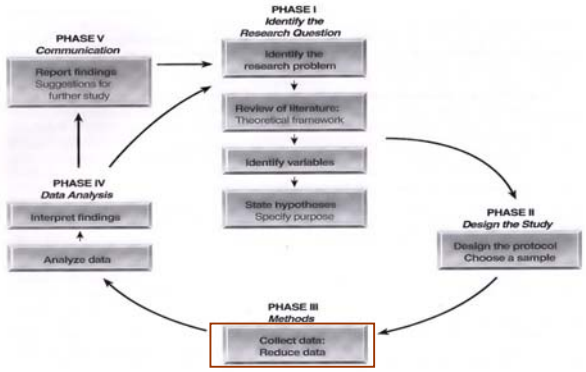
Culegerea şi stocarea datelor
Analiza datelor I

"To consult the statistician after an experiment is finished is often merely to ask hin to conduct a *post mortem* examination. He can perhaps say what the experiment died of."

Presidential Address to the First Indian Statistical Congress, 1938

Cuprins

- **Culegerea şi stocarea datelor:**
  - Baze de date Excel
  - Tipuri de date şi formate
- **Analiza datelor I**



Culegerea şi stocarea datelor

Scop:

- Organizarea datelor în formatul care să permită sumarizarea şi prelucrarea acestora

Aplicabilitate pentru teză:

- Stocarea în format electronic va permite sumarizarea şi prelucrarea statistică

Culegerea datelor

- Constituie un element esenţial al cercetării
- De realizează pe fişe de culegere a datelor:
  - pe suport de hârtie
  - electronic
- Pentru prelucrarea datelor e necesar suportul electronic
- Conţinutul fişei de culegere a datelor:
  - Pentru fiecare pacient toate caracteristicile urmărite
    - exemplu: numele, vârsta, greutatea, înălţimea, tensiunea arterială sistolică, ...

Gestiunea datelor medicale cu Microsoft Excel

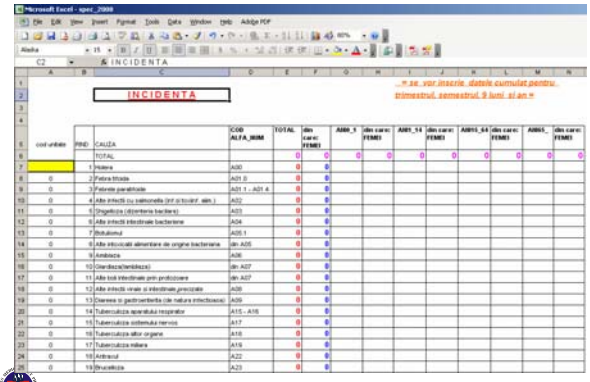
- Pachet de programe destinat:
  - tratării datelor în formă tabelară
  - prelucrării statistice şi reprezentării grafice a informaţiei conţinută în aceste tabele
  - este dotat cu o funcţie pentru tratarea datelor tabelei ca baza de date

Baza de date Excel

- Regiune compactă de date care ocupă ca suprafaţă cel puţin două rânduri (denumite articole) şi două coloane (denumite câmpuri) adiacente, coloane consecutive şi rânduri ne-consecutive sau rânduri consecutive şi coloane ne-consecutive
- Prima linie dintr-o astfel de regiune poartă numele de antet şi conţine denumiri de câmpuri

Baza de date Excel

- Dimensiunea maximă a unei baze de date Excel este dată de:
  - dimensiunea unei foi de calcul: 256 coloane × 65536 rânduri
  - numărul maxim de foi de calcul: 256



Manipularea foilor de calcul

- **Clic dreapta de mouse pe foaia de calcul şi:**

[Insert]	**adăugarea unei noi foi de calcul**
[Delete]	**ştergerea foii de calcul selectată**
[Rename]	**schimbarea denumirii foii de calcul**
[Move or Copy]	**schimbarea ordinii foii de calcul selectate sau copierea acesteia**
[Select All Sheets]	**selectarea tuturor foilor de calcul**
[Format - Sheet - Background]	**definirea background-ului pe pagina de lucru**
[Format - Sheet - Hide]	**ascunderea foii de calcul selectate**

Editarea foilor de calcul

- Inserare unui rând sau a unei coloane:
  - clic dreapta de mouse pe eticheta de rând sau coloană şi activarea opţiunii [Insert]
- Ştergere:
  - clic dreapta de mouse pe eticheta de rând sau coloană şi activarea opţiunii [Delete]
- Definirea antetelor şi a notelor de subsol: [View - Header and Footer] sau [View - PageSetup] – vizibile la Print Preview şi la printare

Editarea foilor de calcul

- Selectarea:
  - unei celule: clic stânga pe celulă
  - unei coloane: clic stânga pe eticheta coloanei (idem şi pentru selectarea unui rând, de această dată clic stânga pe eticheta rândului)
  - unui domeniu aleator de celule: clic stânga pe o celulă, simultan activarea tastei CTRL în combinaţie cu clic stânga celulele dorite a fi selectate

Editarea foilor de calcul

- Căutarea şi înlocuirea datelor din celule:
  - [Edit - Find] pentru căutare
  - [Edit - Find - Replace] pentru înlocuire
- Copiere: [Edit - Copy - Paste]
- Mutarea: [Edit - Cut - Paste]
- Ştergerea: [Edit - Cut]

  Atenţie! Este necesară selectarea prealabilă a datelor!

[Edit - Paste Speciale...]

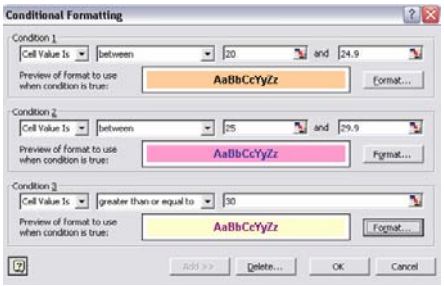
[Edit - Paste Speciale...]	[Edit - Paste Speciale...]
Conţinutul şi toate proprietăţile acestuia	All
Formulele de calcul	Fomulas
Valorile selecţiei fără formulele care au generat aceste valori	Values
Proprietăţile de formatare	Formats
Comentariile anexate celulelor selectate	Comments
Regulile de validare a datelor din selecţie	Validation
Totul cu excepţia formatărilor de chenare	All except borders
Doar formatările de lăţime a coloanelor	Column widths
Formulele din celulele selectate şi formatarea datelor	Formulas and number formats
Valorile din celulele selectate şi formatarea datelor	Values and number formats

Formatare celule: [Format - Cells ...]

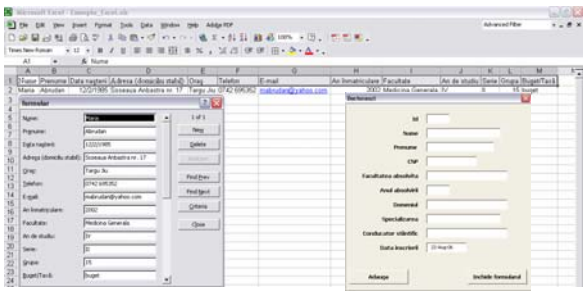
Denumire	Tip	Observaţii
Number	numeric	aliniate implicit la dreapta
Date	calendaristic	aliniate implicit la dreapta
Time	timp	aliniat implicit la dreapta
Text	non numeric	aliniate implicit la stânga
Scientific	1.0E-01	aliniate implicit la

- **Transformarea unui grup de celule într-o singură celulă: [Format - Cells ... - Alignment - Merge cells]**

Formatale celule: [Format - Conditional Formatting...]



Crearea unei baze de date Excel - Formulare



Validarea datelor dintr-o bază de date Excel

- [Data - Validation...]:
  - crearea unei liste de opţiuni
  - limitarea datelor de intrare la un anumit tip sau de o anumită mărime

Validarea datelor dintr-o bază de date Excel

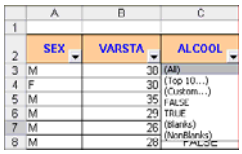
- **Criteriile de validare:**

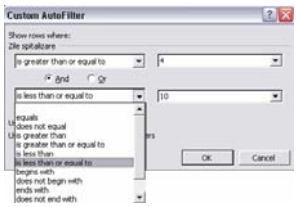
Criteriul	Semnificaţie
Whole Number	În funcţie de operatorul ales: incluziune (minim - maxim), excluziune (minim - maxim), egal cu, diferit de, mai mare decât, mai mic decât, mai mare sau egal cu, mai mic sau egal cu
Decimal	Includerea sau excluderea unui şir de numere care îndeplinesc condiţia operatorului (vezi operatorii anteriori)
List	Includerea unei liste de opţiuni sau legătura cu o listă de opţiuni creată anterior
Date	Includerea sau excluderea de date tip calendaristic (dd/mm/yyyy) în funcţie de operatorul ales
Time	Includerea sau excluderea de date tip timp (hh:mm:ss AM/PM) în funcţie de operatorul ales
Text Length	Specificarea lungimii textului prin folosirea operatorilor
Custom	Folosirea formulelor de validare create de către utilizator

Operaţiunea de filtrare

- Afişarea conţinutului care satisface criteriile de filtrare impuse de utilizator, articolele care nu satisfac criteriile fiind ascunse
- Tipuri:
  - auto-filtrarea [Data - Filter - AutoFilter]
  - filtrarea avansată [Data - Filter - Advanced Filter...]

Operaţiunea de filtrare





Operatori în auto-filtrare

Operator	Denumire variabilă	Criteriu	Ce localizează în baza de date? înregistrări în care
egal	SEX	M	sexul este masculin (M)
ne-egal	ALCOOL	FALSE	variabila Alcool are valoarea TRUE
mai mare	VARSTA	40	persoanele au vârsta mai mare de 40 de ani
mai mare sau egal	DATA ANGAJĂRII	2/11/1987	data angajării este 2 octombrie 1987 (inclusiv)
mai mic	IMC	21	indicele de masă corporală are valoare mai mică de 21
mai mic sau egal	SALAR	1000	valorile din câmpul salar sunt mai mici de 1000
începe cu	TIPUL LAPTELUI	mi	valorile corespunzătoare câmpului TIPUL LAPTELUI încep cu mi
nu începe cu	TIPUL LAPTELUI	mi	valorile corespunzătoare câmpului TIPUL LAPTELUI nu încep cu mi
se termină cu	ORAŞ	na	valorile corespunzătoare câmpului ORAŞ care se termină cu na
nu se termină cu	ORAŞ	na	valorile corespunzătoare câmpului ORAŞ care nu se termină cu na
conţine	ETNIE	ag	valorile corespunzătoare câmpului ETNIE care conţin literele ag
nu conţine	ETNIE	ag	valorile corespunzătoare câmpului ETNIE care nu conţin literele ag

Aplicaţii biomedicale

- Management medical
- Gestiunea datelor medicale
- Gestiunea în domeniul medical
  - Salarizare
  - Contabilitate
  - Facturarea serviciilor medicale

De reţinut!
Gestiunea datelor cu Microsoft Excel

- Ieftin şi uşor de implementat!
- Pentru utilizare e nevoie de abilităţi minime de lucru cu calculatorul.
- Permite organizarea datelor după placul utilizatorului.
- Permite importul de date în alte programe:
  - De gestiune a informaţiilor: Microsoft Acess
  - De prelucrare statistică: EpiInfo

Culegerea datelor

- Fişa de culegere a datelor în format electronic:
  - Fişier Microsoft Excel
  - Primul rând conţine obligatoriu denumirea variabilei şi unitatea de măsură (dacă este cazul)
  - Abrevierile folosite se introduc ca şi comentariu la denumirea variabilei
  - În fiecare rând se introduc datele corespunzătoare unui singur pacient
  - Dacă nu avem toate datele specificăm în celulele corespunzătoare "lipsă"

Culegerea datelor

- Fişa de culegere a datelor în format electronic:
  - Se culeg **doar** date primare (măsurate sau observate)
  - Nu se culeg date derivate din date perimare
    - Exemplu:
      - Dată primară = înălţime, greutate
      - Dată secundară = indicele de masă corporală
    - Exemplu:
      - Dată primară = tensiunea arterială sistolică, tensiunea arterială diastolică
      - Dată secundară = status hipertensiv – hipotensiv - normotensiv

De reţinut! Culegerea datelor

- Culegerea datelor trebuie să se realizeze în conformitate cu protocolul de cercetare.
- Pentru sumarizarea şi analiza statistică a datelor este necesară stocarea acestora în format electronic.
- Se culeg doar date primare! Datele secundare se obţin în urma prelucrării datelor primare.

Analiza datelor

Scop:

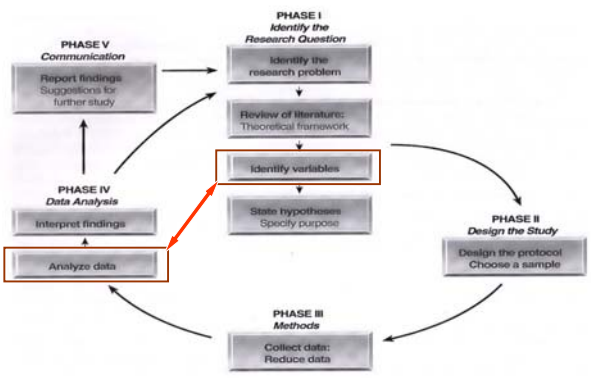
- Sumarizarea datelor
- Verificarea ipotezelor de cercetare

Finalitate:

- Generarea de noi ipoteze de cercetare

Aplicabilitate pentru teză:

- Capitolul Rezultate



De ce analiza statistică?

- 2 scopuri:
  - Descriptiv (statistica descriptivă):
    - Modalităţi de sumarizare a caracteristicilor importante ale unui set de date medicale
  - Inferenţial (statistica inferenţială):
    - Cum (şi când) generalizăm rezultatele obţinute pe un eşantion la populaţia generală

"There are three kinds of lies: lies, damned lies, and statistics."
Benjamin Disraeli

- Popularizată în SUA de Mark Twain
  - "...statement refers to the persuasive power of numbers, the use of statistics to bolster weak arguments, and the tendency of people to disparage statistics that do not support their positions..."

SUMARIZAREA TABELARĂ ŞI REPREZENTAREA GRAFICĂ A DATELOR

Statistica descriptivă

Conţinut

- **Principii de sumarizare tabelară**
- **Principii de reprezentare grafică**
- **Sumarizarea tabelară şi/sau reprezentarea grafică a datelor:**
  - Atribut (calitative): o variabilă
  - Atribut (calitative): două variabile
  - Numerice (cantitative): o variabilă
  - Numerice (cantitative): două variabile

Principii de sumarizare tabelară

1. Simple: de preferat 2/3 tabele mai mici în loc de unul încărcat
2. Informative prin ele însele
   - Abrevieri sau simboluri explicate la subsolul tabelului
   - Etichete de rând şi coloană
   - Unităţi de măsură
   - Titlul: ce? când? Unde?
   - Linii şi/sau coloane de sinteză (total)
3. Dacă datele nu sunt originale trebuie să se menţioneze sursa lor într-o notă de subsol

Principii de reprezentare grafică

- Orice reprezentare grafică trebui să aibă:
  - Titlul
  - Definirea axelor
  - Unităţi de măsură pentru fiecare axă (dacă este cazul)
  - Legendă (dacă este cazul)
  - O reprezentare grafică trebuie să se "înţeleagă" singură!
    - Fără a se citi textul!!!

Principii de reprezentare grafică

- Scopul unei reprezentări grafice este de a transmite o informaţie
- Când construim o reprezentare grafică trebuie să răspundem la întrebarea: Care este scopul acestei reprezentări?
- Datele trebuie reprezentate grafic în aşa fel încât să fie utile în înţelegerea fenomenului clinic
- Atenţie la compoziţia culorilor (nu puneţi fundaluri colorate) şi la dimensiunea caracterelor!

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

- Se ordonează datele crescător
- Se determine frecvenţa fiecărei valori
- Se includ valorile distincte şi frecvenţele într-un tabel pe două coloane:
  - Frecvenţa absolută (numărul de cazuri care îndeplinesc criteriul)
  - Frecvenţa relativă = raportul dintre frecvenţa absolută şi volumul eşantionului/populaţiei (simbol = n). Valorile se pot prezenta şi procentual.

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

- Se pot alcătui tabele de frecvenţe cu mai multe coloane care să cuprindă:
  - frecvenţe absolute
  - frecvenţe absolute cumulate crescător / descrescător
  - frecvenţe relative
  - frecvenţe relative cumulate crescător / descrescător
- Microsoft Excel:
  - funcţia COUNTIF
  - Tabele Pivot [Data - Pivot Table and Pivot Chart Report ...]

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

Frecvenţa absolută | Frecvenţa relativă

Diagnostic	Nr. persoane	Procent (%)
Asfixia la naştere	527	26,1
Traumatisme obstreticale	92	4,6
Stare septică	7	0,3
Pneumonie	181	9,0
Diaree	8	0,4
Malformaţii congenitale	598	29,6
Alte cauze	606	30,0
Total	2019	100

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

Suma frecvenţelor relative ale tuturor valorilor seriei care sunt mai mici sau egale decât x/n

Suma frecvenţelor absolute ale tuturor valorilor seriei care sunt mai mici sau egale decât x

Diagnostic	f_a	f_r	f_a cumulat ↑	f_r cumulat ↑
Asfixia la naştere	527	26.10	527	26.10
Traumatisme obstreticale	92	4.56	619	30.66
Stare septică	7	0.35	626	31.01
Pneumonie	181	8.96	807	39.97
Diaree	8	0.40	815	40.37
Malformaţii congenitale	598	29.62	1413	69.99
Alte cauze	606	30.01	2019	100
Total	2019	100		

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

- **Pentru seria statistică 5, 6, 7, 7, 8, 8, 5, 7, 8, 7 cărei din valorile de mai jos îi corespunde frecvenţa relativă cumulată crescător de 0.7:**

A. 8
B. 6
C. 5
D. 7
Nici un răspuns nu este corect

Variabile calitative: 1 variabilă
Tabelul de frecvenţă

- **Pentru seria statistică 5, 6, 7, 7, 8, 8, 5, 7, 8, 7 cărei din valorile de mai jos îi corespunde frecvenţa relativă cumulată crescător de 0.7?**

Valoare	f_a	f_r	f_a cc	f_r cc
5	2	0,20	2	0,20
6	1	0,10	3	0,30
7	4	0,40	7	0,70
8	3	0,30	10	1
Total	10	1		

Variabile calitative: 2 variabile
Tabelul de contingenţă

	TBC=da	TBC=nu	Total
sex=F	2	10	12
sex=M	24	54	78
Total	26	64	90

Variabile calitative: n variabile
Tabel de frecvenţă

Tabelul 1. Distribuţia patologiilor pulmonare asociate silicozei

	BrC	BPOC	Emfizem	CPC	TBC	Total
silicoza grad I	12	20	0	0	14	46
silicoza grad I/II	1	5	1	1	1	9
silicoza grad II	3	7	1	1	7	19
silicoza grad II/III	0	1	0	0	0	1
silicoza grad III	0	3	0	0	4	7
Total	16	36	2	2	26	82

BrC = bronşită cronică; BPOC = bronho-pneumopatie cronică obstructuvă;
CPC = cord pulmonar cronic; TBC = tuberculoză pulmonară

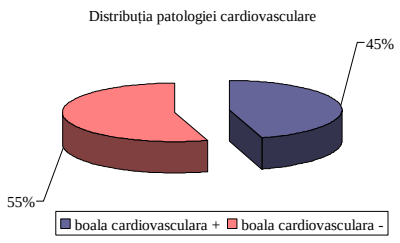
Variabile cantitative: 1 variabilă
Tabele pe clase de frecvenţă

Greutate (g)	f_a	f_r	f_r cumulată ↑
(2800 – 3200]	151	18,60	18,60
(3200 – 3400]	299	36,82	55,42
(3400 – 3600]	300	36,95	92,37
(3600 – 3800]	0	0,00	92,37
(3800 – 4000]	62	7,64	100
Total	812	100	

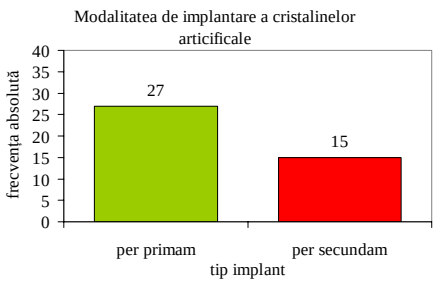
Reprezentarea grafică: 1 variabilă
Plăcinta (PIE)

- Variabile calitative sau cantitative. Dacă este cantitativă trebuie să fie clase de frecvenţe.
- Se foloseşte pentru a reprezenta frecvenţe absolute sau relative:
  - Vizualizarea prevalenţei relative a unui fenomen de sănătate
- Datele se culeg ca frecvenţe absolute

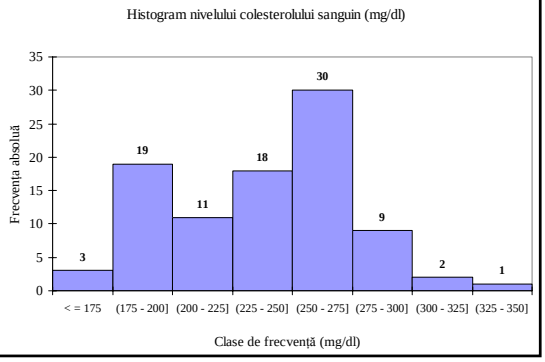
Reprezentarea grafică: 1 variabilă
Plăcinta (PIE)



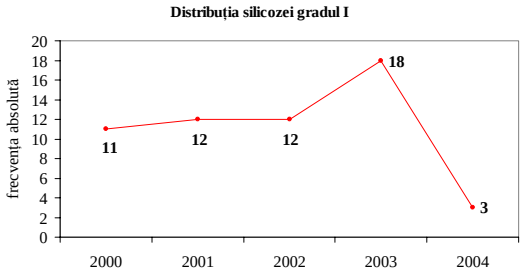
Reprezentarea grafică: 1 variabilă
Coloane (COLUMN)



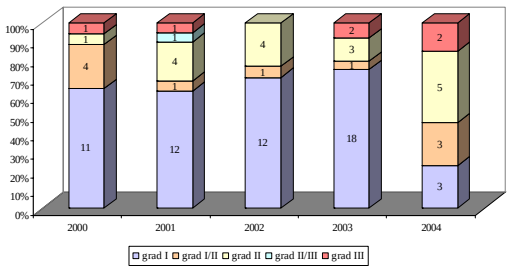
Reprezentarea grafică: 1 variabilă
Histograma (COLUMN)



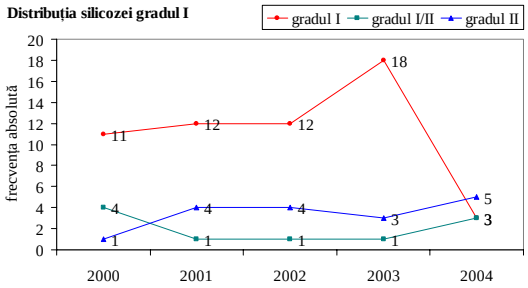
Reprezentarea grafică: 1 variabilă
Linie (LINE)



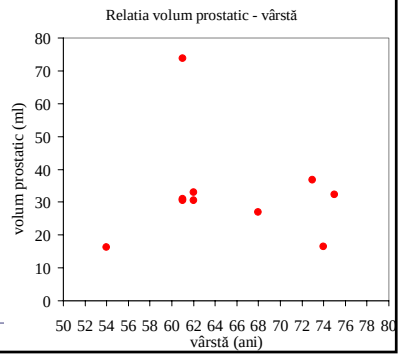
Reprezentarea grafică: 2 variabile
calitative - Coloane (COLUMN)



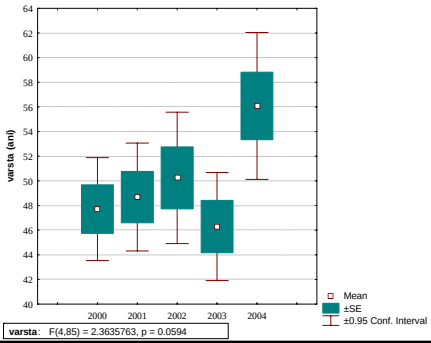
Reprezentarea grafică: 2 variabile
calitative – Linie (LINE)



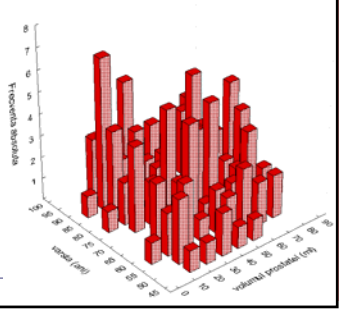
Reprezentarea grafică: 2 variabile
Norul de puncte (SCATTER)



Reprezentarea grafică: 2 variabile
Box-and-whisker



Reprezentarea grafică: 2 variabile
Coloane (Histograme) bi-dimensionale



De reţinut!

- Sumarizarea tabelară şi reprezentarea grafică se realizează cu scopul transmiterii de informaţii.
- În realizarea lor trebuie să ţinem cont de scop (putem distrage atenţia privitorului de la ceea ce dorim să transmitem).
- Asiguraţi-vă că aveţi titluri informative, denumiri de rânduri şi coloane; totaluri pe rânduri şi/sau coloane.

De reţinut!

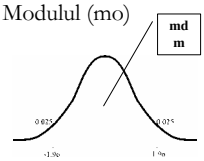
- **Asiguraţi-vă că axele au denumiri şi unităţi de măsură.**
- **Minimizaţi numărul de culori.**
- **Evitaţi reprezentările grafice 2D şi 3D:**
  - **Ceea ce se reprezintă ocupă o dimensiune mai mică din grafic.**
  - **Poate distorsiona imaginea în compararea a două distribuţii.**

Parametrii statistici descriptivi

Date numerice, o variabilă

Măsuri ale tendinţei centrale

- Cvartile
- Media aritmetică (m)
- Mediana (md)
- Modulul (mo)



Curba lui Gauss (distribuţia normală)

Măsuri ale dispersiei sau variabilităţii

- Amplitudinea
- Varianţa
- Eroarea standard a mediei

Caracteristici ale unei serii de date în jurul valorilor de mijloc

Date numerice, o variabilă: cuartile

- **Divizăm setul de valori în 100 părţi (maxim - minim): Percentile**
  - Percentila de 90% = valoarea sub care regăsim 90% din valorile seriei **[=PERCENTILE(argument)]**
  - **[Insert – Function…]**
- **Divizăm scala în 10 părţi: decile**
- **Divizăm scala în 4 părţi: cvartile [=QUARTILE(argument)]**
- Permit aprecierea distribuţiei datelor analizate
  - Exemplu: fie Q_1 (1/4), Q_2 (1/2) şi Q_3 (3/4) primele 3 cvartile. Dacă $Q_2 - Q_1 \approx Q_3 - Q_2$ distribuţia datelor este aproximativ simetrică. Dacă nu distribuţia este asimetrică (spre dreapta sau spre stânga)

Date numerice, o variabilă: cuantile

- Cvartila 1: 3150 (Q_1)
- Cvartial 2: 3300 (Q_2)
- Cvartila 3: 3450 (Q_3)

- Metoda de calcul: funcţia QUARTILE
- Microsoft Excel
- Argumentele funcţiei:
  - Array: selectăm coloane unde avem date
  - Quart: 1 (cvartial 1), 2 (cvartial a doua), 3 (cvartila a treia)

Greutate la naştere (g)
2800
3100
3200
3300
3400
3500
3800

$Q_2 - Q_1 \approx Q_3 - Q_2$

$150 = 150$

Date numerice, o variabilă: Media aritmetică

- **Simboluri standard (universal recunoscute):**
  - **μ** = media aritmetică a unei populaţii **[=AVERAGE(argument)]**
  - **m** = media aritmetică a unui eşantion ($\bar{x}$)
    - Au aceeaşi formulă de calcul
    - Diferenţa constă în semnificaţia lui ***n*** (volumul populaţiei sau volumul eşantionului)
  - ∑ = sumă **[=SUM(argument)]**
    - Exemplu: fie variabila zile de supravieţuire (notată cu 'x') cu trei valori 1, 2, 4. $\sum x = 1+2+4 = 8$
    - $m(x) = \sum x / n = 8/3$

Date numerice, o variabilă: Mediana

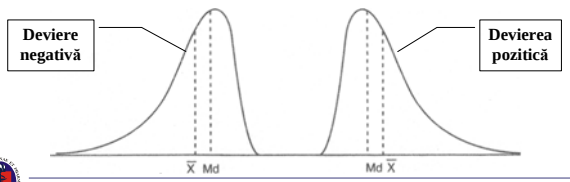
- **Nu are simbol standard (uneori se notează cu 'md')**
- Calculare **(= MEDIAN(argument))**
  - Aşezăm valorile datelor în ordine crescătoare
  - Mediana este egală cu valoarea datei din mijlocul seriei dacă volumul eşantionului este impar
  - Valoarea medianei este egală cu media aritmetică a celor două valori din mijlocul seriei dacă volumul eşantionului este par

Date numerice, o variabilă: Mediana

- **Exemplu: calcularea medianei**
  - Antigenul prostatic specific (simbol: PSA, unitate de măsură: ng/ml) pentru un eşantion de 10 pacienţi (volumul eşantionului n = 10) cu manifestări prostatice: 7,6; 4,1; 5,9; 9,0; 6,8; 8,0; 7,7; 4,4; 6,1; 7,9
  - aranjăm datele în ordine crescătoare:
    - 4,1; 4,4; 5,9; 6,1; 6,8; 7,6; 7,7; 7,9; 8,0; 9,0
    - n = 10 → md = (6,8 + 7,6)/2 = 7,2

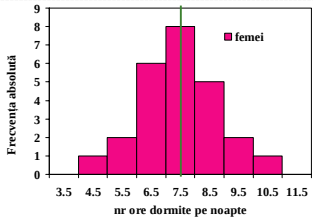
Date numerice, o variabilă: Media Aritmetică & Mediana

- Nu toate datele urmează o distribuţie normală:
  - Deviere negativă: coadă la stânga
    - md > m (m < md < mo)
  - Deviere pozitivă: coadă la dreaptă
    - m > md (mo < md < m)



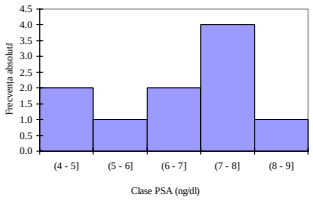
Măsuri de simetrie

- **Într-o distribuţie simetrică**
  media aritmetică = mediana = valoarea modală



Date numerice, o variabilă: Modulul

- **Nu are simbol standard (uneori se foloseşte 'mo')**
  - Valoarea cea mai frecventă a seriei
  - Pe graficul de tip bare: modulul este valoarea barei cu frecvenţa cea mai mare



De reţinut! Măsuri de centralitate

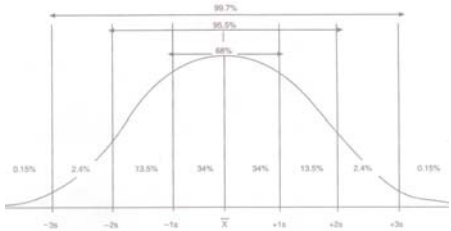
	+++	---
MODUL	▪ uşor de calculat ▪ utilă pentru datele nominale	▪ slabă stabilitatea de eşantionare
MEDINANA	nu e afectată de valorile extreme	▪ Într-o oarecare măsură slabă stabilitate de eşantionare
MEDIA	▪ stabilitate de eşantionare ▪ în legătură cu varianţa	▪ Nu este utilă pentru datele discrete ▪ E afectată de distribuţia asimetrică a datelor

Date numerice, o variabilă: Variaţia & Deviaţia Standard

- **Simbol standard varianţa populaţiei: σ^2**
  - Formula de calcul: $\sigma^2 = \sum(x- \mu)^2/n$
  - Alternativ: $\sigma^2 = [\sum(x^2- n\times\mu^2)]/n$
- **Simbol standard varianţa populaţiei: s^2**
  - Formula de calcul: $s^2 = [\sum(x^2- n\times m^2)]/(n-1)$
  - Exemplu:
    - $m_{PSA} = 6.75$
    - $s^2 = (74.1^2 + 4.4^2 + \ldots + 9.0^2 - 10\times6.75^2)/9 = 2.585$
- **Deviaţia standard: simbol s**
  - Formula $s = \sqrt{s^2}$ **[= SQRT(argument)]**
  - Exemplu: $s = \sqrt{2.585} = 1.6078$

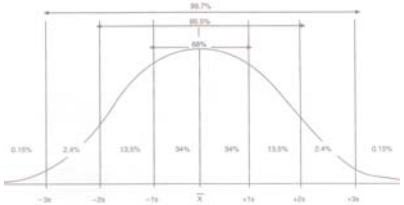
Date numerice, o variabilă: Media & Deviaţia standard

- **Procentul de cazuri cu 1, 2 şi 3 deviaţii standard ale mediei în distribuţia normală**



Date numerice, o variabilă: Media & Deviaţia standard

- Pe o curbă de distribuţie normală care este procentul cazurilor care au valori mai mici decât $m+2\times s^2$?
- Pe o curbă de distribuţie normală care este procentul cazurilor care au valori mai mici decât $m+1\times s^2$?



Date numerice, o variabilă: Eroarea standard a mediei

- **Simbol standard:**
  - Pentru media populaţiei: σ_m
  - Pentru media eşantionului: s_m
  - Formule de calcul:
    - Populaţie: $\sigma_m = \sigma/\sqrt{n}$
    - Eşantion: $s_m = s/\sqrt{n}$
  - Exemplu PSA:
    - $s_m = s/\sqrt{n} = 1,6078/\sqrt{10} = 0,5084$

Date numerice, două variabile: Covarianţa

- **Analiza dependenţei sau independenţei dintre două variabile măsurabile**
  - **Covarianţa:**
    - Când urmărim două variabile măsurabile care variază simultan, una în relaţie cu cealaltă
  - **Coeficientul de corelaţie:**
    - Cuantificarea legăturii dintre cele două variabile

Date numerice, două variabile: Covariaţia

- Necesită înregistrarea unor date perechi
- Formula de calcul:
  - Populaţie: $\sigma_{xy} = (xy - n \times \mu_x \times \mu_y)/n$
  - Eşantion: $s_{xy} = (\sum xy - n \times m_x \times m_y)/(n - 1)$
  - Exemplu:
    - $s_{xy} = (75 \times 32.3 + 68 \times 27.0 + \ldots 74 \times 16.4 - 10 \times 65.1 \times 32.73)/(10\text{-}1) = -10.6478$
  - Excel (doar pentru **populaţie**):
    - **[=COVAR(variabila1, variabila2)]**

Date numerice, două variabile: Covariaţia

- **Interpretare:**
  - Dacă o variabilă creşte odată cu cealaltă variabilă (aşa cum este tensiunea arterială sistolică şi diastolică) covarianţa este pozitivă şi valoarea acesteia este mare.
  - Dacă o variabilă creşte în timp ce cea de-a doua variabilă scade (cum este de exemplu vârsta şi volumul prostatei) covarianţa este negativă şi valoarea acesteia este mare.
  - Dacă creşterea sau descreşterea unei variabile nu este în legătură cealaltă variabilă de interes, valoarea covarianţei este mică.

Măsuri de împrăştiere: Coeficientul de variaţie

- **Interpretarea omogenităţii:**

Coeficient de variaţie (CV)	Interpretare: populaţia poate fi considerată
CV < 10%	omogenă
10% ≤ CV < 20%	relativ omogenă
20% ≤ CV < 30%	relativ eterogenă/relativ heterogenă
> 30%	eterogenă/heterogenă

Date numerice, două variabile: Coeficientul de corelaţie

- Formula de calcul:
  - Populaţie: $\varrho_{xy} = \sigma_{xy}/(\sigma_x \times \sigma_y)$
  - Eşantion: $r_{xy} = s_{xy}/(s_x \times s_y)$
- Interpretare:
  - Dacă o variabilă în relaţie perfect lineară directă cu cea de-a doua variabilă coeficientul de corelaţie ia valoarea +1 (variabila 1 ↑ + variabila 2 ↑)/-1 (variabila 1 ↑ + variabila 2 ↓ sau variabila 1 ↓ + variabila 2 ↑)
  - Regulile lui Colton

Date numerice, două variabile: Coeficientul de corelaţie

- Interpretare: Regulile lui COLTON:
  - un coeficient de corelaţie de la -0.25 la 0,25 înseamnă o corelaţie slabă sau nulă,
  - un coeficient de corelaţie de la 0.25 la 0.50 (sau de la -0.25 la -0.50) înseamnă un grad de asociere acceptabil
  - un coeficient de corelaţie de la 0.5 la 0.75 (sau de la -0.5 la -0.75) înseamnă o corelaţie moderată spre bună
  - un coeficient de corelaţie mai mare decât 0.75 (sau mai mic decât -0.75) înseamnă o foarte bună asociere sau corelaţie

Date numerice, două variabile: Coeficientul de corelaţie

- Exemplu:
  - $s_{x(vârsta)}$=6,9992
  - $s_{y(volum)}$=15,9351
  - s_{xy}=-10,6478
  - r_{xy}=-10,6478/(6,9992×15,9351)=-0,0956
  - Acest rezultat ne spune că pe eşantionul studiat format din 10 pacienţi, volumul prostatei tinde să descrească odată cu creşterea vârstei, dar relaţia dintre cele două variabile este foarte slabă

Sumarizarea şi analiza datelor

Sumarizarea şi analiza datelor

- **Generalizarea rezultatelor obţinute pe eşantion asupra populaţiei**

1. Intervalul de confidenţă (pentru medie, pentru frecvenţe)
2. Statistică inferenţială

Sumarizarea şi analiza datelor: intervalul de confidenţă

- **Definiţie. Scop**
- **Interpretare**
- **Intervalul de încredere pentru medie**
- **Intervalul de încredere pentru frecvenţă**

Sumarizarea şi analiza datelor: intervalul de confidenţă
De ce intervalul de încredere?

- **Estimarea punctuală**
  - **= o valoare pentru parametrul teoretic estimat**
  - **Influenţată de fluctuaţiilor de eşantionare**
  - **poate fi la o mare distanţă de valoarea reală a parametrului estimat**
- **Este recomandabil să se estimeze un parametru teoretic nu printr-o singură valoare ci printr-un interval, numit** interval de încredere **(în care să se poată afirma că parametrul estimat se găseşte cu o probabilitate ridicată).**

Sumarizarea şi analiza datelor: intervalul de confidenţă
Definiţie

- Un şi de valori al unui estimator de interes calculat astfel încât pentru o probabilitate de eroare aleasă să includă valorile adevărate ale variabilei.
- P[valoarea critică inferioară < estimatorul < valoarea critică superioară] = 1-α
  - unde α = nivelul de semnificaţie
- Intervalul definit de valorile critice va cuprinde estimatorul populaţiei cu o probabilitate de 1-α
- Se aplică în cazul variabilelor distribuite normal!

Sumarizarea şi analiza datelor: intervalul de confidenţă Interpretare

- Dacă intervalul de încredere pentru diferenţa dintre o medie observată şi una teoretică cuprinde valoarea 0, datele sunt compatibile cu o diferenţă a mediei populaţiei egală cu 0.
- Dacă intervalul de încredere pentru diferenţa dintre o medie observată şi una teoretică nu cuprinde valoarea 0, datele nu sunt compatibile cu egalitatea mediilor populaţiei.

Sumarizarea şi analiza datelor: intervalul de confidenţă

- **Se calculează în funcţie de:**
  - **Talia eşantionului sau a populaţiei**
  - **Variabila de studiat (calitativă, cantitativă)**
- **Formula de calcul cuprinde 2 părţi:**
  - **Un estimator al calităţii eşantionului pe baza căruia estimatorul populaţiei s-a calculat (eroarea standard)**
  - **Gradul de încredere (confidenţă) al intervalului specificat (scorul Z_α)**
- **Cel mai frecvent utilizat este intervalul de încredere pentru** medie

Sumarizarea şi analiza datelor: intervalul de confidenţă Intervalul de încredere pentru medie

- **Eroarea standard a mediei este egală cu deviaţia standard împărţită la radicalul volumului eşantionului**
  - **Dacă deviaţia standard este mare, şansa de eroare în estimator este mare**
  - **Dacă volumul eşantionului este mare, şansa erorii în estimator este mică.**

$$\left[\overline{X} - Z_\alpha \frac{s}{\sqrt{n}}, \overline{X} + Z_\alpha \frac{s}{\sqrt{n}}\right]$$

Sumarizarea şi analiza datelor: intervalul de confidenţă Intervalul de încredere pentru medie

- **Scorul Z este scorul distribuţiei normale de medie 0 şi deviaţie standard de 1. Orice distribuţie poate fi transformată în scorul Z utilizând formula:**

$$Z = \left(X - \overline{X}\right)/s$$

  - **Scorul pozitiv este mai mare decât media**
  - **Scorul negativ este mai mic decât media**
  - Pentru intervalul de confidenţă de 95%: $Z_{5\%} = 1{,}96$
  - **Pentru intervalul de confidenţă de 99%: $Z_{1\%} = 2{,}58$**

$$\left[\overline{X} - Z_\alpha \frac{s}{\sqrt{n}}, \overline{X} + Z_\alpha \frac{s}{\sqrt{n}}\right]$$

Sumarizarea şi analiza datelor: intervalul de confidenţă Intervalul de încredere pentru medie

- **Media glicemiei la un eşantion de 121 pacienţi este de 105 iar variaţia de 36. Care este intervalul de încredere al mediei glicemiei în populaţia din care s-a extras eşantionul cu un prag de semnificaţie α=0,05, considerând că glicemia este normal distribuită şi pentru acest prag Z = 1,96.**

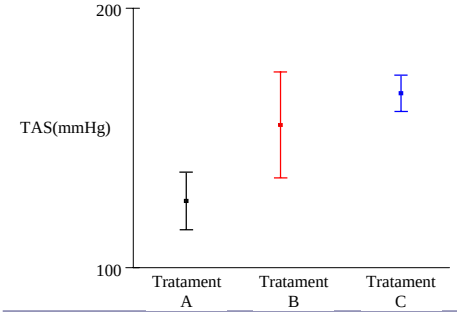
- **n = 121**
- **s^2 = 36**
- **s = 6**

$\overline{X} = 105$

$$\left[105 - 1{,}96\frac{6}{\sqrt{121}}; 105 + 1{,}96\frac{6}{\sqrt{121}}\right]$$

- **[105-1.07, 105+1.07]**
- **[103.93 – 106.07]**
- **[104-106]**

Sumarizarea şi analiza datelor: intervalul de confidenţă Compararea mediilor cu ajutorul intervalului de încredere



Sumarizarea şi analiza datelor: intervalul de confidenţă
Intervalul de încredere pentru frecvenţe

- Dacă n p > 10

$$\left[f - Z_\alpha \sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

Sumarizarea şi analiza datelor: intervalul de confidenţă
Intervalul de încredere pentru frecvenţe

- Suntem interesaţi în estimarea frecvenţei cancerului de sân la femeile între 50 şi 54 de ani care au antecedente familiale pozitive. Într-un studiu randomizat la care au participat 10000 de femei, s-a constatat că 400 dintre acestea au fost diagnosticate cu cancer de sân.
- Care este intervalul de încredere de 95% asociat frecvenţei observate?

- f = 400/10000 = 0.04

$$\left[0{,}04 - 1{,}96 \sqrt{\frac{0{,}04 \cdot 0{,}96}{10000}}; 0{,}04 + 1{,}96 \sqrt{\frac{0{,}04 \cdot 0{,}96}{10000}} \right]$$

- [0,04-0,004; 0,04+0,004]
- [0,036; 0,044]

$$\left[f - Z_\alpha \sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

Sumarizarea şi analiza datelor: intervalul de confidenţă

- **Rata şansei, respectiv riscul relativ**
  - Dacă intervalul de confidenţă al ratei şansei sau al riscului relativ conţine valoarea 1 → expunerea nu este factor de risc pentru patologia de interes

Sumarizarea şi analiza datelor: intervalul de confidenţă
De reţinut!

- Estimarea corectă a unui parametru statistic se face cu ajutorul intervalului de încredere.
- Intervalul de încredere depinde de volumul eşantionului şi de eroarea standard.
- Cu cât eroarea standard este mai mare cu atât intervalul de încredere este mai larg.
- Cu cât volumul eşantionului este mai mic cu atât intervalul de încredere este mai larg.