LUCRAREA DE LABORATOR Nr. 4
TEMA: Reprezentarea datelor statistice.

1. SCOPUL LUCRĂRII:

Introducere în analiza statistică a datelor.

2. PREZENTAREA LUCRĂRII:

Lucrarea se va realiza utilizând softul matematic Wolfram Mathematica, posibilitățile Excel, Geogebra, sau alte instrumente potrivite pe care le cunoașteți.

Numărul variantei coincide cu numărul dvs. în registrul profesorului.

1. Să se perfecteze o foaie de titlu în stil obișnuit, folosind instrumentele de tehnoredactare din Wolfram Mathematica, cu indicarea disciplinei, numelui, grupei, facultății, specialității și numărului variantei.

2. Să se efectueze un Page Break

3. În continuare, se scrie condiția exercițiului (format text) și rezolvarea acestuia(format input).

4. Se salvează fișierul cu extensia .nb cu numele fișierului: NumePrenume.grupa.nb

Lucrarea se prezintă în cadrul orei de laborator și se plasează pe platforma ELSE.

Recomandare. Instrumente utile găsiți în Wolfram Documentation, compartimentul Data Manipulation & Analysis. Alte instrumente pot fi:

https://www.geogebra.org/t/diagrams?lang=en

https://www.rapidtables.com/tools/line-graph.html

3. SARCINA DE BAZĂ

Sarcina 1.

Accesați următorul link și set de date, după cum urmează:
https://statbank.statistica.md/PxWeb/pxweb/ro/20%20Populatia%20si%20procesele%20demografice/

>Populatia stabila si procesele demografice

>>Nascuti

>>>Nascuti-vii pe raioane si medii, 2004-2018

Descărcați setul de date pentru anii 2004-2018, pentru localitățile ce corespund variantelor:

1. Dubăsari, Călărași, Telenești
2. Criuleni, Ștefan Vodă, Anenii Noi
3. Hîncești, Căușeni, Ocnița
4. Hîncești, Glodeni, Leova
5. Dubăsari, Telenești, Dondușeni
6. Glodeni, Taraclia, Cimișlia
7. Soroca, Rezina, Ialoveni
8. Glodeni, Șoldănești, Florești
9. Călărași, Florești, Căușeni
10. Hîncești, Leova, Dondușeni
11. Căușeni, Cahul, Călărași
12. Telenești, Taraclia, Edineț
13. Glodeni, Nisporeni, Edineț
14. Criuleni, Florești, Drochia
15. Glodeni, Drochia, Orhei
16. Fălești, Rezina, Ștefan Vodă
17. Briceni, Florești, Ialoveni
18. Briceni, Telenești, Ocnița
19. Telenești, Ialoveni, Anenii Noi
20. Rezina, Ștefan Vodă, Orhei
21. Soroca, Glodeni, Taraclia
22. Basarabeasca, Șoldănești, Cantemir
23. Ungheni, Cimișlia, Anenii Noi
24. Rezina, Drochia, Ștefan Vodă
25. Căușeni, Taraclia, Ștefan Vodă

Pentru datele extrase să se realizeze următoarele operații:

1. Să se calculeze pentru fiecare **regiune** în parte:
    a. valoarea medie a nașterilor
    b. dispersia
2. Să se calculeze pentru fiecare **an** în parte:
    a. valoarea medie a nașterilor
    b. dispersia
3. Să se reprezinte histogramele
4. Să se construiască poligonul frecvențelor

Sarcina 2.

Având la dispoziție rezultatele din exercițiul 7, lucrarea de laborator nr.3, să se introducă datele într-un tabel (ex. Tabelul1, suportul teoretic) și să se obțină următoarele rezultate:

a) histograma
b) poligonul frecvențelor

Înălţimea unui bărbat este o v.a. cu repartiţia normală. Presupunem că această repartiţie are parametrii $m=175+(-1)^n/n$ cm şi $\sigma=6-(-1)^n/n$ cm. Să se formeze programul de confiţionate a costumelor bărbăteşti pentru o fabrică de confecţii care se referă la asigurarea cu costume a bărbaţilor, înălţimile cărora aparţin intervalelor: [150, 155), [155, 160), [160, 165), [165, 170), [170, 175), [175, 180), [180, 185), [185, 190), [190, 195), [195, 200], *n* fiind numarul variantei, *n=1,2,…30*.

Suport teoretico-practic

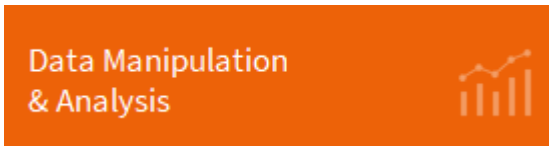
Reprezentarea grafică a datelor statistice - Consideraţii generale

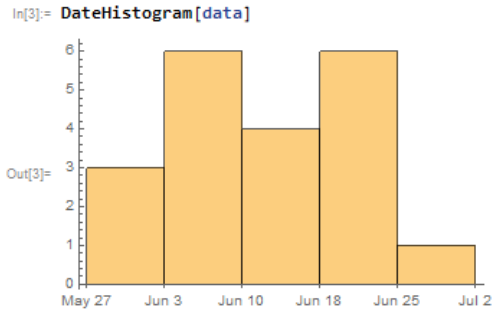
Sunt două metode de bază în statistică: **numerică şi grafică**. Folosind metoda numerică putem calcula statistici ca media şi deviaţia standard. Aceste statistici poartă informaţie despre tendinţa centrală şi variabilitate, altele poartă alt tip de informaţie. Metoda grafică este mai potrivită decât cea numerică pentru identificarea vizuală a tendinţei datelor. Metoda numerică este mai obiectivă şi mai precisă. De vreme ce se completează una pe alta, este util să le folosim combinat.

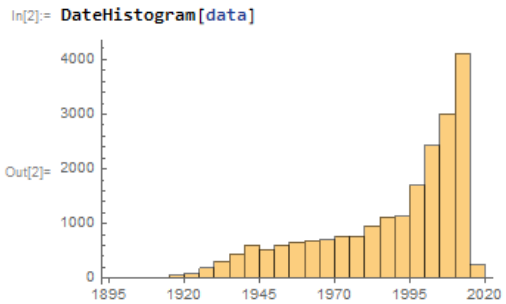
Informaţia conţinută în date culese şi înregistrate este greu de sintetizat pentru a avea o imagine cât mai clară despre situaţia pe care acestea o reflectă. Indicatorii statistici oferă o sinteză mai mult sau mai puţin fidelă a informaţiei, pierzând inerent din informaţie. Totuşi, pierderea de informaţie datorată înlocuirii unei serii de valori prin indicatorii săi nu este totdeauna o pierdere de care să ne ferim, din contră, de cele mai multe ori, indicatorii statistici oferă o imagine mai utilă decât datele în sine. De obicei, pierderea de informaţie este un rău necesar.

De la începuturile statisticii, o metodă de sintetizare a informaţiei mult folosită este reprezentarea grafică a datelor. ***Informaţia prezentată vizual*** este mult mai penetrantă pentru simţuri şi chiar pentru intelect şi de obicei ***"o imagine bună este mai utilă ca o mie de cifre "***. Reprezentarea grafică a datelor se face însă cu mult discernământ căci, aşa cum se va vedea mai jos, nu orice grafic ne spune ceva, iar cantitatea de informaţie care se pierde la reprezentare trebuie foarte atent controlată. De-a lungul timpului au fost folosite multe tipuri de grafice pentru a reprezenta cât mai bine informaţia conţinută în date. Cele mai des folosite grafice sunt histograma, graficul cu bare, poligonul frecvenţelor, graficul liniar de evoluţie în timp, diagrame, grafice punctuale etc. Pentru o mai bună înţelegere să discutăm întâi cazul unui tip de grafic care a făcut carieră în toate domeniile de aplicabilitate ale statisticii: ***histograma.***

ATENŢIE! În sistemul Wolfram Mathematica, construim histograme cu ajutorul funcției **DateHistogram, BarChart.**

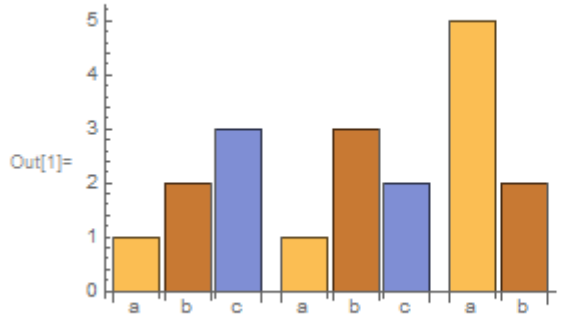






```
In[1]:= BarChart[{{1, 2, 3}, {1, 3, 2}, {5, 2}}, ChartLabels → {"a", "b", "c"}]
```

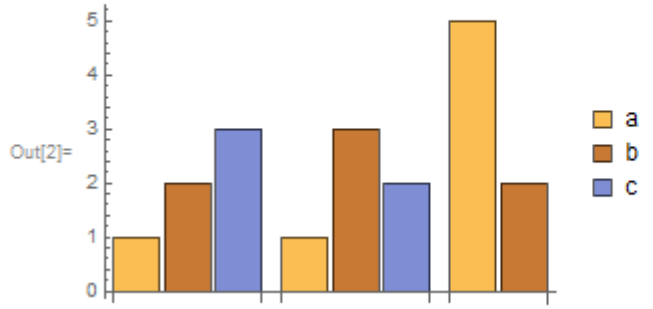
Out[1]=



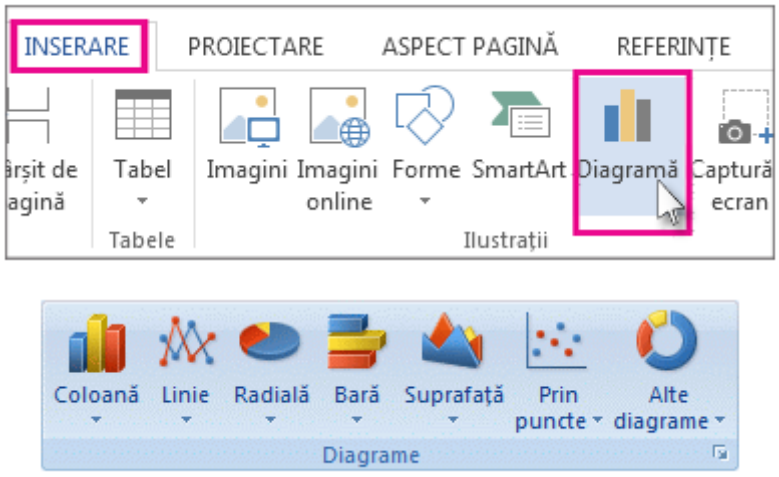
Categorical legends:

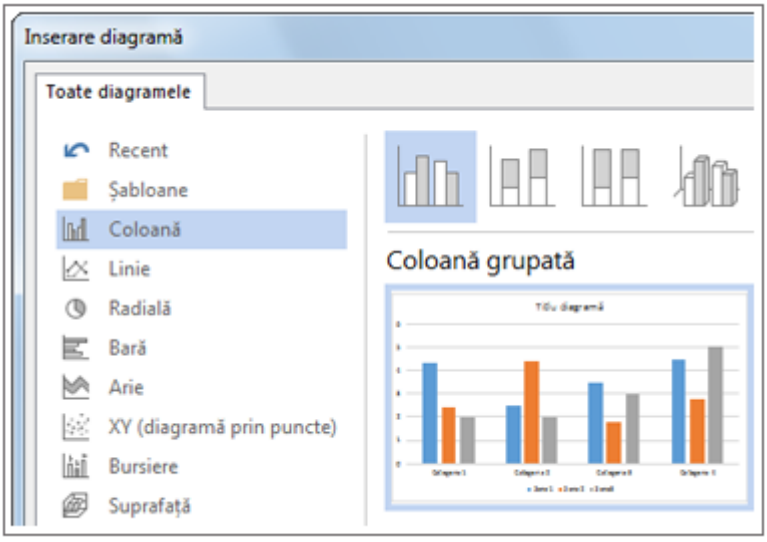
```
In[2]:= BarChart[{{1, 2, 3}, {1, 3, 2}, {5, 2}}, ChartLegends → {"a", "b", "c"}]
```

Out[2]=



Observație. Setul ”data” a fost declarat și ulterior apelat.





Ca şi concept, histograma este de fapt echivalentul grafic al tabelului de frecvenţe. Mai întâi să lucrăm pe un exemplu concret şi apoi să urmărim problemele specifice care pot face din histogramă un instrument util de lucru sau un balast. Avem mai jos un tabel care sintetizează situaţia parametrului Greutate corporală la 1014 pacienţi cu diferite afecţiuni:

Tabelul 1 Greutatea corporală a 1014 pacienţi cu diferite afecţiuni, pe clase din 5kg în 5kg

Clasa	Greutate(Kg)	Frecvenţa (Nr indivizi)
1	35..40	17
2	40..45	46
3	45..50	84
4	50..55	108
5	55..60	130
6	60..65	136
7	65..70	160
8	70..75	113
9	75..80	106
10	80..85	54
11	85..90	29
12	90..95	12
13	95..100	9

Acum să privim graficul din figura 1, care reprezintă situaţia din tabel:

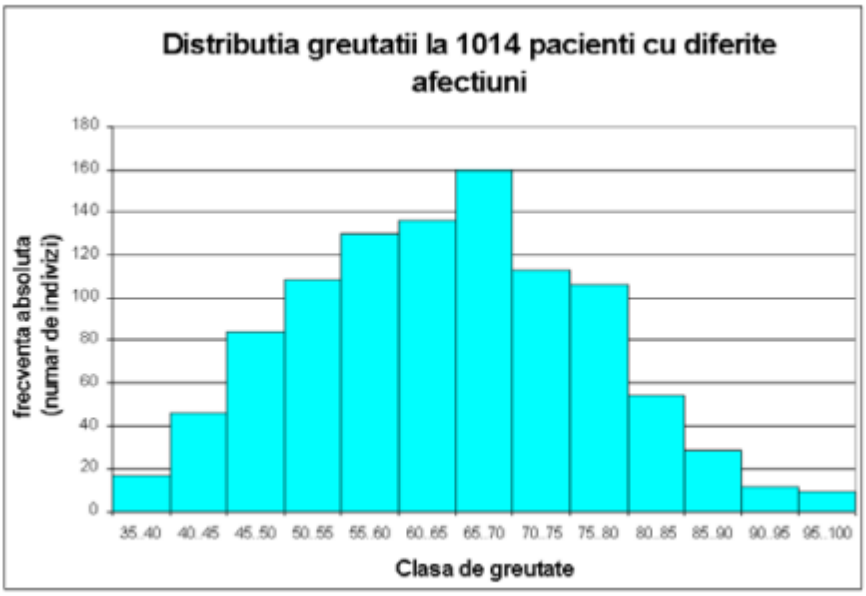


Figura 1 Histograma greutăţilor corporale a 1014 pacienţi cu diferite afecţiuni

Mai întâi, ce s-a reprezentat de fapt? Se oservă că pe orizontală sunt figurate clasele din tabel în ordine, fiecăreia fiindu-i alocat un segment de aceeaşi lungime, iar pe verticală, dreptunghiurile au înălţimi proporţionale cu frecvenţele absolute ale claselor. Mulţimea barelor verticale este cea care ne dă impresia vizuală pe care trebuie să o interpretăm în sensul sitetizării informaţiei. Observăm: Din stânga se începe cu bare scunde care cresc în înălţime pe măsură ce ne apropiem de clasa din centru, după care are loc un proces invers. Este tendinţa naturală la cele mai multe situaţii. Datele au de cele mai multe ori tendinţa de a se situa în stânga şi drepta mediei, din ce în ce mai puţine pe măsură ce ne depărtăm de medie. Pe acest grafic nu este figurată media dar este de bun simţ să ne gândim că este situată undeva în clasele de mijloc. Indivizii care au sub 35 Kg şi cei peste 100 Kg, probabil foarte puţini, nu au fost luaţi în calcul. Se obişnuieşte totuşi ca ei să fie luaţi în considerare prin introducerea a două clase speciale. În acest caz, clasele speciale de introdus ar fi fost: clasa ″sub 35″ şi clasa ″peste 100″. De obicei aşa este bine să se procedeze. Modul cum cresc barele este diferit de modul cum descresc. Aceasta este ceea ce se numea la indicatorii statistici asimetria. Această histogramă arată o uşoară asimetrie la dreapta. Dacă indivizii de la care s-au cules datele ar fi fost normali, histograma ar fi avut un aspect mai simetric. Asimetria acestei hitograme ne arată că în clasele de la 40 la 65 kg sunt mai mulţi indivizi decât în clasele simetrice lor de la 75 la 90 kg. Având în vedere că majoritatea lor sunt bărbaţi, acestă asimetrie ne spune că un număr de indivizi au greutatea mai mică decât ar fi normal. Acest lucru este explicabil în acest caz, deoarece cei mai mulţi au afecţiuni hepatice grave ca ciroză hepatică, cancer hepatic, şi sunt într-o stare fizică mult slăbită. În acest caz, am explicat forma histogramei pe baza realităţii. De obicei însă se întâmplă exact invers. Histograma este aceea care ne ajută să înţelegem mai bine realitatea.

Pentru a realiza diferenţa dintre o distribuţie simetrică şi una asimetrică, să transpunem într-o histogramă situaţia din tabelul 2, care sistematizează situaţia supravieţuirilor în cazurile de cancer mamar pe un lot de 2456 de pacienţi.

Tabelul 2 Situaţia supravieţuirilor în cazurile de cancer mamar. Gruparea în clase de 12 luni

Nr.Crt	Perioada	Nr.cazuri	Procent %	Procent Cumulat %
1	0..12 luni	672	27.36	27.36
2	12..24 luni	446	18.16	45.52
3	24..36 luni	368	15.00	60.52
4	36..48 luni	249	10.14	70.66
5	48..60 luni	196	8.00	78.66
6	60..72 luni	172	7.00	85.66
7	72..84 luni	126	5.13	90.79
8	84..96 luni	98	4.00	94.79
9	96..108 luni	45	1.83	96.62
10	108..120 luni	31	1.26	97.88
11	Peste 120 luni	52	2.12	100.00

În figura 2, este reprezentată histograma corespunzătoare pentru tabelul 2. Se observă că barele histogramei au înălţimi descrescătoare întocmai ca şi frecvenţele absolute ale claselor.

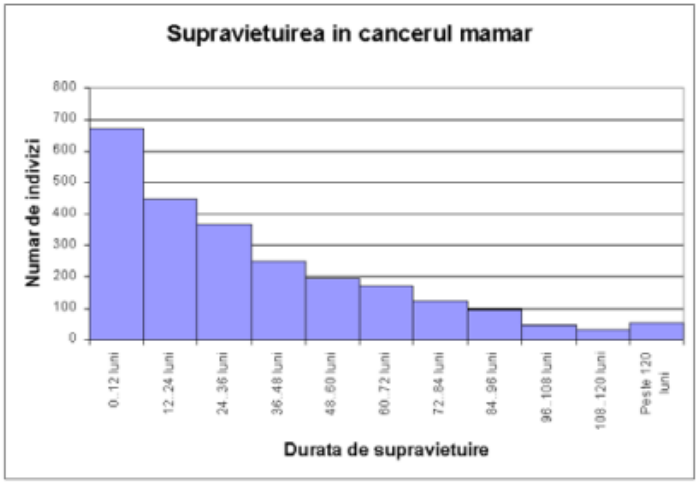
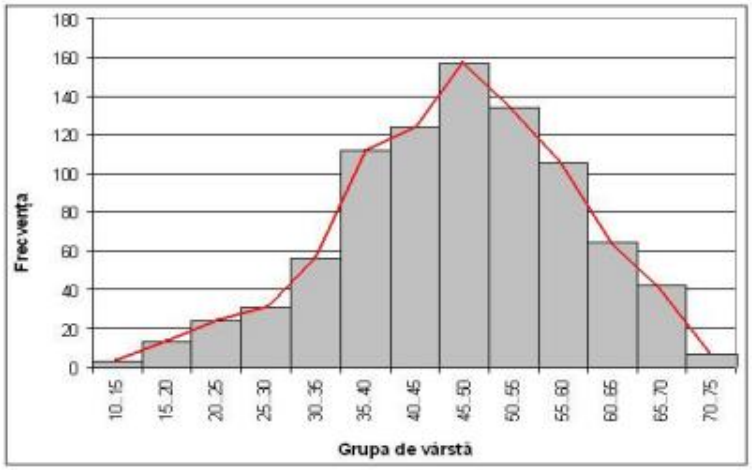


*Figura 2 Histograma corespunzătoare pentru **tabelul 2**. Se observă că barele histogramei au înălţimi descrescătoare întocmai ca şi frecvenţele absolute ale claselor*

Se observă la această histogramă că are o asimetrie foarte puternică spre dreapta. Vom considera totdeauna (ca o convenţie), să spunem că o histogramă arată asimetria spre partea unde descreşterea este mai lentă. Tendinţa observată în această histogramă este normală, având în vedere fenomenul surprins. Procesele de supravieţuire sunt de obicei marcate de o distribuţie a valorilor cu excentricitate spre dreapta, adică spre supravieţuiri lungi.

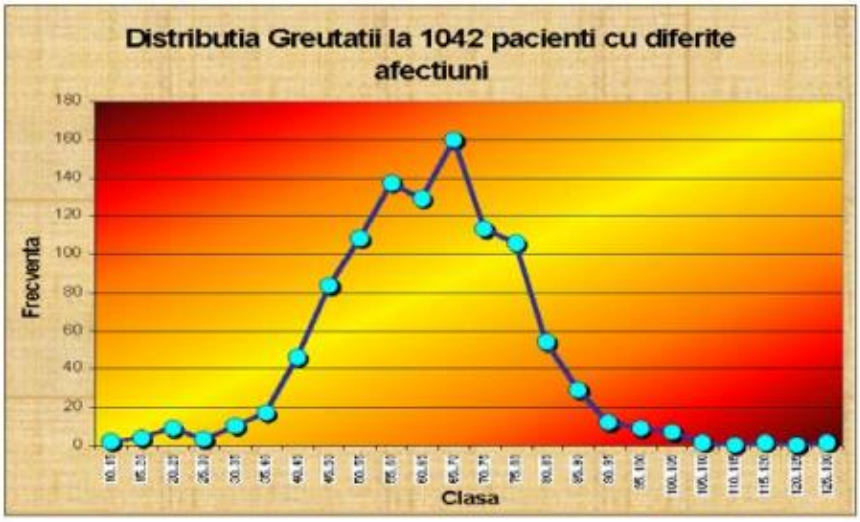
Poligonul frecvenţelor

Este un grafic care reprezintă frecvenţele absolute dintr-un tabel de frecvenţă printr-o linie frântă. Clasele se realizează ca şi la histogramă. Linia frântă, leagă puncte din plan care au ca ordonate frecvenţele de reprezentat, iar ca abscise, mijloacele claselor. Graficul se poate realiza şi din histogramă, prin unirea mijloacelor laturilor superioare ale barelor. În figura este reprezentat un exemplu de modul cum se obţine poligonul frecvenţelor din histogramă.



Poligonul frecvenţelor obţinut prin unirea mijloacelor laturilor superioare ale barelor unei histograme.

În figura 4 este reprezentat poligonul frecvenţelor pentru greutatea a 1042 de pacienţi cu diferite afecţiuni, din 5 în 5 Kg.



8 Poligonul frecvenţelor pentru greutatea a 1042 de pacienţi cu diferite afecţiuni, cu clase din 5 în 5 Kg.

Deşi oferă o imagine vizuală foarte bună a modului cum sunt distribuite valorile din serie pe clase, poligonul frecvenţelor este mai puţin folosit decât histograma, care oferă şi ea tot informaţia despre distribuţia valorilor din serie pe clase. Aceasta deoarece histograma pare ochiului un grafic mai bogat. În realitate, între cele două grafice, nu există o diferenţă calitativă. Ele oferă aceeaşi informaţie.

ATENŢIE! Graficul histogramă şi graficul poligonul frecvenţelor, conţin exact aceeaşi cantitate de informaţie, dacă au la bază acelaşi tabel de frecvenţe.

Semnificaţia statistică a histogramei

Histograma este influenţată de factori aleatori în ce priveşte forma, deci ne poate da o informaţie mai mult sau mai puţin valoroasă în funcţie de aceşti factori. Ca şi în cazul celorlalţi indicatori statisitici, vom considera histograma ca având înmagazinată

informaţie cu atât mai corectă cu cât avem un număr mai mare de indivizi în lotul pe care ea îl reprezintă.

Alegerea numărului de clase.

De obicei, programele de calculator realizează histograme după ce utilizatorul a furnizat lungimea clasei. Pentru a nu ajunge în situaţii când un astfel de tabel are un număr total neindicat de clase, de obicei se calculează lungimea aproximativă a unei clase în aşa fel încât numărul de clase să fie cel dorit. Acest lucru se poate realiza dacă se caută cea mai mică şi cea mai mare valoare din seria de date (notate mai jos cu min şi max), şi se ia ca lungime a unei clase, aproximativ rezultatul următorului calcul:

$$L = \frac{\max - \min}{nr.clase}$$

De exemplu, dacă în seria vârstelor unor pacienţi, cel mai tânăr pacient are 26 de ani, iar cel mai vârstnic are 78, pentru a obţine 6 clase (număr de clase indicat pentru vârste de adulţi), avem L= (78 - 26) / 6 = 8,6. Deci este indicat să se ia clase de 10 ani, prin rotunjire. Dacă însă se doresc mai multe clase, să zicem 10, atunci obţinem: L = (78 - 26) / 10 = 5,2 şi este indicat să se ia clase din 5 în 5 ani. Prima clasă va fi [25,30), iar următoarele: [30, 35), [35, 40),….[75, 80). Numărul de clase nu este neapărat 10, el se alege de fapt de către cel care face histograma, astfel ca să se piardă cât mai puţină informaţie, dar şi numărul de clase să nu fie prea mare căci atunci luăm în considerare aspecte prea nesemnificative.

Ca regulă generală, este bine să se reţină că:

• Se pierde cu atât mai multă informaţie cu cât numărul de clase este mai mic. Nu se recomandă histograme cu 2-4 clase

• Un număr prea mare de clase duce la o ascundere a esenţialului de către aspectele nesemnificative

Întrucât cei care nu au experienţă nu ştiu cum să aleagă numărul de clase, recomandăm: • Pentru câteva zeci de valori, să se aleagă maximum 6 – 8 clase • Pentru câteva sute de valori, să se aleagă între 10 şi 15 clase • Pentru câteva mii de valori, să se aleagă peste 15 clase Nu se recomandă folosirea a mai mult de 20 – 30 de clase decât în cazuri speciale, în studii cu multe mii de cazuri. Nici mai puţin de 4 – 6 clase nu este recomandat să se folosească. Nu se recomandă folosirea histogramelor dacă nu avem cel puţin câteva zeci de valori. De exemplu, pentru o serie de 15 valori, nu se face o histogramă.

Curba densităţii de probabilitate

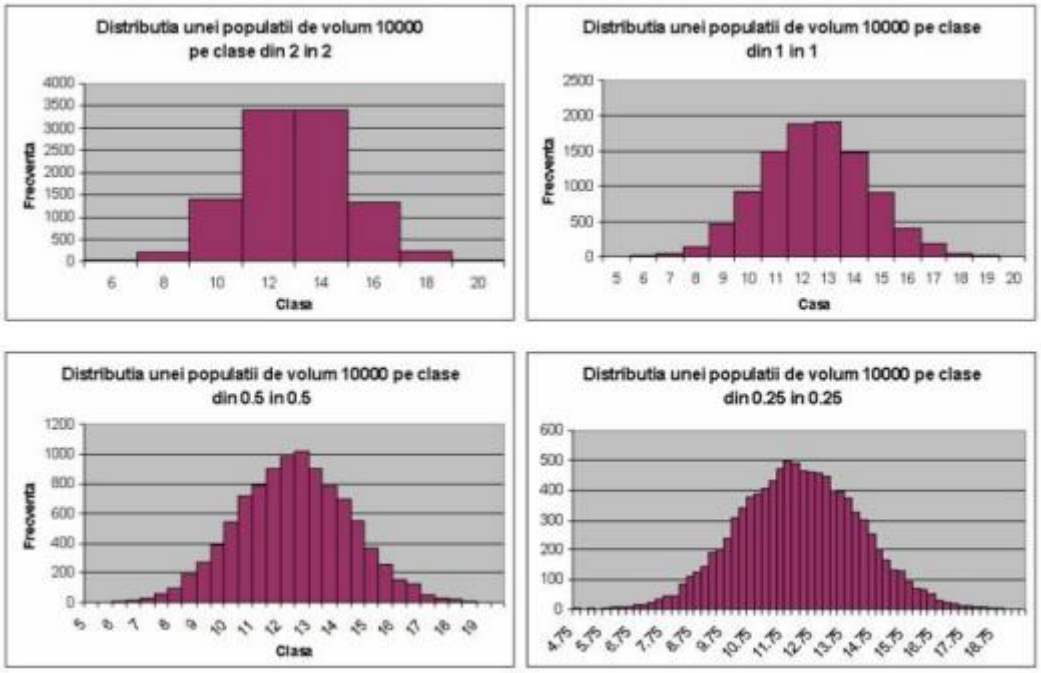
S-a văzut că histograma este un grafic care dă informaţii despre repartizarea valorilor dintr-o serie de valori, care arată dacă valorile din serie sunt repartizate simetric sau asimetric şi dacă repartiţia are un singur vârf sau este multimodală. Să ne imaginăm că pe măsură ce mărim indefinit numărul de valori din serie, lungimea claselor scade foarte mult, astfel încât obţinem histograme din ce în ce mai “fine”. Ce se obţine prin acest proces? O apropiere din ce în ce mai accentuată de repartiţia reală a datelor, repartiţie

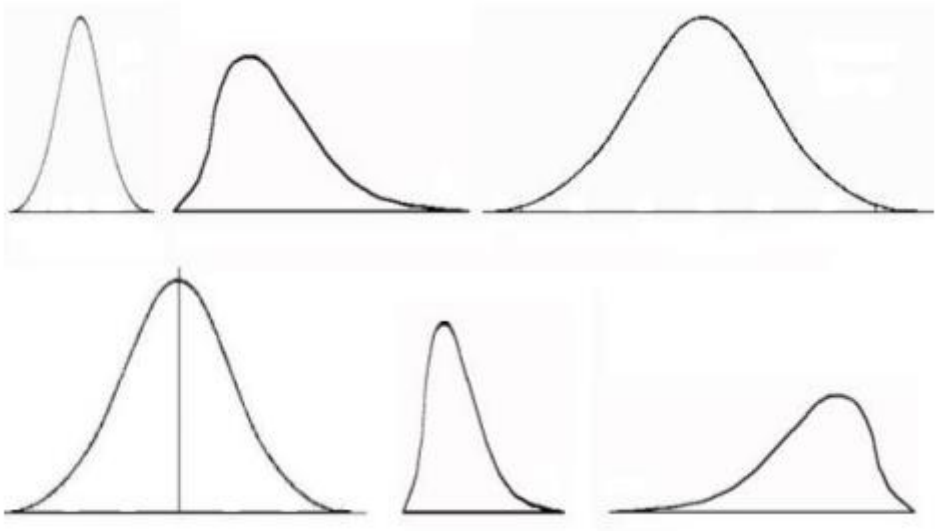
pe care histogramele o aproximează din ce în ce mai bine. Histogramele oferă imaginea repartizării valorilor dintr-o serie, deci o imagine incompletă a realităţii. Într-adevăr, valorile dintr-o serie de date sunt culese pe un eşantion sau lot, care este de obicei extras dintr-o populaţie mult mai numeroasă. Ceea ce ne interesează de obicei însă, este modul cum se repartizează valorile din întreaga populaţie.

Pe măsură ce histogramele devin din ce în ce mai fine, ele tind să se asemene cu o curbă. Dacă volumul seriei ar fi mult mai mare, asemănarea cu o curbă ar fi atât de clară încât ochiul nu ar mai putea observa aspectul de ″treaptă″. Acest proces este vizibil în special atunci când în locul histogramelor folosim poligoane ale frecvenţelor

Acest mod de a ajunge la o curbă a densităţii de probabilitate (sau o curbă de repartiţie) este instructiv prin faptul că oferă o imagine intuitivă a diferenţei dintre o histogramă sau un poligon al frecvenţelor şi o curbă de repartiţie. În plus, oferă o ideie despre cum arată curba de repartiţie. Strict vorbind însă, noţiunea de curbă a densităţii de probabilitate, trebuie introdusă folosind un aparat teoretic mai complex. Deoarece o introducere fundamentată ar depăşi nivelul cursului de faţă, vom considera, intuitiv, fără a pretinde că aceasta este o definiţie riguroasă, că: O curbă a densităţii de repartiţie este curba care are acelaşi aspect cu curba către care tinde poligonul frecvenţelor relative, atunci când numărul de valori dintr-o serie tinde la infinit, iar lungimea fiecărei clase tinde la 0. Pentru o exprimare mai clară, atunci când nu există pericolul unor confuzii, în locul termenului de curbă a densităţii de probabilitate, vom folosi termenul de curbă de repartiţie, sau mai simplu, repartiţie.

Pe măsură ce statistica a evoluat ca ştiinţă, s-a demonstrat că unele din curbele densităţii de probabilitate joacă un rol central în ştiinţă în general şi în medicină în special. Astfel, multe fenomene din ştiinţă se petrec astfel încât deviaţiile stânga-dreapta de la medie ale măsurătorilor pe care le facem sunt repartizate simetric şi nu oricum, ci tind să fie repartizate foarte asemănător cu o anumită curbă, mult studiată, care se numeşte curba densităţii normale sau curba Gauss.





Diverse forme ale curbei densităţii de probabilitate