Universitatea de Nord Baia Mare
Facultatea de Stiinte
Catedra de Chimie Biologie

Denumirea disciplinei: **Prelucrarea datelor experimentale**

Profilul: Chimie
Specializarea: Chimie
An de studiu: 3
Semestrul: 6
Tipul de evaluare: C
Numar de credite: 6
Titular disciplinei: Sef lucrari dr. ing. Leonard Mihaly Cozmuta

Obiectivele disciplinei:

Cursul are ca obiectiv principal modul de prelurare a datelor experimetale. Se vor urmari analizarea surselor de erori care intervin pe parcursul procesului analitic, a modului de evidentiere a erorilor, utilizarea unui aparat matematic in sistematizarea si caracterizarea rezultatelor experimentale. Se va avea in vedere aplicarea diferitelor teste statistice pentru eliminarea datelor necoresunzatoare. In partea finala a cursului se vor aborda problemele legate de studiul relatiilor dintre variabile, regresia liniara simpla si multipla, necesare in intelegerea comportarii, evolutiei si simularii comportarii unui sistem fizico-chimic.

Programa analitica		
Tipul activitatii	Continutul	Ore alocate
Curs		
	1. Marimi fizice. Principii de masurare	2
	2. Surse de erori in procesul analitic	2
	3. Parametrii statistici care caracterizeaza distributia datelor	2
	4. Sistematizarea si prezentarea datelor statistice	2
	5. Legi de repartitie Repartitia normala (clopotul lui Gauss)	2
	6. Indici de asimetrie si boltire a distributiei datelor	2
	7. Esantionarea. Distributia de esantionare. Intervale de incredere	2
	8. Testarea omogenitatii dispersiilor (testul Cochran). Compararea a doua esantioane (raportul lui Fisher)	2
	9. Estimarea preciziei si controlul preciziei (testul χ^2)	2
	10. Estimarea exactitatii (testul Link si Walace)	2
	11. Controlul de calitate in analiza chimica	2
	12. Eliminarea datelor necorespunzatoare	2
	13. Studiul relatiilor dintre variabile. Regresia liniara simpla	2
	14. Regresia liniara multipla	2
	TOTAL	**28**
Lucrari de laborator		
	1. Surse de erori in analiza fizico-chimica. Aplicatie in analiza volumetrica	2
	2. Precizia si exactitatea unei metode de analiza. Estimarea preciziei si exactitatii unor metode diferite de determinare a cuprului.	2
	3. Parametrii statistici utilizati in prelucrarea datelor experimentale	
	4. Metode de evidentiere a surselor de erori in analiza fizico-chimica	2
	5. Repartitia normala. Clopotul lui Gauss	2
	6. Eliminarea datelor experimentale necorespunzatoare. Utilizarea si compararea diferitelor teste de eliminare a datelor necorespunzatoare.	2
	7. Trasarea dreptei de etalonare. Utilizarea metodei celor mai mici patrate	2
	8. Trasarea graficelor semilogaritmice si logaritmice	
	9. Prelucrarea semnalului in analiza potentiometrica, conductometrica si polarografica	4
	10. Trasarea si interpretarea curbei granulometrice pentru un material solid	2
	11. Studiul relatiilor dintre variabile. Coeficient de corelatie	4
	12. Finalizarea laboratorului. Evaluarea rezultatelor	2
	TOTAL	**28**

Bibliografie

1. FOO-TIM CHAU, YI-ZENG LIANG, JUNBIN GAO, XUE-GUANG SHAO - Chemometrics From Basics to Wavelet , Transform, Published by John Wiley & Sons, Inc., Hoboken, New Jersey, 2004, ISBN 0-471-20242-8
2. Johan Gasteiger, Chemoinformatics, *Chemoinformatics: A Textbook.* Edited by Johann Gasteiger and Thomas Engel, 2003 Wiley-VCH Verlag GmbH & Co. KGaA., ISBN: 3-527-30681-1
3. J.W. Einax, H.W.Zwanyiger, S. Geib, - Chemometrics in environmental analysis, 1997, Wiley-VCH verlag GmbH, ISBN 3-527-28772-8

2. MARIMI FIZICE. PRINCIPII DE MASURARE. SURSE DE ERORI

2.1. Marimi fizice, unitati de masura

Studiul oricarui proces, fenomen fizic sau chimic are la baza marimile fizice. Pentru a fi masurabila o marime trebuie sa fie definita calitativ si cantitativ. Orice marime fizica prezinta doua componente: valoare si natura.

Valoarea marimii fizice reprezinta un raport intre marimea ei si o marime de referinta considerata a fi egala cu unitatea. Totalitatea valorilor pe care le poate lua o marime fizica corespunde **multimii starilor sau intensitatilor**. Marimea de referinta corespunde elementului considerat egal cu unitatea din multimea starilor. Aceasta multime trebuie sa fie **strict ordonata si trebuie sa se stabileasca o relatie biunivoca cu multimea numerelor reale**. Astfel fiecarui element din multimea starilor sa-i corespunda un numar real si invers fiecarui numar real sa-i corespunda un anumit element din multimea starilor (sau intensitatilor). Pe aceasta baza se stabileste o **scara de masurare si o unitate de masura**. O marime fizica poate avea aceeasi **valoare**, intensitate sau stare dar poate avea unitate de masura diferita functie de marimea unitatii de referinta.

Marimile pot fi aditive (masa) sau neaditive (pH, temperatura). In cazul marimilor neaditive scara de marime se alege conventional, in general prin alegerea a doua valori care determina un interval. Trecerea de la o scara la alta se va face prin **interpolare**. Un exemplu reprezentativ pentru acest caz ar constitui doua segmente de dreapta de lungimi diferite. Fiecarui punct de pe segmentul scurt o sa-i corespunda un unic punct situat pe segmentul lung. Pe acest considerent pe ambele segment ar trebui sa existe acelasi numar de puncte. Aparent pe segmentul lung trebuie sa existe

mai multe puncte decat pe cel scurt. Realitatea este aceea ca ambele segmente de dreapta indiferent de lungime contin un numar infinit, deci identic de puncte.

Natura marimii defineste aprecierea calitativa a acesteia si se exprima printr-un simbol: L – lungime; M – masa; T – timp. Pentru fiecare natura exista o multime a starilor sau intensitatilor.

Marimile fizice pot fi:

- extensive – prezinta proprietati de ordonare si sumabilitate
- intensive – prezinta doar proprietati de ordonare
- scalare – sunt determinate doar de valoarea lor numerica
- vectoriale – care asociaza fiecarei coordonate cate un vector (forta)

O alta clasificare a marimilor fizice le imparte in fundamentale (independente alese conventional) si derivate sau secundare (care se definesc in baza celor fundamentale). Alegerea unitatilor fundamentale este arbitrara, dar se prefera utilizarea unui sistem unitar cu scopul exprimarii marimilor derivate fara interventia unor factori de proportionalitate. Totalitatea marimilor fundamentale si derivate utilizate intr-un domeniu alcatuiesc un sistem de unitati de masura. Pe parcursul dezvoltarii cunoasterii au existat sisteme de unitati diferite care difera fie prin natura marimilor fundamentale fie prin valorile unitatilor de masura. Cele mai importante sisteme de unitati de masura sunt: CGS (centimetrul, gram, secunda), MKfS (metrul, kilogram forta, secunda) si sistemul international de unitati de masura adoptat in 1961 MKS (metrul, kilogram, secunda).

2.2. Principii de masurare

Principiul de masurare descrie procesele (reactiile chimice) prin care se actioneaza asupra materialului supus analizei in scopul de a obtine un semnal analitic corelat cu elementul care se doreste a fi masurat. Prin prelucrarea ulterioara a acestui semnal se realizeaza masurarea efectiva.

Metoda de analiza descrie toti pasii necesari a fi efectuati intre pregatirea probei si obtinerea rezultatelor sau a datelor analitice. Procedura contine toti acesti pasi inclusiv interpretarea datelor exeprimentale. Metodele de analiza pot fi clasificate in metode directe (volumetria directa, gravimetria) in care se masoara direct elementul a fi dozat si metode indirecte (instrumentale in general) prin care se masoara efectul acestui

element sau se aplica un stimul acestuia cu scopul masurarii efectului selectiv si caracteristic provocat. Acest semnal analitic este corelat cu cantitatea elementului prezent in proba si prin utilizarea unei asa numite curbe de calibrare se poate determina valoarea acestuia. Trebuie mentionat faptul ca trasarea curbei de calibrare are la baza utilizarea unor etaloane (de obicei greu de fabricat) in care se cunoaste cu exactitate continutul elementului urmarit. Problemele care mai apar in etapa de calibrare sunt stabilirea domeniului calibrarii (de cele mai multe ori liniar) si in trasarea celei mai bune drepte sau curbe de calibrare. Aceste probleme sunt elegant rezolvate in baza unor metode matematice, utilizate in chemometrie.

Caracteristica principala a principiului de masurare este aceea ca trebuie sa duca la obtinerea unui **semnal selectiv si caracteristic** pentru elementul urmarit a fi dozat si sa asigure evitarea sau minimalizarea interferentelor altor elemente prezente in proba cu scopul evitarii erorilor. In marea majoritate a cazurilor se stie faptul ca erorile variaza cu continutul elementului a fi dozat in sensul ca la scaderea concentratiei cresc erorile.

Posibilitatea determinarii elementului urmarit in baza relatiei stimul - semnal caracteristic, a dus la o puternica permanenta dezvoltare a aparatelor de analiza, in timp ce metodele directe de analiza au atins un anumit nivel de stagnare.

O anumita metoda de analiza poate fi descrisa de exactitate si precizie. **Exactitatea** unei metode arata masura in care aceasta permite obtinerea unui rezultat apropiat de realitate. O metoda este cu atat mai exacta cu cat rezultatul obtinut este mai apropiat de valoarea reala. Pentru evaluarea exactitatii unei metode analitice este necesar utilizarea etaloanelor.

Precizia unei metode analitice reda gradul de dispersie al rezultatelor, obtinute pe aceeasi proba in conditii similare de lucru, in jurul valorii medii. Cu cat rezultatele sunt mai apropiate de valoare medie cu atat metoda este mai precisa. Aprecierea preciziei unei metode de analiza se realizeaza prin utilizarea unor procedee statistice.

O metoda de analiza poate fi foarte precisa dar mai putin exacta sau invers. O analogie la aceasta problema ar putea fi reprezentata de un concurs de tir. Un sportiv care are loviturile mai grupate dar pe marginea tintei va fi mai precis si mai putin exact decat alt sportiv care are loviturile mai imprastiate pe suprafata tintei dar cu media lor mai apropiata de punctajul maxim, sportiv care este mai putin precis dar mai exact. In practica analitica

se urmareste utilizarea unor metode care sa asigure atat o exactitate mai ridicata cat si o precizie mai mare.

2.3. Surse de erori in procesul analitic

Datele obtinute in urma unui proces analitic de masurare a unui element continut intr-o anumita proba sunt afectate intr-o masura mai mare sau mai mica de o serie de erori. Acestea provin din mai multe surse si sunt de diferite naturi. Rezultatul final este afectat de erori care pot interveni in fiecare pas al procesului analitic, incepand de la modalitatea de recoltare si pana la calculul final si interpretarea rezultatului.

Pentru diminuarea sau eliminarea erorilor cauzate de factorul uman se recomanda ca atat analiza cat si recoltarea probei sa fie realizata de catre operatorul chimist, respectand cu strictete atat normele de prelevare cat si metodologia metodei de analiza. **Erorile grosolane** duc la obtinerea unor rezultate care nu au nimic de-a face cu analiza urmarita si cauzeaza costuri legate de consumul inutil de resurse. Acestea se imputa persoanei in cauza. **Erorile subiective** sunt cauzate tot de factorul uman si constau in erori de citire sau de apreciere a volumului de echivalenta, de realizare a dilutiilor, etc. si genereaza vicierea rezultatului.

Erorile de procedura sunt erorile care apar inerent datorita neaplicarii, a aplicarii incorecte sau a imposibilitatii de aplicare exacta a normelor si standardelor legate de recoltare, reducerea probei si de pregatirea probei in vederea analizarii ei. Problema cea mai importanta care apare aici este legata de obtinerea unei probe reprezentative. In situatia in care se urmareste determinarea continutului unui element pe o suprafata sau dintr-un volum ridicat este necesar recoltarea unui numar mare de probe individuale de pe suprafete determinate prin impartirea suprafetei mari in suprafete uniforme sau concentrate functie de natura si sursa de poluare. Acestea vor fi amestecate ulterior iar prin metode specifice cum ar fi metoda dreptunghiului sau a conului var fi aduse la un volum mai redus. Erorile care apar in pregatirea probei tin de modalitatea de dezagregare sau de aducere in solutie. Majoritatea analizelor se efectueaza in solutie. Dezagregarea materialului solid trebuie sa asigure trecerea complecta a elementului urmarit a fi dozat din starea solida (sub forma de compusi insolubili) in stare lichida sub forma de saruri solubile. In mod similar in cazul analizei gazelor sistemele de adsorbtie – absorbtie trebuie sa asigure in final trecerea

cantitativa a elementului analizat din starea gazoasa in starea lichida. Din aceasta cauza si pregatirea probei constituie o bogata sursa de erori.

Erorile de analiza pot aparea in cazul in care se utilizeaza o metoda de analiza cu un principiu corect dar nu se asigura respectarea conditiilor de lucru care sa asigure indepartarea interferentelor. Alte tipuri de erori de analiza apar in cazul etalonarii aparatelor si instrumentelor in cazul in care se utilizeaza etaloane necorespunzatoare sau nu se obtine cea mai buna curba de calibrare. O alta sursa de erori de acest tip viciaza rezultatul in cazul in care extrapolarea se realizeaza la limitele domeniului de calibrare.

Dupa natura lor erorile pot fi intamplatoare sau sistematice, independente sau dependente de continutul elementului urmarit a fi dozat din proba. Erorile intamplatoare pot fi de tipul erorilor datorate factorului uman, cele **sistematice** de obicei sunt datorate metodei sau aparatului. Erorile sistematice deplaseaza valorile obtinute intr-un singur sens, fie spre valori mai mari fie spre valori mai mici decat valoarea reala. Realizarea dezagregarii incomplecte a unei probe solide are ca urmare trecerea in solutie partiala a elementului urmarit generand erori sistematice cu obtinerea unor rezultate totdeauna mai mici decat valorile reale. Un alt tip de eroare sistematica independenta de marimea masurata poate fi considerate si lipsa de puritate a unui reactiv utilizat pe parcursul analizei. Trasarea unei curbe de calibrare cu panta mai mare (sau mai mica) decat cea optima genereaza aparitia unor erori sistematice dependente de concentratie elementului. In cazul extrapolarii cu cat ne vom apropia de limita superioara a domeniului de calibrare cu atat erorile obtinute vor fi mai mari sau mai mici.

In analiza chimica evidentierea erorilor intamplatoare se poate realiza prin utilizarea metodelor statistice de analiza a datelor. Deasemenea se pot aplica diferite teste de eliminare a datelor necorespunzatoare care ar vicia valoarea mediei. Aici intervine chemometria: in scopul de a obtine cele mai bune rezultate atat prin interpretarea datelor dar si prin optimizarea procesului analitic.

3. BAZA STATISTICA A MASURATORILOR

Consideram ca am efectuat in conditii similare o analiza chimica utilizand aceeasi metoda cu acelasi principiu si am obtinut un numar de n rezultate sau date experimentale. Totalitatea datelor obtinute alcatuieste o populatie de date sau o populatie statistica. Populatia statistica se defineste ca fiind o multime definita de obiecte de aceeasi natura. Elementele populatiei se numesc unitati statistice sau indivizi. Numarul elementelor

definesc volumul sau efectivul populatiei. O submultime de elemente a populatiei constituie un esantion. In tratarea statistica a datelor se utilizeaza o serie de parametrii care caracterizeaza tendinta de centrare sau imprastiere a datelor experimentale.

3.1. Parametrii statistici care caracterizeaza distributia datelor

3.1.1. Indicatori ai tendintei de centrare a datelor

1. Media aritmetica. Aceasta reprezinta valoarea medie a rezultatului . Aceasta estimeaza tendinta de centrare a datelor dar este puternic influentata de catre valorile extreme.

$$x_m = \bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

In cazul in care in sirul de date apar elemente care se repeta, definim notiunea de **frecventa** ca fiind numarul de repetitie al acesteia in cadrul populatiei obtinute in urma masuratorii. **Frecventa relativa** reprezinta raportul dintre frecventa individuala si suma frecventelor individuale a tuturor datelor. In acest caz media aritmetica se va determina cu formula:

$$x_m = \bar{x} = \sum_{i=1}^{i=n} \frac{f_i \cdot x_i}{f_i}$$

2. Mediana sau valoarea de mijloc se obtine prin ordonarea crescatoare a datelor si identificarea datei situate la mijlocul seriei. In cazul in care aceasta serie contine un numar impar de date, mediana va fi considerata valoarea situate la mijlocul seriei. In cazul in care aceasta serie contine un numar par de date, mediana va fi considerata media aritmetica a celor doua date situate la mijlocul seriei. Mediana nu mai este influentata de catre valorile extreme.

$x_1 \leq x_2 \leq \ldots\ldots\ldots\ldots \leq x_m \leq \ldots\ldots \leq x_n$

$x_1, x_2, \ldots\ldots\ldots\ldots x_m, \ldots\ldots\ldots\ldots x_n$ daca n este impara $n=2k+1$ $m = (n+1)/2$

$Me = x_m$

$x_1, x_2, \ldots\ldots\ldots\ldots x_m, x_{m+1}, \ldots\ldots\ldots\ldots x_n$ daca n este para $n=2k$ $m = n/2$

$Me = (x_m + x_{m+1})/2$

3. Modulul reprezinta valoarea care apare cu frecventa cea mai mare. Functie de acest parametru populatia de date poate fi clasificata in unimodala sau polimodala. O functie polimodala arata neomogenitatea datelor, adica arata faptul ca datele obtinute nu fac parte din aceesi populatie.
Exemple:

- pentru sirul de date: 1,2,3,4,4,4, 5,6,7,8,9 modulul este $M_O = 4$
- pentru sirul de date: 1,2,3,4,4,5,6,6,6,6,7,8,9 cele doua module sunt $M_O^1 = 6$ si
  $M_O^2 = 4$

3.1.2. Indicatori ai tendintei de imprastiere a datelor

1. Amplitudinea sau domeniul datelor reprezinta diferenta dintre cea mai mare si cea mai mica valoare. Cu cat amplitudinea va fi mai mica cu atat valorile vor fi mai apropiate si frecventa de aparitie a unei valori individuale mai mare.

$A = x_n - x_1$

2. Abaterea medie patratica sau abaterea standard sau deviatia standard este parametrul principal care exprima imprastierea rezultatelor in jurul valorii medii, fiind un indicator al preciziei (al reproductibilitatii rezultatelor). Deasemenea abaterea patratica standard este un indicator de punere in evidenta a erorilor intamplatoare care afecteaza procesul de analiza. In cazul unei distributii normale a datelor, se calculeaza cu formula:

$$s = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - x_m)^2}{n-1}}$$

3. Dispersia sau varianta reprezinta patratul abaterii standard si masoara gradul de împrăştiere a eşantionului în jurul mediei de sondaj. Presupunând că există n elemente în eşantion, cu valorile $\{x_1, x_2, \ldots, x_n\}$, având media $M = (x_1 + x_2 + \ldots + x_n)/n$, atunci dispersia este:

$s^2 = [(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2]/(n-1)$

4. Deviatia medie a datelor reprezinta media aritmetica a valorilor absolute a deviatiilor individuale a datelor in jurul valorii medii. Deviatia individuala reprezinta valoarea absoluta a diferentei intre valoarea individuala si valoarea medie aritmetica a acestor valori.

$d_i = \left| x_i - x_m \right|$

$$d_m = \frac{\sum_{i=1}^{i=n} \left| x_i - x_m \right|}{n}$$

Mentionam faptul ca suma deviatiilor calculate in valori reale si nu absolute este nula. In acest caz deviatiile pozitive vor anula deviatiile negative ale valorilor individuale fata de valoarea medie aritmetica.

5. Coeficientul de variatie sau variabilitate (coeficientul de variatie al lui Pearson) Este utilizat în scopul stabilirii gradului de omogenitate a unui esantion si se obtine prin raportarea abaterii standard la media esantionului. Rezultatul obtinut se raporteaza apoi în procente.

$$V = \frac{s}{x_m} \cdot 100$$

Spre exemplu, daca $x_m = 11,40$, iar $s = 2,7$, vom avea:

$V = (2,7/11,4)*100 = 23,68\%$

Interpretarea coeficientului de variabilitate se face în functie de valorile obtinute:

- daca coeficientul este cuprins între 0 si 15%, înseamna ca împrastierea datelor este foarte mica, iar media este reprezentativa, deoarece esantionul masurat este omogen;

- daca valoarea lui este între 15 si 30%, împrastierea datelor este mijlocie, media fiind înca suficient de reprezentativa;

- daca coeficientul depaseste 30%, media aritmetica nu este reprezentativa pentru esantionul în cauza, fiind recomandata utilizarea medianei din cauza lipsei de omogenitate a grupului.

OBSERVATIE: Acest coeficient este aplicabil doar în cazul variabilelor masurate pe scala de raport, cu origine naturala zero.

3.2. Sistematizarea si prezentarea datelor statistice

Sistematizarea constituie o etapă în cadrul prelucrării datelor statistice în vederea prezentării acestora sub formă de serie statistică (tabele statistice).

Datele obţinute ca urmare a procesului de observare statistică, în forma lor brută, permit o caracterizare amănunţită a fiecărei unităţi din populaţia considerată. Deoarece, datele rezultate din observare se prezintă sub formă dezorganizată nu permit o caracterizare a populaţiei în ansamblu.

În vederea atingerii scopului cercetării statistice întreprinse şi anume acela de a da o caracterizare de ansamblu a populaţiei considerate este necesar ca datele rezultate din observare să fie supuse unor operaţii de sistematizare şi prezentare în vederea deducerii a ceea ce este esenţial, tipic şi general în legătură cu populaţia.

Deoarece în prelucrarea statistică primul pas îl constituie prezentarea datelor observate sub forma de serie (tabel), pentru construirea seriilor statistice se aleg variabilele care trebuie să fie în strânsă dependenţă cu scopul cercetării şi cu natura fenomenului cercetat.

3.2.1. Quantile. Quartile. Decile. Centile. Box ploturi

Ansamblul quantilelor de ordinul k impart setul de date in k parti egale din punct de vedere al numarului de valori. Similar se definesc decilele si centilele, care impart setul de valori in 10, respective 100 de parti egale, care contin acelasi numar de date.

Quartilele impart setul de date in patru parti din punct de vadere al numarului de valori. Acestea prezinta cea mai mare importanta. Quartilele prezinta 5 ordine:

- quartila de ordin 0 este identica cu valoarea minima inregistrata din sirul de date
- quartila de ordin 4 este identica cu valoarea maxima
- quartila de ordin 2 este identica cu valoarea medianei
- quartila de ordin 1 este identica cu valoarea medianei elementelor cuprinse sub valoarea medianei
- quartila de ordin 3 este identica cu valoarea medianei elementelor mai mari decat valoarea medianei centrale

In baza quartilelor se construiesc **box ploturile**. O diagramă de tip boxplot reflectă grafic rezumarea setului de date experimentale prin cele 5 valori a unei distribuţii: valoarea minimă, prima quartilă, mediana, a treia quartilă şi valoarea maximă. Pe aceste tipuri de grafice se poate reprezenta si limitele inferioare si superioare in afara carora datele se considera a fi aberante. Distanta dintre prima si a treia quartila se numeste interval interquartil (contine 50% din date). Limitele inferioara si superioara se stabilesc functie de acest interval interquartil D. Valorile aberante (care nu apartin populatiei) se considera a fi acele valori mai mari decat LS si mai mici decat LI.

$D = q_{0.75} - q_{0.25}$

$LS = q_{0.75} + 1.5 \cdot (q_{0.75} - q_{0.25})$

$LI = q_{0.25} - 1.5 \cdot (q_{0.75} - q_{0.25})$

3.2.2. Impartirea datelor experimentale in clase

Operaţia de **stabilire a claselor** presupune împărţirea unităţilor unei populaţii în clase distincte în raport cu una sau mai multe variabile şi aranjarea claselor rezultate într-o anumită ordine. În urma unei asemenea operaţii, fiecare unitate trebuie să se găsească în una şi numai una din clasele rezultate. Această operaţie nu trebuie să conducă la pierderi de unităţi, modificand doar ordinea initiala de obtinere a adatelor experimentale.

Omogenitatea constituie o proprietate de bază pe care trebuie să o aibă clasele. Se spune că o clasă este omogenă dacă, pentru unităţile care fac parte din ea, variabila de grupare înregistrează variaţii nesemnificative.

Problemele care apar in impartirea pe clase a datelor sunt:
- determinarea lungimii intervalelor - de lungime egale
  - de lungimi diferite
- stabilirea formei de scriere a acestor intervale

Stabilirea numărului de intervale de variaţie trebuie să asigure satisfacerea următoarelor condiţii:

- sa nu existe pierderi de informatie prin grupare, evitarea divizarii excesive a populatiei

- media aritmetică a fiecărei grupe să fie cât mai aproape de centrul intervalului de variaţie respectiv
- să nu existe grupe vide
- reprezentarea grafică a seriei rezultate să permită conturarea unei regularităţi a fenomenului de studiat din cadrul populaţiei. Trebuie remarcat că acest lucru nu este posibil nici în cazul unui număr mic de intervale deoarece se pierd prea multe date, nici în cazul unui număr prea mare de intervale, populaţia fărâmiţându-se prea tare.

Statisticianul american H.A. Struges a stabilit, in cazul distributiilor normale, următoarea expresie de calcul a lungimii clasei, in cazul in care setul de date se imparte intr-un numar egal de clase:

$$l_x = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \ \lg N}$$

(numarul de clase in care se imparte populatia de date este1+3,322 Lg(n). Seria de intervale de lungime egală după care se imparte sirul de date este:

$$X: \begin{pmatrix} [x_{\min}; (x_{\min} + l_x)) & \dots & [x_{\min} + (k-1)l_x; (x_{\min} + kl_x)) & \dots & [x_{\min} + (R-1)l_x; (x_{\min} + Rl_x)) \\ N_1 & & N_k & & N_R \end{pmatrix}$$

Numeroase sunt cazurile practice în care studiul unei populaţii în raport cu o variabilă sau mai multe, presupune împărţirea domeniilor de variaţie ale acestora în intervale de lungime neegală. În asemenea cazuri nu există o relaţie de calcul în acest sens. Stabilirea intervalelor de variaţie se face în directă legătură cu variaţia variabilelor şi distribuirea unităţilor în raport cu acestea.

Dacă la baza seriei în cauză stau două sau mai multe variabile calitative sau cantitative atunci clasele se stabilesc în raport cu fiecare din variabilele considerate prin stările acestora, in acest caz avem de-a face cu serii bidimensionale sau multidimensionale.

Nu este recomandat ca numărul variabilelor în raport cu care se studiază populaţia să fie prea mare, deoarece aceasta duce la o divizare exagerată a populaţiei pierzându-se din vedere aspectele principale, generand pierderi de informatii.

După ce clasele au fost definite, are loc repartizarea unităţilor populaţiei în clasele respective, folosind în acest scop un algoritm adecvat.

Pentru elaborarea şi prezentarea seriilor statistice se apelează la pachete de programe statistice cum ar fi: S.P.S.S. (Statistical Package for the Social Sciences), STATISTICA, S.A.S. (Statistical Analysis System), STATGRAPHICS, etc.

3.2.3. Constructia histogramelor

Histogramele prezinta datele aparute cat si frecventele lor de aparitie si pot fi reprezentate si prin diagrame in care frecventele de aparitie ale datelor individuale se insumeaza. Pentru rezumarea datelor continue (reprezentări grafice), este necesară uneori gruparea datelor. Aceasta se realizează prin divizarea domeniului in care au aparut valorile individuale în intervale disjuncte, numite intervale de clasă (sau intervale de grupare), astfel încât fiecare valoare să fie conţinută într-un interval de clasă.

Exemple:

- seria 1 pentru sirul de date: 1,2,3,4,4,4, 5,6,7,8,9
- seria 2 pentru sirul de date: 1,2,3,4,4,5,6,6,6,6,7,8,9

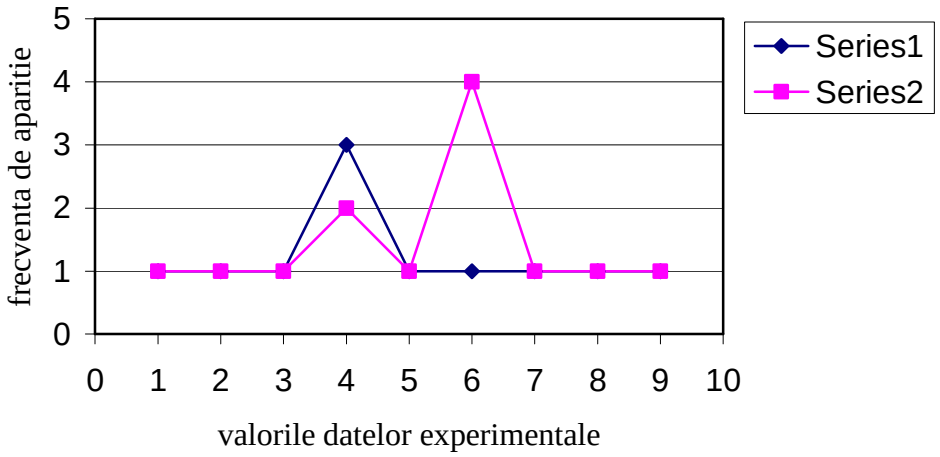

Figura nr. 3.2.3.1. Histograma datelor experimentale.

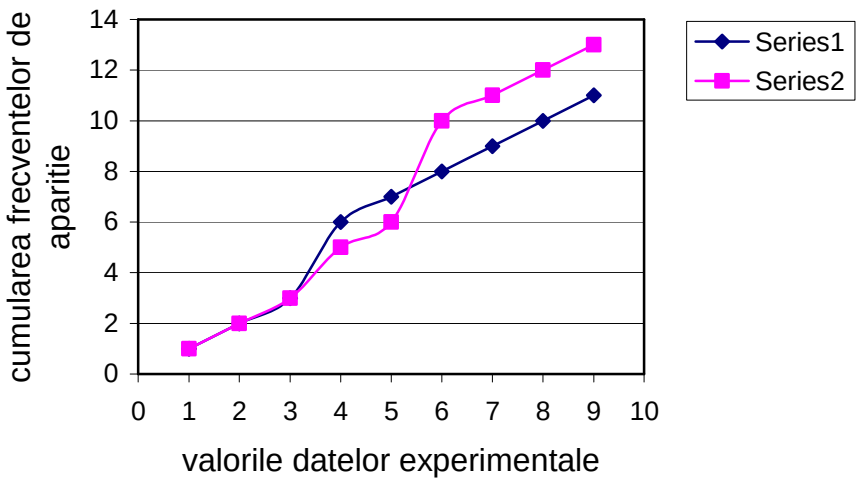

Figura nr. 3.2.3.2. Cumularea frecventelor pentru valorile masurate.

Prezentarea histogramelor sub forma profilului dreptunghiurilor se realizeaza prin construirea a cate unui dreptunghi pe fiecare interval (sau clasa) a carui inaltime este proportionala cu frecventa clasei (sau in cazul in care intr-o anumita clasa avem date care au frecvente diferite inaltimea dreptunghiului va fi dat de frecventa absoluta a clasei (a datelor din interval). Pentru cazul unei distributii normale a datelor populatiei, histograma prezinta aspectul clopotului lui Gauss. Pentru aceasta varianta de distributie, aspectul de clopot este dat de faptul ca se considera ca intr-o distributie normala numarul indivizilor clasei va fi cu atat mai mare cu cat clasa care-i contine este mai apropiata de clasa care contine valoarea medie aritmetica.

Cercurile de structura permit vizualizarea structurii datelor prin reprezentarea sub forma de sectoare de cerc a submultimilor populatiei de date. Aranjarea datelor in submultimi se poate realiza functie de o serie de caracteristici. Unghiul unui sector de cerc care caracterizeaza o anumita submultime va fi dat de procentul datelor cuprinse de submultimea respective care apartin polulatiei.