- Octav Onicescu:

  “Până în secolul XX, ştiinţa a trăit iluzia mecanicistă.”

  – singura dificultate părea rezolvarea de noi şi noi ecuaţii diferenţiale.
- Acum o sută de ani au început să fie puse sub semnul întrebării

  – determinismul cauzal imuabil

  – imaginea lumii ca mecanism uriaş

CUNOAŞTEREA STATISTICĂ

- Arsenalul de atac pentru cunoaştere se îmbogăţeşte cu
  - valori statistice tipice
  - frecvenţe
  - dispersii
  - corelaţii…
- Limitele cunoaşterii:
  - principiul nedeterminării al lui Heisenberg
  - teoria cuantelor (limite ale măsurărilor posibile)

NICHOLAS GEORGESCU-ROEGEN

- **Evoluția sferei noțiunii "statistică"**
- Teoria probabilităților – Blaise Pascal XVII
- "statistik": Gottfried Achenwall – 1749
  - statistica: stabilirea naturii informațiilor despre stat, a cadrului în care sunt expuse
    - a evoluat spre Economie Politică, Sociologie, Demografie
    - 1874, Rumelin: "Statistica nu este decât o metodă"
- Şcoala engleză: abordare calculatorie
  - Graunt: raportul 14/13 dintre numărul naşterilor de băieți şi de fete
  - "aritmicieni politici"
- "ştiința numărătorii", "ştiința numerelor mari", "ştiința valorilor medii". Istorie de 150 de ani.

DEX: “STATISTICA”

1. Evidență numerică referitoare la diverse fenomene; numărătoare
2. Culegere, prelucrare şi valorificare a unor date
3. **Ştiință care culege, sintetizează, descrie şi interpretează date referitoare la fenomene generale**
4. Statistică matematică: ramură a matematicii care elaborează noțiunile şi metodele folosite de Statistică (1., 2., 3.)
5. Teorie fizică ce urmăreşte şi descrie comportarea unui sistem format din numeroase particule.

FENOMENE COLECTIVE

- Fenomene naturale:
  - tipice : $P = U \cdot I$
  - colective (“generale”) : variația prețului petrolului, rate de schimb, variația incidenței unei boli etc.
- Fenomene tipice: în condiții identice sau similare, se produc în aceeaşi formă
    - caracteristice pentru anumite niveluri ale lumii anorganice
- Fenomene colective: nu se pot reproduce identic aproape niciodată
  - fenomene sociale
  - fenomene biologice
  - unele fenomene anorganice (meteorologie)

FENOMENE COLECTIVE

- Înainte de apariția științei, toate fenomenele păreau atipice (“colective” avant la lettre)
- Granița dintre fenomenele considerate tipice şi cele considerate colective s-a modificat mereu
  - prin identificarea, observarea şi măsurarea a noi factori ce influențează respectivele fenomene
  - analog graniței dintre rezolvabil şi nerezolvabil
- Există fenomene “absolut colective”
  - număr foarte mare de cauze / factori care le influețează
  - importanță variabilă a fiecărei cauze în diferite instanțieri ale aceluiaşi fenomen

METODA

- Fenomene tipice:
  - experiența de laborator, modelarea matematică
- Fenomene colective:
  - observarea (rareori repetabile prin experiment: meteorologie, economie, sociologie etc.)
  - trebuie observate multe repetări
    - pentru a distinge **tipicul** de **accidental**
- Metodă de studiu cu o altfel de modelare matematică. Noțiunile noi:
  - **variabilă aleatoare**
  - **lege stochastică**

LEGE STOCHASTICĂ

- Fenomene tipice: legi rigide

  "Spațiul parcurs este egal cu produsul dintre viteza de deplasare şi timpul de deplasare"

- Fenomene colective: legi stochastice

  "Din stejar, stejar răsare"

- Spiritul uman, prin abstractizare, tinde să rețină ce este tipic, general şi să ignore excepțiile.

STABILITATEA FRECVENȚELOR

- Cum se descoperă legi stochastice?
- Ce le face adevărate?
- Punctul comun al teoriei statisticii şi al realității:
- **<u>Axioma stabilității frecvențelor</u>**

  Dacă într-o serie de observații conținând $N_1, N_2, \ldots, N_p$ cazuri, obținute sub influența aceluiaşi complex de cauze, numărul de cazuri prezentând calitatea A este de $f_1, f_2, \ldots, f_p$, atunci raporturile $f_1/N_1, f_2/N_2, \ldots, f_p/N_p$ nu diferă prea mult între ele.

  – Frecvențele relative ale evenimentului A.

  – Bernoulli "Ars conjectandi", 1713: "demonstrație".

STABILITATEA FENOMENELOR COLECTIVE

- Nu doar la experimente artificiale (moneda)
- Halley: tabele de mortalitate.
  - Exemplu: din 100 000 de bărbați de 30 de ani, 698 vor deceda înainte de a avea 31 de ani
    - în **medie**, pe mai mulți ani
  - Societăți de asigurare: pariază implicit, prin suma asigurată, că raportul este acesta (de fapt, mai mare…)
    - Schimbarea dramatică a factorilor (război etc.) nu se asigură!
    - Factorii care influențează vitalitatea populației pot varia de la generație la generație, comportarea medie rămâne aproximativ aceeaşi.
- Raționamentul statistic operează cu noțiunea de **stabilitate**, nu de **constanță**.
    - stabilitatea are loc în jurul unei valori care poate varia în timp.

CAUZE ALE ABATERILOR DE LA LEGI STOCHASTICE

- Populațiile mici prezintă particularizări ale factorilor şi legea poate să nu fie respectată.
  - sondaj viciat de interogarea la telefon (1932).
- O lege stochastică este valabilă doar pentru populații ce prezintă toate variațiile de cazuri, fiecare cu proporția sa
  - "(sub)populații cu structură completă"
  - dualitatea dintre maximizarea / minimizarea sub-populației considerate
- Şi fenomenele colective evoluează
    - frecvențele sunt stabile pe o perioadă de stabilitate a factorilor
    - exempu: modificarea ratei mortalității la vârste peste 50.

DISPERSIA

- Stabilitatea frecvențelor nu se exprimă prin constanță a valorilor
  - valori “în jurul” celei aşteptate.
- **Dispersia**: abaterea valorilor reale de la valoarea medie.
- Dispersia poate caracteriza acolo unde media nu distinge
- Exemplu: temperaturi anuale medii egale în oraşe cu tipuri diferite de climă.

- Statistica descriptivă
  - sintetizarea şi prezentarea datelor
  - informează, aranjând datele pentru decizii
- Statistica inferențială (matematică): modele şi tehnici pentru
  - a obține **concluzii** din datele colectate
  - a face estimări de **parametri**
  - a verifica **ipoteze statistice**

FOLOSIRE IMPROPRIE A STATISTICII

- Un anumit mod de a prelucra statistic impune un anumit mod de a colecta datele – nu pot fi mixate
  - Date culese pentru a fi prelucrate într-un mod anumit nu pot fi prelucrate corect în alt mod.
- Colectare incorectă a datelor
- Analize statistice superficiale

- “Există 500 000 de analfabeți în România”
- “Venitul mediu anual pe cap de locuitor este de $1200”
- “”Speranța” de viață este de 70,1 ani”
- Se pot face deducții privind un singur locuitor?
- Evident nu, dar se reprezintă sintetic o întreagă populație.
- **Statistici**:
  - valori punctuale (numerice)
  - calculate folosind un **eşantion**
  - pot estima valorile corespunzătoare pentru **populație**.

ELEMENTELE DEFINITORII ALE UNUI STUDIU STATISTIC

- **Populație**: o colecție de obiecte (entități elementare, indivizi), posedând toate o anumită caracteristică.
  - finite / infinite; concrete / abstracte
  - definirea populației este esențială
- **Eşantion**: o submulțime a populației definite.
- **Atribut variabil**: o caracteristică ce prezintă valori ce pot diferi de la un individ la altul.
  - cantitative / calitative (sortabile / nesortabile).
- **Observație**: valoare a unui atribut variabil pentru un anumit individ.

EŞANTIONARE ALEATOARE

- Eşantionare subiectivă (exemple: selecția rocilor, pacienți pentru tratamente diferite)
- **<u>Eşantionare aleatoare</u>: fiecare individ din populație are aceeaşi şansă de a fi selectat.**
  - metoda selecției aleatoare (etichetarea tuturor indivizilor)
  - selecția sistematică (din k în k; periodicități?)
  - selecția stratificată (proporțiile straturilor)
  - selecția pe grupe: străzi, careuri de teren, circumscripții
  - selecția ierarhică: aleator județe → comune → străzi → persoane.

PROIECTAREA EXPERIMENTELOR

- **Nu** se caută structuri mici în date foarte numeroase.
- Prelucrarea statistică începe după analizarea atentă a datelor (“familiarizarea” cu datele).
- La dimensiunile actuale, *Data mining*
- Colectarea datelor: numai în conformitate cu analiza statistică ulterioară.
- Surse de erori - datele:
    - **pot lipsi** (cei cu durerile cele mai mari se tratează)
    - pot fi **greşit înregistrate** (cifre semnificative lipsă)
    - pot fi din **altă populație**: definire, eşantion ne-aleator

FRECVENȚĂ

- **Frecvența unei observații în eşantion**: numărul de apariții ale acelei observații (valori) în eşantion.
- **Frecvența relativă a unei observații în eşantion**: raportul dintre numărul de apariții ale observației în eşantion şi numărul total de observații (*dimensiunea eşantionului*)
- **Distribuția frecvențelor** (atribut variabil discret): mulțimea tuturor observațiilor distincte, împreună cu frecvențele lor relative în eşantion.
  - Exemplu:

fumat	Intens	Rar	Nu	Total
» F_abs	7149	2818	6563	16500
» f_rel	0.433	0.170	0.397	1.00

ATRIBUTE CONTINUE

- **Clasă interval**: un subinterval inclus între valorea minimă şi cea maximă.
- **Frecvența clasei interval**: numărul de observații ce aparțin clasei respective.
- **Distribuția frecvențelor** unui atribut variabil continuu: mulțimea claselor interval împreună cu frecvența relativă a fiecăreia.

REPREZENTAREA GRAFICĂ A DISTRIBUȚIEI FRECVENȚELOR

- **Histograme**:
  - X – axa valorilor;
  - Y – axa frecvențelor;
  - aria fiecărui dreptunghi – proporțională cu frecvența relativă respectivă.
- **Poligonul frecvențelor**: se unesc centrele laturilor superioare ale dreptunghiurilor din histogramă.
- **Frecvențe cumulate**: suma frecvențelor valorilor mai mici decât o valoare dată
  - variabile continue.

VALORI TIPICE ÎNTR-UN EŞANTION

- De la structură la număr
  - calitate → cantitate
  - simplificare, pentru a reprezenta succint o trăsătură tipică.
- Se descrie un eşantion printr-o valoare unică
  - atribut variabil numeric (cel puțin sortabil)

1. **Tendința centrală** (mediana, medii, mod)
2. **Împrăştierea** (amplitudine, quartile…, deviații, dispersie)

MEDIANA

- Descriere printr-**o observaţie** (sau prin media a două observaţii) din eşantion.
- Eşantionul **se sortează** după variabila studiată.
- **Definiţie**: *Mediana* unui set de N observaţii ordonate crescător este egală cu
  - valoarea de pe poziţia k+1, dacă N=2k+1
  - media dintre valorile de pe poziţiile k şi k+1, dacă N=2k.
- *Stabilitate*: schimbarea valorii unei observaţii, dar nu şi a rangului ei, nu afectează mediana.

MEDIA ARITMETICĂ (1)

1.- *Pentru atribute discrete:* $M = (x_1+\ldots+x_n) / n$

- Depinde de toate observațiile.
- Dacă valoarea x_i se repetă de p_i ori:

$$M = (p_1x_1+\ldots+p_nx_n) / (p_1+p_2+\ldots+p_n)$$

- Notând $f_i = p_i / n$: $M = f_1x_1 + \ldots + f_nx_n$

2.- *Pentru frecvențe distribuite pe intervale* - media ponderată a centrelor intervalelor de grupare:

- se alege mijlocul fiecărui interval (presupunând distribuție omogenă pe interval / principiul erorii minime)
- se înmulțeşte cu numărul de observații pe interval
- se sumează după toate intervalele şi se împarte la numărul de observații

MEDIA ARITMETICĂ (2)

- **Stabilitate**:
  - valorile aberante o afectează
  - mici modificări ale sumei nu o afectează
  - reaşezări de intervale nu o afectează prea mult
- **Liniaritate**: $M(a\mathbf{x}+b) = aM(\mathbf{x}) + b$
- **Abaterile** în raport cu media aritmetică:

$$\sum_i (x_i - \bar{x}) = 0$$

- **Definiția variațională**: media aritmetică este numărul M care minimizează expresia $\sum_i (x_i - \bar{x})^2$
  - legătura cu definirea dispersiei.

MEDIA ARMONICĂ

- Un automobil parcurge distanța Iaşi – Paşcani de mai multe ori, respectiv cu vitezele de 80 km/h, 90 km/h, 120 km/h, 60 km/h. Care a fost viteza sa medie?

$$M = 87,5 \text{ km/h}$$

- În realitate:

$$H = 4/(1/80 + 1/90 + 1/120 + 1/60) = 82,3 \text{km/h}.$$

- Utilizată la calcule bursiere ($H \leq G \leq M$) – distribuții în J.

MEDIA GEOMETRICĂ

- Populația SUA:
  - 1840: 17 069 000
  - 1850: 23 192 000
  - 1860: 31 443 000
- Dacă nu am avea observația din 1850:
- Media aritmetică M = 24 256 000
- Media geometrică G = 23 167 000

MÓDUL

- Valoarea dominantă (cea mai frecventă).
- Vârful poligonului frecvențelor. În cazul intervalelor:

  Mod = $L + i*(f_z - f_l) / ((f_z - f_l) + (f_z - f_h))$

  - i – lungimea intervalului
  - L – marginea inferioară a clasei modale
  - $f_{z,l,h}$ – frecvențele claselor modală, imediat inferioară ei şi imediat superioară

  – tipic spiritului uman să extindă calitatea cel mai des întâlnită la toate elementele observate.
- **Antimódul**: clasa (valoarea) de frecvență minimă.

COMPARAȚIE ÎNTRE MEDIANĂ, MEDIE ARITMETICĂ ŞI MOD

- La o distribuție simetrică, coincid.
- Media aritmetică nu se poate calcula pentru distribuții deschise (ultimul interval nemărginit);
- Mediana – da.
- Pentru distribuții asimetrice, módul dă impresia cea mai reală.
- Mediana şi módul nu au proprietăți de liniaritate.

AMPLITUDINE

- Măsură grosieră a variabilității.
- **Definiție**: diferența dintre cea mai mare și cea mai mică valoare ale observațiilor.
- Exemple:
  - amplitudinea salariilor;
  - amplitudinea temperaturii (pentru concediu);
  - amplitudinea notelor (relevanța unui test).

QUARTILE

- **Definiție**: Pentru un set de observații, *quartilele* (q_1, q_2, q_3), sunt valorile din şirul ordonat al tuturor observațiilor, pentru care numărul de valori mai mici reprezintă 25%, 50%, respectiv 75% din numărul total de observații.
  - q_2 este mediana;
  - q_1 este mediana valorilor din stânga medianei;
  - q_3 este mediana valorilor din dreapta medianei.

MĂSURI ALE ÎMPRĂŞTIERII

- **Definiție**: *Amplitudinea (intervalul) semi-inter-quartilă* este $0,5*(q_3 - q_1)$.
- Între q_1 şi q_3 se găsesc 50 % dintre valori.
- *Sumarul celor 5 valori*: $(\min,q_1,q_2,q_3,\text{Max})$
- **Definiție**: *Decilele* D_k, k=1..9, sunt valorile din şirul ordonat crescător la stânga cărora se află 10*k % dintre observații.
- **Definiție**: *Percentilele* P_k, k=1..100, sunt valorile din şirul ordonat crescător la stânga cărora se află k % dintre observații
- Importante sunt P_1 , P_5 , P_{95} , P_{99} .

DEVIAȚII ȘI DISPERSIE

- *Deviație medie*: media abaterilor absolute față de media aritmetică. Rar folosită.

$$dm = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}\right|; \qquad dm\_interval = \frac{\sum f_i\left|x_i - \bar{x}\right|}{\sum f_i}$$

- *Dispersia* a N observații: $V = \frac{\sum (x_i - \bar{x})^2}{N}$
- *Deviația standard a unui eșantion*: $SD = \sqrt{V}$
- Pe intervale: $SD = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$

COEFICIENTUL DE DISPERSIE

- Dispersia raportată la medie: $CV = \frac{SD}{\bar{x}}$
  - adimensional; comparabil pe atribute diferite.
- **Exemplu**. Eşantion de manageri;
  vârsta ($medie_1 = 51$, $SD_1 = 11{,}74$);
  IQ ($medie_2 = 125$, $SD_2 = 20$).
  Ce atribut are împrăştiere mai mare?
- $CV_1 = 11{,}74 : 51 = 0{,}23$
- $CV_2 = 20 : 125 = 0{,}16$.
- Concluzie: mai multă variaţie la vârstă.

MOMENTE

$$m_1 = \frac{\sum (x_i - \bar{x})}{N} = 0 \qquad m_2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{N} \qquad a_3 = \frac{m_3}{SD^3}$$

- =0: simetric;
- <0: asimetric negativ (mod dreapta);
- >0: asimetric pozitiv.

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{N} \qquad a_4 = \frac{m_4}{SD^4}$$

- <3: plat(ikurtic); >3: cu vârf ascuțit (leptokurtic)