RECAPITULARE

ȘL.Dr. Lavinia Moleriu

DISTRIBUȚII STATISTICE . STATISTICĂ DESCRIPTIVĂ 1.Introducere în biostatistică

- **Statistica:** *gruparea, analizarea și interpretarea datelor referitoare la fenomene de masă*
- **Biostatistica:** *ramura a statisticii, care studiază, interpretează și analizează date din domeniul biologic și medical*
- ***Variabilitatea:*** reprezintă principala caracteristică a proceselor biologice (variabilitatea implică un anumit grad de incertitudine)
- **Individ:** *un element din populație, localizat în spațiu și timp* (unitate statistică)
- **Populație:** *ansamblul indivizilor dintr-un studiu, având cel puțin o proprietate comună* (de asemenea este localizată in spațiu și timp)
- **Volumul populației:** *numărul indivizilor din populație*

2. Metode de studiu

Recensământ: este o metodă exactă, bine definită în timp și spațiu, dar e și foarte costisitoare

Screening: în acest tip de studiu nu este necesară localizarea în timp și este relativ costisitoare

Selecție (eșantionare): în aceste studii se alege o submulțime din populație, un eșantion, iar toată analiza se face la nivel de eșantion. În acest studiu costurile sunt reduse.

3.Biostatistică

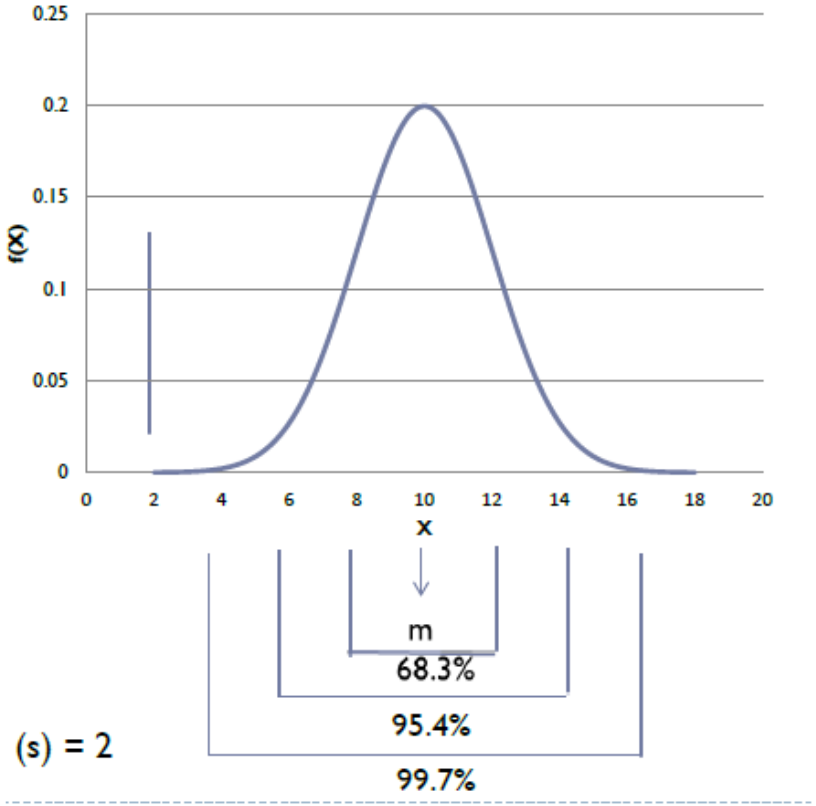
- **Statistică descriptivă:** cuprinde descrierea variabilelor
  - calcularea indicatorilor statisticii descriptive
  - reprezentări grafice
  - table de frecvențe
- **Statistică inferențială:** *generalizarea caracteristicilor unui eșantion pentru întreaga populație*
- **Eșantion reprezentativ:** *eșantion care cuprinde toate straturile populației în proporții similare și are toate caracteristicile populației*
- **Criterii de selecție a unui eșantion reprezentativ:**
  - *echiprobabilitate:* toți indivizii popuației să aibă aceeași probabilitate de a fi selectați în eșantion
  - *independență:* alegerea unui individ în eșantion să fie independentă de alegerea altui individ
- **Bias:** *orice condiție care influențează procesul de selecție, influențează reprezentativitatea eșantionului*

Biostatistică

- **Variabile statistice:**
  - *variabile numerice:* exprimată printr-un număr (are unitate de măsură)
  - *variabile ordinale (rang):* exprimate printr-un număr asociat unei scări convenționale
  - *variabile nominale (calitative):* se definesc mai multe clase, corespunzătoare valorilor posibile ale calităților. Se folosesc numere (proporții) corespunzătoare fiecărei clase
  - *Variabileilele dichotomice:* variabilele care au doar doua valori posibile

4. Statistică descriptivă

- **Indicatori ai tendinței centrale**
  - *Media:* media aritmetică (cel mai folosit indicator al tendinței centrale)
  - *Mediana:* valoarea care împarte lotul de valori ordonate în două subclase egale ca număr de valori
  - *Moda:* cea mai frecventă valoare
- **Indicatori de dispersie**
  - *domeniul de valori:* amplitudinea intervalului (range-ul)
  - *eroarea medie absolută (abatere centrală) :* reprezintă distanța unei valori individuale fața de valoarea medie
  - *deviația standard:* reprezintă gradul de împrăștiere al valorilor individuale în jurul mediei eșantionului
  - *eroarea standard a mediei:* arată gradul de împrăștiere a mediilor eșantioanelor în jurul mediei populației. Mediile eșantioanelor au o distribuție normală în juru mediei populației. (eroarea standard a mediei scade când dimensiunea eșantionului crește)



5. Distribuția normală (Gauss)

- **Proprietăți:**
  - *este simetrică față de valoarea medie*
  - *lățimea este dependentă de deviația standard*
  - *aria totală de sub curbă este egală cu 1*
  - *”cozile”graficului se întind spre infinit dar nu ating niciodată planul orizontal*

6. Intervale de încredere. Estimarea valorilor / estimarea mediilor de valori

- **valorile individuale** se vor găsi cu o anumită probabilitate în intervalele (unde m = media, iar s=deviația standard):
- **(m-s ; m+s)** în 68 % din cazuri (mai exact 68,3%)
- **(m-2*s ; m+2*s)** în 95% din cazuri (mai exact 95,4%)
- **(m-3*s ; m+3*s)** în 99,7% din cazuri
  - media valorilor se va găsi cu o anumită probabilitate în intervalele (unde m = media, iar sx=eroarea standard, sx=s/√n, n=volumul populației):
- **(m-sx ; m+sx)** în 68 % din cazuri (mai exact 68,3%)
- **(m-2*sx ; m+2*sx)** în 95% din cazuri (mai exact 95,4%)
- **(m-3*sx ; m+3*sx)** în 99,7% din cazuri

TESTE STATISTICE
1. Definiții

- **Testele statistice:** reprezintă un procedeu prin care se stabilește, cu un anumit nivel de încredere, dacă diferențele între parametrii statistici sunt sau nu sunt semnificative.
- **Diferențe nesemnificative:** sunt acele diferențe, care au o probabilitate mare să apară din întâmplare și se datorează variabilității de eșantionare
- **Diferențe semnificative:** sunt acele diferențe, care au o probabilitate mică să apară din întâmplare
- **Prag de semnificație α:** reprezintă valoarea convențională sub care începem să considerăm diferențele ca semnificative. Uzual, α=5%

TESTE STATISTICE
1. Definiții

- **Ipoteze de cercetare:** certifică existența unei diferențe între grupurile studiate, sau o asociere între factori
- **Testarea statistică a ipotezei:** permite cuantificarea riscului de eroare implicat în mecanismul inferenței statistice
- **Ipoteza de zero (H0):** propoziție ce afirmă că diferențele observate sunt nesemnificative (de regulă, se construiește pe o negație),
  - matematic : H0: $\mu_1 = \mu_2$
- **Ipoteza alternativă (Ha):** se construiește prin negarea iotezei de nul,
  - matematic se scrie: $\mu_1 \neq \mu_2$- ipoteza alternativă bilaterală,

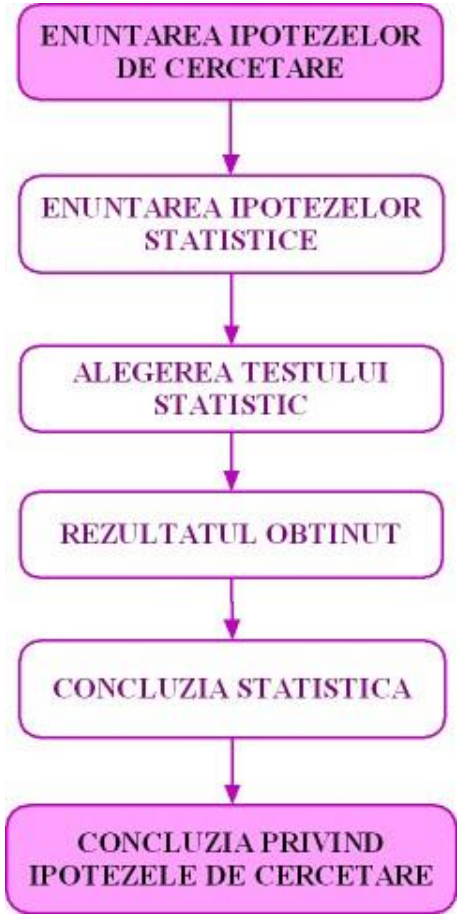
    respectiv $\mu_1 > \mu_2$ sau $\mu_B < \mu_F$ - ipoteze alternative unilaterale

TESTE STATISTICE
Definiții

- **Valoarea “p” calculată prin teste:** reprezintă probabilitatea ca diferențele observate să fi apărut din întâmplare
- **Decizia testelor statistice:** formularea deciziei se face în funcție de valoarea **”p”**:
  - dacă $p \geq \alpha$ acceptăm H0 → diferențele sunt nesemnificative
  - dacă $p < \alpha$ respingem H0 → diferențele sunt semnificative
  - pentru $\alpha = 0{,}05$ regiunea de respingere se împarte în 3 subregiuni:
    - dacă $p < 0{,}05$, atunci diferențele sunt semnificative
    - dacă $p < 0{,}01$, atunci diferențele sunt foarte semnificative
    - dacă $p < 0{,}001$, atunci diferențele sunt extrem de semnificative

TESTE STATISTICE
Definiții

- **Eroarea statistică:** știind că decizia unui test statistic are caracter probabilistic există riscul de a avea erori în decizia noastră. Erorile statistice posibile se împart în două clase:
  - erori de tip I: când respingem H0, deși este adevărată
  - erori de tip II: când acceptăm H0, deși este falsă
- Cu α se notează probabilitatea erorii de tip I, iar cu β se notează eroarea de tip II.
- **Nivelul de încredere**: reprezintă capacitatea de a accepta o ipoteză când aceasta este adevărată
- Mărimea $1-\alpha$ se numește *nivel de încredere* sau *nivel de confidență* a testului, unde α reprezintă pragul de semnificație, sau probabilitatea erorii de tip I
- **Puterea testului:** reprezintă capacitatea de a respinge o ipoteză când aceasta este falsă
- Mărimea $1-\beta$ se numește *puterea testului*, unde β reprezintă probabilitatea erorii de tip II
- **OBS:** Cele două caracteristici variază invers proporțional

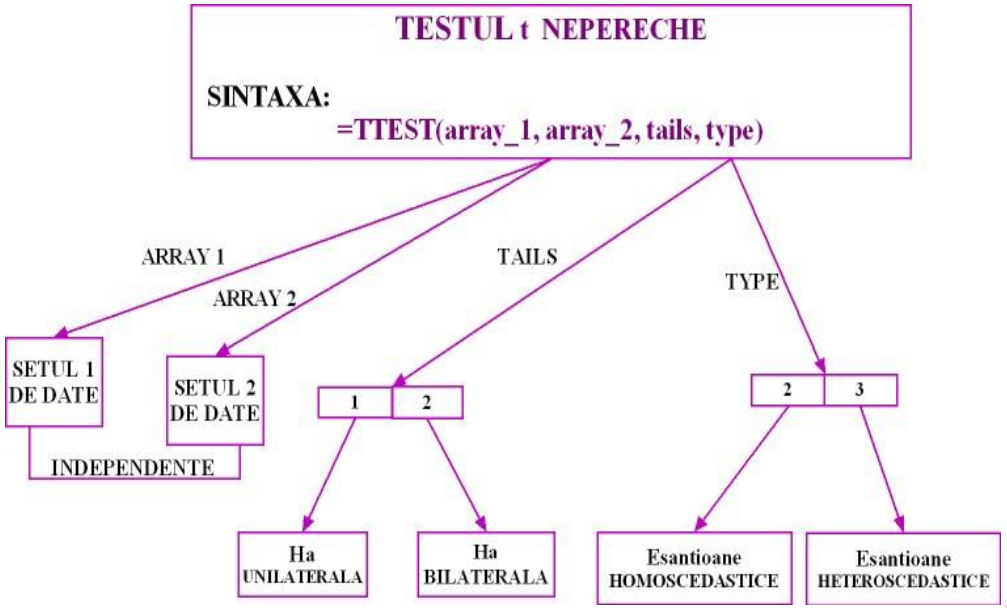


TESTE STATISTICE
Etapele unui test statistic

TESTE STATISTICE

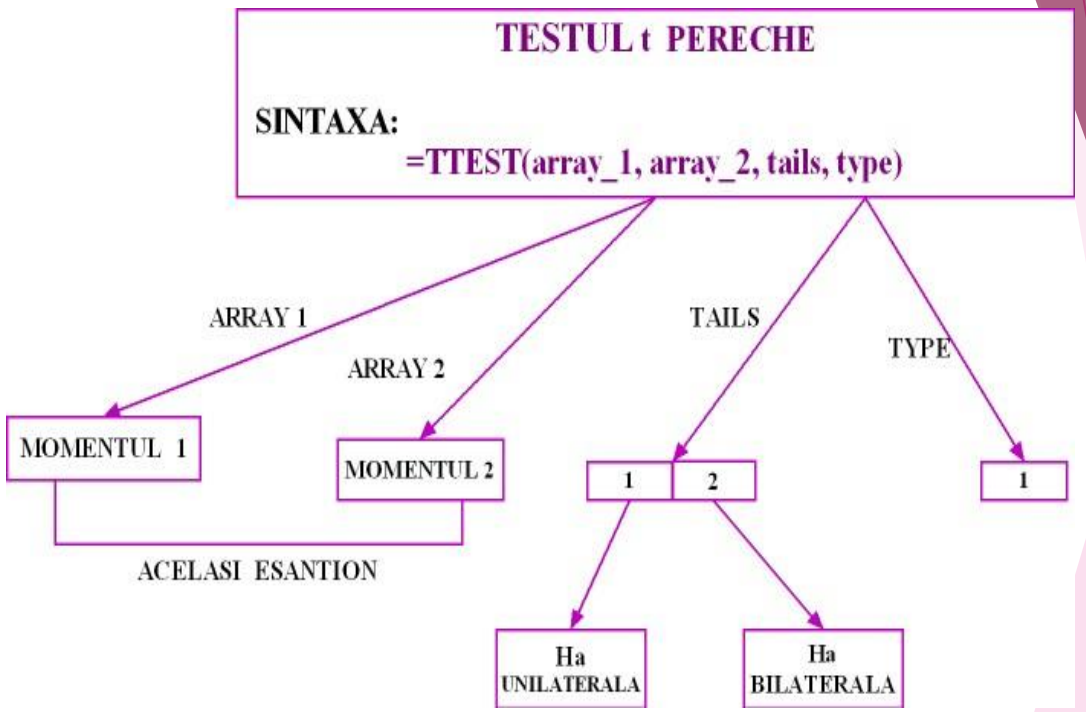
- **Teste statistice uzuale:**
  - Testul z
  - Testul t
  - Testul t nepereche
  - Testul t pereche
  - Testul χ^2
  - ANOVA

TESTE STATISTICE Teste pentru variabile numerice:



- TESTUL t NEPERECHE
  - seriile de valori provin din determinări pe indivizi diferiți, adică seriile sunt independente
  - se compară două valori medii obținute pe serii independente și variabile distribuite normal
  - H0: Între cele două valori medii nu sunt diferențe
    - formula: $\mu_1 = \mu_2$
  - Ha: Între cele două valori medii există diferențe
    - formula: $\mu_1 \neq \mu_2$ ipoteză alternativă bilaterală
    - $\mu_1 > \mu_2$ sau $\mu_1 < \mu_2$ ipoteză alternativă unilaterală

TESTE STATISTICE Teste pentru variabile numerice:



- TESTUL t PERECHE
  - seriile de valori provin din determinări pe aceiași indivizi în condiții diferite, adică seriile sunt dependente
  - se compară două valori medii obținute pe serii perechi și variabile distribuite normal
  - H0: Între cele două valori medii nu sunt diferențe
    - formula: $\mu_1 = \mu_2$
  - Ha: Între cele două valori medii există diferențe
    - formula: $\mu_1 \neq \mu_2$ ipoteză alternativă bilaterală
    - $\mu_1 > \mu_2$ sau $\mu_1 < \mu_2$ ipoteză alternativă unilaterală

TESTE STATISTICE Teste pentru variabile numerice:

- **ANOVA (ANalysis Of VAriance)**
  - este un test de comparare a mediilor mai multor populații cu distribuții normale (se compară ”n” serii de date independente)
  - H0: **Nu** există diferențe între mediile celor ”n” serii
    - formula: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n$
  - Ha: **Există** cel puțin o serie (valoare medie) care diferă față de celelelte
- ”p” valoarea testului se obține cel mai ușor folosind soft-ul Excel: din meniul Data -> Data Analysis -> Anova Single Factor

TESTE STATISTICE

- **TESTUL χ^2**
  - realizează analiza asocierilor ce implică date de tip categorie (frecvența de apariție)
  - se compară diferențele între proporții
  - H0: Între proporții **nu** există diferențe
  - Ha: Între proporții **există** diferențe
  - ”p” valoarea testului se obține cel mai ușor folosind Epi Info, însă se mai pot folosi si alte soft-uri, spre exemplu Excel sau Graphpad

TESTE STATISTICE

- **Concluzia statistică:**
  - Dacă $p \geq 0.05 \Rightarrow$ acceptăm H0 $\Longrightarrow$ nu există diferențe semnificative între...
  - Dacă $p < 0.05 \Rightarrow$ respingem H0 $\Longrightarrow$ există diferențe semnificative între...
  - Dacă $p < 0.01 \Rightarrow$ respingem H0 $\Longrightarrow$ există diferențe FOARTE semnificative între...
  - Dacă $p < 0.001 \Rightarrow$ respingem H0 $\Longrightarrow$ există diferențe EXTREM de semnificative între...

EPIDEMIOLOGIE. ANALIZA RISCULUI

- **FACTOR DE RISC** = o cauză ipotetică (indiferent de natură – comportament, condiţie) ce determină creşterea probabilităţii ca un individ sănătos să dezvolte o anumită boală
- **Clasificare:**
  - Factori de mediu
    - factori poluanţi
    - toxine
    - microorganisme infecţioase, etc
  - Factori comportamentali (obiceiuri)
    - fumat
    - alcool
    - droguri
    - nerespectarea măsurilor de protecţie a muncii, etc.
  - Factori sociali
    - evenimente familiale tragice
    - divorţ, etc.
  - Factori genetici
    - hipercolesteromia, etc

EPIDEMIOLOGIE. ANALIZA RISCULUI

- **Tipuri de expunere la acţiunea factorului de risc**
  - Expunere punctuală
    - accidente
  - Expunere cronică
    - cea mai frecventa -> se estimează “doza” curentă, cumulată şi durata expunerii
- **Relaţie factor risc / boală**
  - Factor cauzal
    - se atribuie factorului o acţiune curentă
  - Factor favorizant (marker)
    - creşte probabilitatea de adezvolta o boală

EPIDEMIOLOGIE. ANALIZA RISCULUI

	B+	B-	Total
E+	a	b	a+b
E-	c	d	c+d
Total	a+c	b+d	a+b+c+d

- **Prezentarea datelor** -> tabele de contingență

- unde
  - **a** reprezintă pacienții expusi la un factor de risc care dezvoltă boala
  - **b** reprezintă pacienții expusi la un factor de risc care nu dezvoltă boala
  - **c** reprezintă pacienții ne-expusi la un factor de risc care dezvoltă boala
  - **d** reprezintă pacienții ne-expusi la un factor de risc care nu dezvoltă boala
  - **a+b** reprezintă pacienții expusi la un factor de risc
  - **c+d** reprezintă pacienții ne-expusi la un factor de risc
  - **a+c** reprezintă pacienții care dezvoltă boala
  - **b+d** reprezintă pacienții care nu dezvoltă boala
  - **a+b+c+d** reprezintă totalul pacienților din studiu (volumul eșantionului)

EPIDEMIOLOGIE. ANALIZA RISCULUI Metode de studiu

- Experimentale: **cele mai exacte studii** -> investigatorul are control complet asupra factorului de risc. Nu prea este recomandat din considerente etice şi deontologice .
- Observaţionale: studiile se efectuează pe loturi unde expunerea nu s-a întâmplat la cererea investigatorului.
  - Cohortă prospectivă (follow up) (E+ / E-)
  - Cohorta retrospectivă (E+ / E-)
  - Case control (B+ / B-)
  - Transversal (cross – sectional)

EPIDEMIOLOGIE. ANALIZA RISCULUI
Indicatorii din analiza riscului:

- ***Riscul Relativ (RR)*** = probabilitatea de apariţie a afecţiunii la cei expuşi faţă de probabilitatea de apariţie a afecţiunii la cei neexpuşi

RISCUL RELATIV

$$RR = \frac{\textbf{incidenta afectiunii in grupul expus}}{\textbf{incidenta afectiunii in grupul neexpus}} = \frac{P(D+|E+)}{P(D+|E-)} = \frac{a/(a+b)}{c/(c+d)}$$

- RR > 1 ∃ risc
- RR ≈ 1 afectiunea si exp. la presupusul factor de risc nu sunt corelate
- RR < 1 asociere negativa (efect protectiv)

EPIDEMIOLOGIE. ANALIZA RISCULUI
Indicatorii din analiza riscului:

- ***Raportul Odds (OR)*** = arată de câte ori este mai mare şansa de îmbolnăvire la lotul expus față de lotul neexpus, raportat la cei care nu se îmbolnăvesc

RAPORTUL INDICILOR "ODDS"

$$O = \frac{P(\text{evenimentul sa se produca})}{1 - P(\text{evenimentul sa se produca})} = \frac{P(\text{evenimentul sa se produca})}{P(\text{evenimentul sa NU se produca})}$$

Indicii "*Odds*" (de paritate, "succes/esec"):

$$O_{D+|E+} = \frac{P(D+|E+)}{P(D-|E+)} = \frac{a}{b}$$

$$O_{D+|E-} = \frac{P(D+|E-)}{P(D-|E-)} = \frac{c}{d}$$

"odds ratio" $\quad OR = \frac{O_{D+|E+}}{O_{D+|E-}} = \frac{a*d}{b*c}$

- $OR > 1$
- $OR \approx 1$
- $OR < 1$

EPIDEMIOLOGIE. ANALIZA RISCULUI

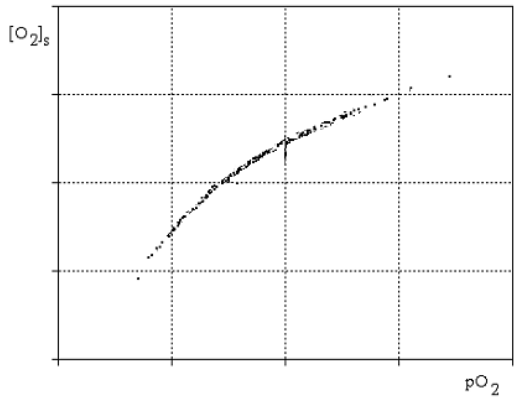
- **Testarea ipotezelor privind riscul:**
- **H0:** nu există factor de risc (semnificativ)
  - $\boldsymbol{R = 1}$
- **Ha:** există factor de risc (semnificativ)
  - $\boldsymbol{R \neq 1}$
- **OBS:** Uzual OR > RR
- **Concluzia statistică:**
  - Dacă $p \geq 0.05 \Rightarrow$ acceptăm H0 $\Longrightarrow$ nu există un factor de risc semnificativ între...
  - Dacă $p < 0.05 \Rightarrow$ respingem H0 $\Longrightarrow$ există un factor de risc semnificativ între...
  - Dacă $p < 0.01 \Rightarrow$ respingem H0 $\Longrightarrow$ există un factor de risc FOARTE semnificativ între...
- Dacă $p < 0.001 \Rightarrow$ respingem H0 $\Longrightarrow$ există un factor de risc EXTREM de semnificativ între...

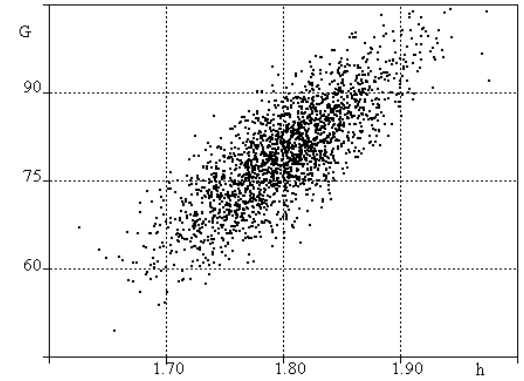
ANALIZA DE CORELAȚIE ȘI REGRESIE

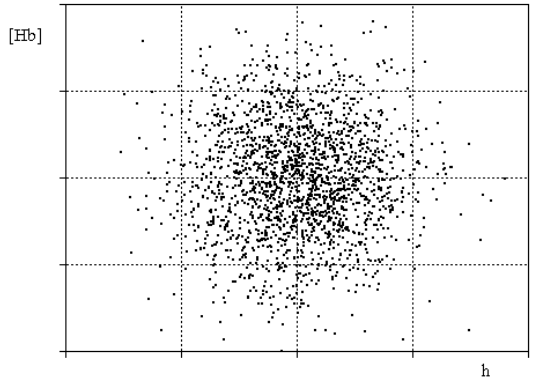
- **Definiție:** *studiază dependența dintre două sau mai multe variabile.*
- **Regresia:** *ne arată cum o variabilă este dependentă de altă variabilă* (una dintre variabile este fixată și cealaltă aleatorie)

- **Obs:** presupune existența unei relații cu sens (o variabilă o influențează pe cealaltă)
  - **Corelația:** *ne arată gradul în care o variabilă este dependentă de o altă variabilă* (măsoară tăria asociației dintre variabile prin calculul coeficientului de corelație, ambele variabile sunt aleatorii)
  - **Diagrame de corelație:** reprezentări grafice având pe axe cele două variabile
    - *în Excel reprezentarea grafică se face din meniul Insert -> opțiunea Scatter*
  - **Analiza bivarată: analiza care urmarește comportamentul a 2 variabile**
    - Variabila X – predictor
    - Variabile Y – răspuns

ANALIZA DE CORELAȚIE ȘI REGRESIE







- **Relatii de dependent**
  - Variabile independentre: repartitia punctelor este aproape simetrica si uniforma (un individ = un punct)
  - Variabile dependente
  - Variabile corelate

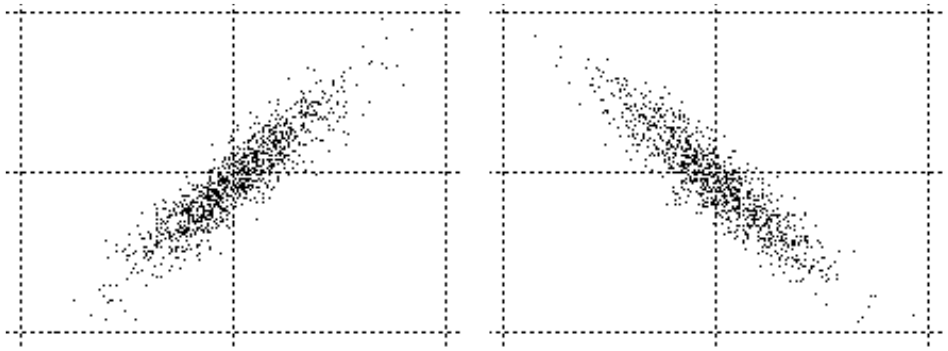
ANALIZA DE CORELAȚIE ȘI REGRESIE

- **Semnificația coeficientului de corelație:** se testează cu testul **t** și se interpretează ca orice test (dacă $p > 0.05$, corelația este nesemnificativă ; dacă $p < 0.05$, corelația este semnificativă)
- **Semnificația statistică:** se testează dacă apariția corelației este întâmplătoare sau e reproductibilă în populație.
  - *Notații:* r – pentru eșantion și ρ – pentru populație
  - *Ipotezele statistice*:
- H0: $\rho = 0$, (cele două variabile sunt independente)
- Ha: $\rho \neq 0$, (cele două variabile sunt corelate)
  - *Testul aplicat:*
- Testul **t** (Student), are o repartiție Student cu N-2 grade de libertate

$$t = t_{calc} = r \cdot \sqrt{\frac{N-2}{1-r^2}}$$

ANALIZA DE CORELAȚIE ȘI REGRESIE

- **Corelația liniară, coeficientul de corelație Pearson:** caracterizează "intensitatea" asociației dintre 2 variabile
  - *Proprietăți:* $r \in [-1, 1]$
- $r > 0 \Rightarrow$ corelație directă (pozitivă)
- $r < 0 \Rightarrow$ corelație indirectă (inversa, negativă)
- $r = 1 \Rightarrow$ corelație directă perfectă
- $r = -1 \Rightarrow$ corelație inversă perfectă
- $r = 0 \Rightarrow$ variabile necorelate



a. Corelație directă

b. Corelație inversă

ANALIZA DE CORELAȚIE ȘI REGRESIE

- În 1974 Colton sugerează următoarele reguli empirice privind interpretarea coeficientului de corelație

- $r \in [-0.25\,, 0.25] \Rightarrow$corelație slabă sau nulă,atunci când **r** se apropie de 0
- $0.25 < |r| < 0.5 \Rightarrow$ corelație slabă spre medie
- $0.5 < |r| < 0.75 \Rightarrow$ corelație medie spre bună
- $|r| > 0.75 \Rightarrow$ corelație foarte bună spre puternică, atunci când **r** se apropie de ± 1
  - **Coeficientul de determinare R^2:** *reprezintă proporția din variația uneia dintre variabile ce poate fi atribuită (sau explicată) de variația celeilalte variabile (oferă informații despre tăria corelației, nu despre sensul corelației)*
    - *Proprietăți:* $R^2 \in [0\,, 1]$

ANALIZA DE CORELAȚIE ȘI REGRESIE

- **Dreapta de regresia:** *”cea mai bună” dreaptă, care trece prin punctele diagramei de corelație*
  - *Parametrii dreptei:* $y = ax + b$, unde a este ordonata la origine (intercept), iar b este panta
  - *Tehnica de fitare* – metoda celor mai mici pătrate (Gauss). Pentru determinarea coeficienților a și b din ecuația dreptei vom considera dreapta care trece printre punctele experimentale ca având suma pătratelor abaterilor minimă

$$SSE = \sum \epsilon_i^2 = min$$

- **Aplicarea testului:**
  - *În excel:* Din meniul Data -> Data Analysis -> Regression
  - *În Epi Info:* Din meniul Analyze Data-> Linear Regression

ANALIZA DE CORELAȚIE ȘI REGRESIE Concluzia statistică

- Dacă $p \geq 0.05 \Rightarrow$ acceptăm H0 $\Longrightarrow$ cele două variabile sunt semnificativ independente
- Dacă $p < 0.05 \Rightarrow$ respingem H0 $\Longrightarrow$ cele două variabile sunt semnificativ corelate
- Dacă $p < 0.01 \Rightarrow$ respingem H0 $\Longrightarrow$ cele două variabile sunt FOARTE corelate
- Dacă $p < 0.001 \Rightarrow$ respingem H0 $\Longrightarrow$ cele două variabile sunt EXTREM de corelate