABELELOR DE FRECVENȚE

Așa cum am prezentat în capitolul 1, metodele utilizate în descrierea unui set de date aparțin statisticii descriptive. Aceasta organizează datele statistice pentru a surprinde trăsăturile esențiale ale lor, pentru a descoperi tendința de concentrare a datelor și a identifica valorile extreme, neobișnuite ale variantelor pe care le poate lua o variabilă statistică.

TABELUL DE FRECVENȚE se realizează prin gruparea tuturor variantelor unei variabile calitative în mai multe categorii (clase) distincte, evidențiind numărul de observații aferent fiecărei categorii în parte.

Grupa de sange a pacientilor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Grupa 0	31	31.0	31.0	31.0
	Grupa A	25	25.0	25.0	56.0
	Grupa B	36	36.0	36.0	92.0
	Grupa AB	8	8.0	8.0	100.0
	Total	100	100.0	100.0	

În setul de date prezentat anterior, sunt evidențiate nouă variabile statistice (vârsta pacienților, sexul, grupa de sânge, RH-ul, numărul de celule albe, prezența HTA, glicemia, colesterolul, prezența diabetului). Variabilele statistice 2,3,4,6 și 9 sunt calitative, iar celelalte patru sunt exprimate cantitativ, numeric.

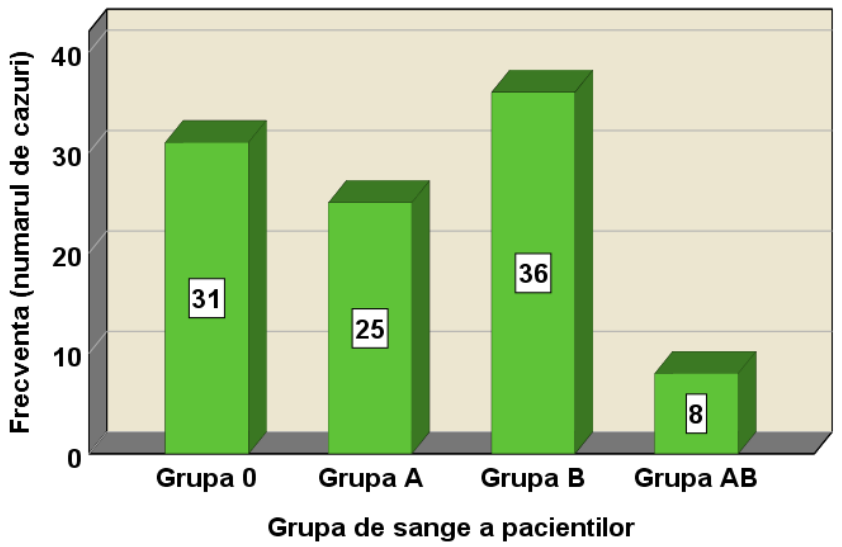
Vom realiza tabelul frecvențelor absolute pentru cei 100 de pacienți, pentru variabila statistică "grupa de sânge". Acest lucru se poate realiza foarte ușor utilizând modulul „Pivot Table" din EXCEL.

Pentru a obține **frecvențele relative** ale fiecărei categorii (clase), se raportează frecvența absolută a fiecăreia (numărul de pacienți pentru fiecare grupă de sânge) la numărul total de observații (numărul tuturor al pacienților din eșantion). Frecvența relativă conturează legătura fiecărei clase de frecvență cu numărul total de observații (ponderea sau procentul de acestea din total).

REPREZENTAREA GRAFICĂ A VARIABILELOR CALITATIVE

DIAGRAMELE PRIN BARE: tipurile calitative (categoriile) sunt prezentate pe axa orizontală, iar clasele de frecvențe pe axa verticală, fiind proporționale cu înălțimea fiecărei bare.

Variabila „grupa de sânge” este măsurată pe o scară nominală, deci ordinea celor 4 grupe de sânge pe axa orizontală nu este importantă. Înălțimea barelor corespunde numărului de pacienți care au o anumită grupă de sânge.



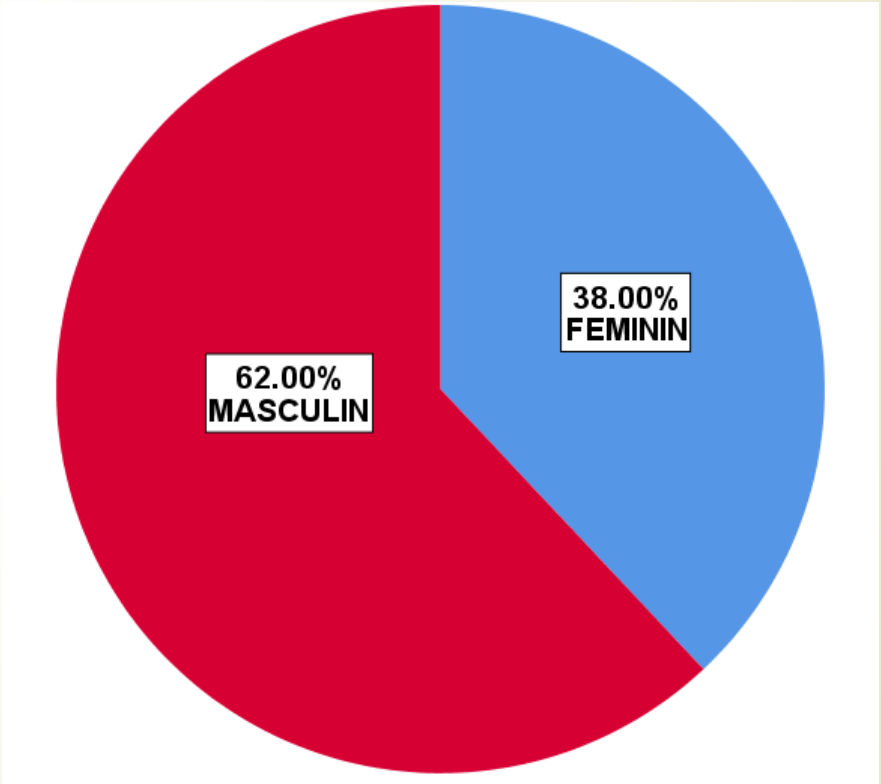
DIAGRAMELE DE STRUCTURĂ: evidențiază proporția (procentul) pe care fiecare categorie îl deține în totalul colectivității studiate (suma frecvențelor absolute).

Pentru a realiza în mod operativ diagrama de structură pentru pacienți în funcție de variabila statistică "sex" (Masculin/Feminin) vom utiliza modulul „Pivot Chart" din EXCEL, selectând de data aceasta pe rânduri, în loc de grupa de sânge cum a fost cazul anterior, sexul pacienților. Vor fi generate automat atât tabelul cu frecvențele absolute și cele relative (procentuale) pentru fiecare tip de vehicul cât și graficul tip „Pie Chart" corespunzător.

Deoarece, sectoarele de cerc reprezintă frecvența relativă a fiecărui sex (M/F), ca procent din numărul total al pacienților, se poate realiza cu ușurință o comparație între ponderile deținute de cele două sexe.

Sexul fiecarui pacient

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FEMININ	38	38.0	38.0	38.0
	MASCULIN	62	62.0	62.0	100.0
	Total	100	100.0	100.0	



CONSTRUIREA DISTRIBUȚIILOR DE FRECVENȚE

Datele prezentate anterior includ și 4 variabile cantitative (numerice): vârsta pacienților, numărul celulelor albe, glicemia și colesterolul. În situația în care se dorește o centralizare a pacienților în funcție de vârsta fiecăruia, avem nevoie să construim o distribuție a frecvențelor.

DISTRIBUȚIA DE FRECVENȚE: gruparea tuturor variantelor unei variabile cantitative în intervale distincte, evidențiind numărul de observații sau variante din fiecare clasă sau interval.

În urma centralizării datelor privind variabila cantitativă „VÂRSTĂ” vom putea răspunde la următoarele întrebări: Care ar fi vârstă tipică a pacienților? Care este cea mai mare / mica vârstă a unui pacient din eșantion? În jurul cărei valori vârstele pacienților tind să se grupeze?

62	49	64	36	40	34	32	47	49	31
62	58	38	55	46	70	58	73	60	51
54	68	25	47	52	69	68	75	51	46
38	50	42	41	41	48	51	42	60	35
57	54	31	59	53	70	32	55	54	22
40	36	40	54	63	28	41	17	27	49
23	29	45	23	28	35	74	37	26	58
48	39	55	22	31	33	59	18	53	41
59	61	64	58	44	48	25	48	49	56
37	64	39	46	55	49	32	43	44	52

Datele prezentate în forma inițială sunt mult mai ușor de interpretat dacă realizăm gruparea și centralizarea acestora prin construirea unei distribuții de frecvențe, parcurgând următoarele etape:

a) Stabilirea numărului de intervale (k): este necesar să determinăm numărul de clase de grupare, utilizând *regula* 2^k → trebuie identificată cea mai mică valoare a lui k pentru care $2^k > n$ (numărul de observații). În cazul exemplului nostru $2^7 = 128 > 100$. Deoarece $2^6 = 64 < 100$, vom lua în considerare valoarea k = 7 intervale de grupare.

b) Determinarea mărimii intervalului de grupare (l):

$$l \geq \frac{Valoarea\ maxima - Valoarea\ minima}{k} = \frac{75 - 17}{7}$$

$$\approx 8.29\ ani$$

În practică, valoarea obținută se rotunjește convenabil (multiplu de 10 sau 100), astfel încât vom considera mărimea intervalului de 10 ani.

c) Precizarea exactă a limitelor fiecărui interval pentru ca fiecare observație să aparțină unei singure clase, evitând astfel suprapunerile.

Vârsta	Frecvența
10 - 20	2
20 - 30	11
30 - 40	18
40 - 50	27
50 - 60	25
60 - 70	12
70 - 80	5
TOTAL	100

Vârsta pacienților este cuprinsa intre 17 și 77 ani, fiind grupată pe intervale egale de mărime 10 ani. Vârstele sunt concentrate în zona 30 - 60 ani (pentru aproape 70% dintre pacienți).

Pentru fiecare interval, putem determina o **vârstă tipică** situată la **mijlocul fiecărui interval** de grupare, la jumătatea distanței dintre limitele superioară și inferioară. Punctul central al fiecărei clase (interval) aproximează vârstele tuturor pacienților din acel interval, fiind astfel o valoare de referință a acestora.

Vârsta celor mai mulți pacienți s-a situat in intervalul 40 – 50 ani, al cărui mijloc este valoarea de 45 ani. Putem astfel afirma, faptul că vârsta tipică în intervalul cu frecvență maximă a fost de 45 ani.

Distribuția frecvențelor relative: în funcție de obiectivele cercetării, poate fi de interes transformarea frecvențelor absolute aferente fiecărui interval de grupare în frecvențe relative, pentru a evidenția proporția din totalul numărului de observații deținută de fiecare clasă. Aceasta se poate realiza prin împărțirea frecvențelor absolute ale fiecărui interval, determinate anterior, la numărul total de observații.

DEZAVANTAJ:
transformarea informațiilor inițiale despre vârste într-o distribuție de frecvențe are ca rezultat pierderea unor informații foarte detaliate, sistematizarea datelor nemaiputând evidenția vârsta exactă a fiecărui pacient analizat.

Vârsta	Frecvența absolută	Frecvența relativă
10-20	2	0.020
20-30	11	0.110
30-40	18	0.180
40-50	27	0.270
50-60	25	0.250
60-70	12	0.120
70-80	5	0.050
TOTAL	100	1.000

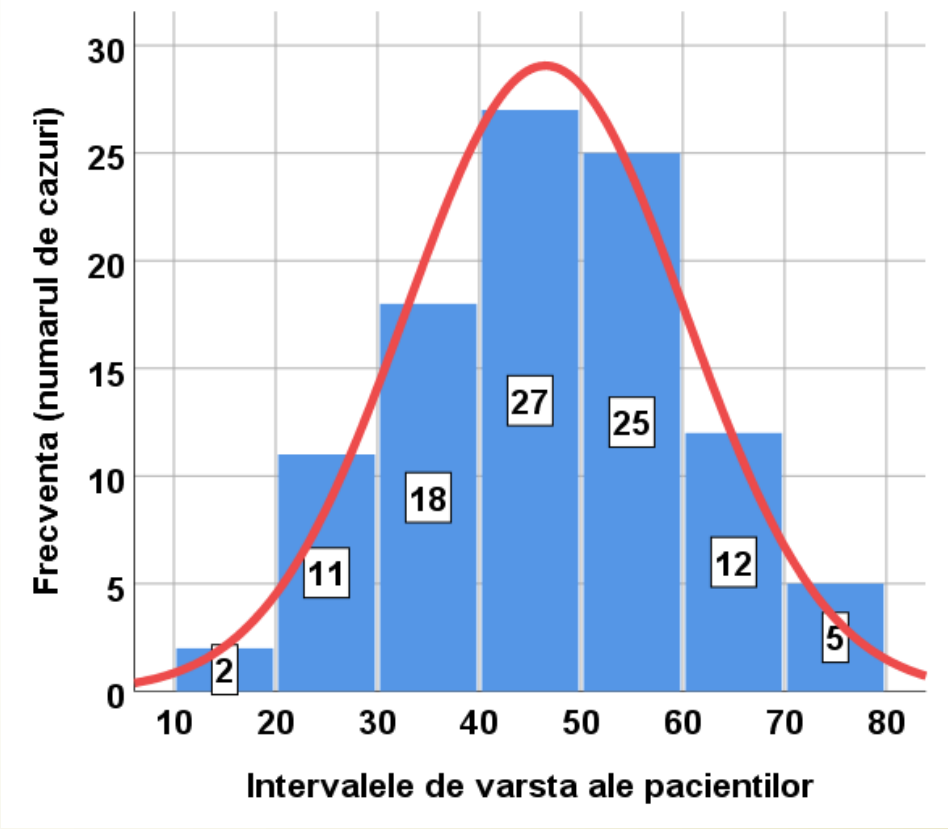
REPREZENTAREA GRAFICĂ A DISTRIBUȚIILOR DE FRECVENȚE

Managerii de spitale, medicii și alți factori de decizie au deseori nevoie de o imagine rapidă asupra distribuțiilor pacienților, prețurilor medicamentelor sau costurilor de spitalizare.

Cele trei tipuri principale de reprezentări grafice care se folosesc în cazul distribuțiilor de frecvențe sunt: histograma, poligonul frecvențelor și poligonul frecvențelor cumulate.

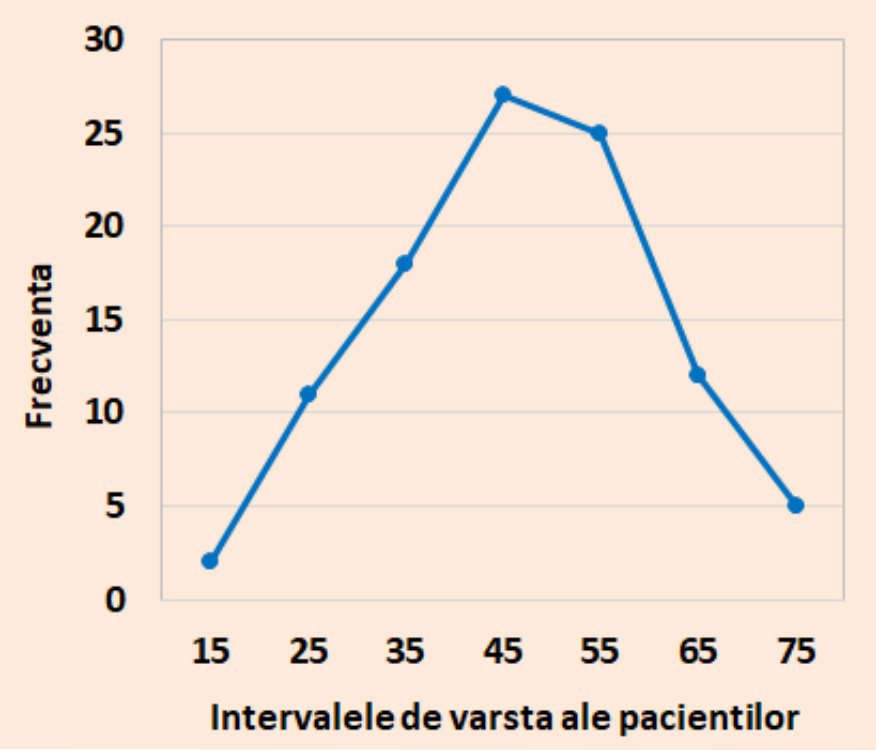
HISTOGRAMA reprezintă un grafic în care intervalele în care este grupată variabila cantitativă sunt poziționate pe axa orizontală și frecvențele aferente fiecărui interval sunt prezentate pe axa verticală sub forma unor benzi având înălțimea proporțională cu frecvența intervalului corespunzător.

Histograma pentru o distribuție de frecvențe este similară cu diagrama prin bare utilizată în cazul distribuției variabilelor calitative. Există și o diferență importantă care ține de natura internă a datelor, variabilele cantitative (numerice) sunt de obicei analizate utilizând scări de măsură care sunt continue, nu discrete.



POLIGONUL FRECVENȚELOR este un grafic care prezintă forma unei distribuții de frecvențe, fiind similar cu histograma. Se obține prin unirea unor segmente de linie care unesc punctele formate la intersecția mijlocului fiecărui interval cu frecvența corespunzătoare a acestuia.

Vârsta	Mijlocul intervalului	Frecvența
10 - 20	15	2
20 - 30	25	11
30 - 40	35	18
40 - 50	45	27
50 - 60	55	25
60 - 70	65	12
70 - 80	75	5
TOTAL	-	100



Atât histograma cât și poligonul frecvențelor ne permit să ne facem o imagine rapidă a caracteristicilor principale ale datelor (valori minime, maxime, puncte de concentrare a valorilor variabilei), fiecare dintre acestea având o serie de avantaje specifice.

DISTRIBUȚIA FRECVENȚELOR CUMULATE

Revenind la exemplul anterior, privind distribuția pacienților în funcție de variabila vârstă. Presupunem faptul că am fi interesați în cunoașterea numărului de pacienți cu vârsta mai mică de 30 de ani. Aceste valori pot fi aproximate prin construirea unui distribuții a frecvențelor cumulate și, ulterior, prin reprezentarea grafică a acesteia în forma poligonului frecvențelor absolute cumulate. Am putea de asemenea dori să cunoaștem numărul de pacienți cu vârsta peste 60 de ani, caz în care va fi utilizat poligonul frecvențelor retrocumulate.

Pentru a construi distribuția frecvențelor absolute cumulate, pornim de la tabelul anterior cu distribuția frecvențelor și observăm că au fost 2 pacienți cu vârsta sub 20 de ani. Prin însumarea celor 2 pacienți din primul interval cu cei 11 din intervalul al doilea obținem un total parțial de 13 pacienți, cu vârsta sub 30 de ani. Frecvența cumulată a următorului interval (al treilea – pacienții sub 40 de ani) care urmează este 31 (2 + 11 + 18).

Vârsta	Frecvența		Frecvența cumulată	
	absolută	relativă	absolută	relativă
20	2	2%	2	2%
30	11	11%	13	13%
40	18	18%	31	31%
50	27	27%	58	58%
60	25	25%	83	83%
70	12	12%	95	95%
80	5	5%	100	100%
TOTAL	100	100%	-	-