ANALIZA DE CORELAȚIE ȘI REGRESIE

ȘL.DR. MOLERIU LAVINIA

CUPRINS

Analiza corelației și regresiei:

problematică, reprezentarea datelor, statistici.

Regresia liniara:

coeficient de regresie și interpretare, estimare, testare statistică.

CORELATIA SI REGRESIA

Analiza asociațiilor ce implica date de tip continuu

ANALIZA CORELAȚIEI ȘI REGRESIEI

(analize bivariate → analize care urmăresc comportamentul a 2 variabile)

- Variabila X – predictor
- Variabile Y – răspuns

TIPURI DE STUDII

Epidemiologice	Experimentale	Corelatie / Regresie
• X – un presupus factor de risc • Y – apariția afecțiunii	• X – variabila independentă (fixata de investigator) • Y – variabila dependenta / răspuns	• Corelatie = ambele variabile sunt aleatorii • Regresie = una dintre variabile este fixată și cealaltă este aleatorie

CORELAȚIA ȘI REGRESIA

- **DEFINIȚII:** *STUDIAZĂ DEPENDENŢA DINTRE DOUĂ SAU MAI MULTE VARIABILE.*
  - **CORELAŢIA***: 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)*
  - **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)

RELAȚII ÎNTRE VARIABILE

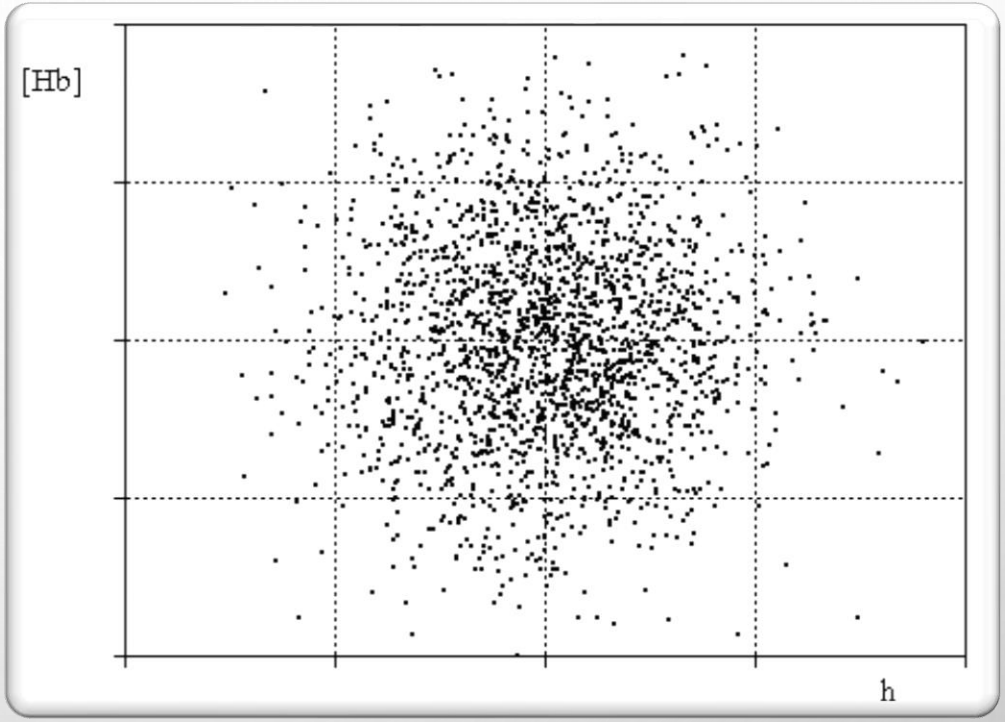
- **RELAȚII DE DEPENDENȚĂ**
  - VARIABILE INDEPENDENTRE: REPARTITIA PUNCTELOR ESTE APROAPE SIMETRICA SI UNIFORMA (UN INDIVID = UN PUNCT)
  - VARIABILE DEPENDENTE
  - VARIABILE CORELATE

REPREZENTĂRI GRAFICE

- **REPREZENTAREA GRAFICĂ**
- **DIAGRAME DE CORELAȚIE:** REPREZENTĂRI GRAFICE AVÂND PE AXE CELE DOUĂ VARIABILE (*ÎN EXCEL REPREZENTAREA GRAFICĂ SE FACE DIN MENIUL INSERT -> OPȚIUNEA SCATTER*)

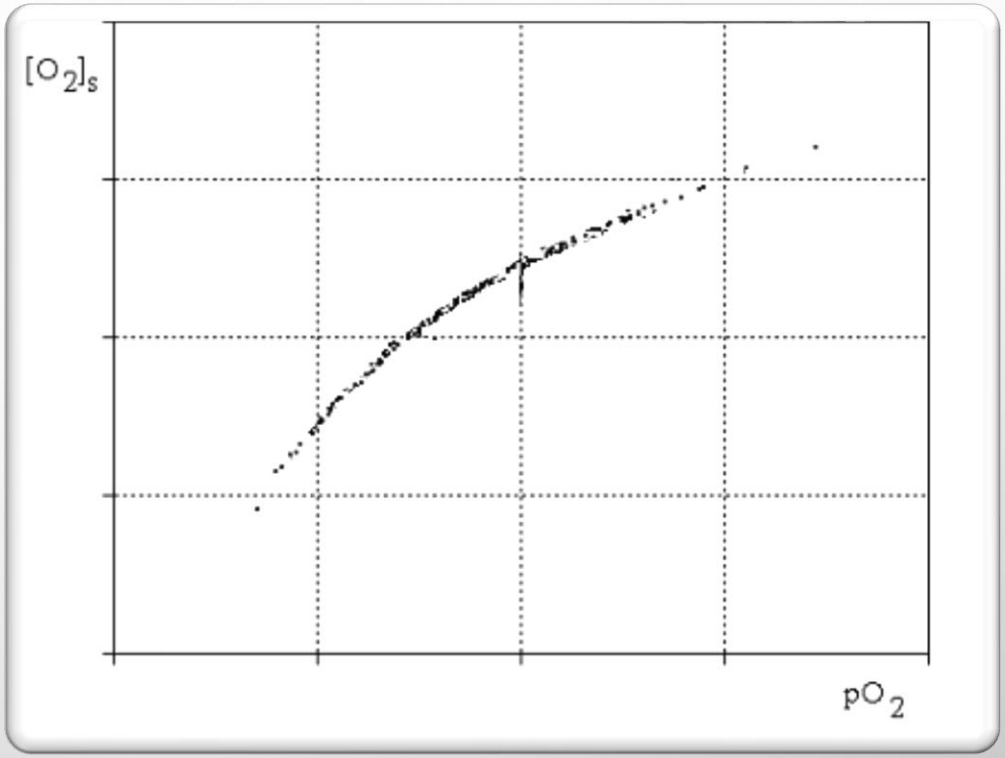
VARIABILE INDEPENDENTE

- EX.:RELAȚIA ÎNTRE ÎNALȚIMEA UNUI INDIVID (H = ÎNALȚIMEA) ȘI CONCENTRATIA HEMOGLOBINEI ÎN SÂNGE (HB)



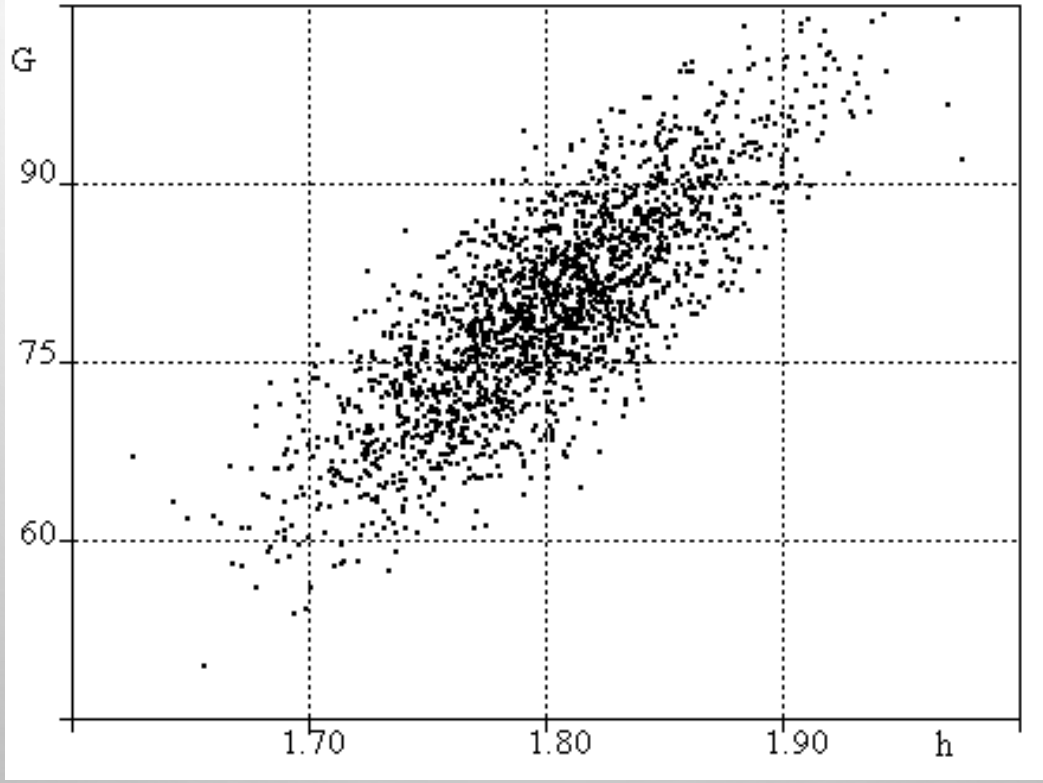
VARIABLE DEPENDENTE

- EX: DACĂ REPREZENTĂM RELAȚIA DINTRE PRESIUNEA PARȚIALA A OXIGENULUI DIN AERUL RESPIRAT ȘI CONCENTRAȚIA OXIGENULUI DIZOLVAT ÎN SÂNGE ($[O_2]$ ÎN SÂNGE - PO_2 ATMOSFERIC) OBȚINEM URMATORUL GRAFIC. RELATIE CAUZALA - MODEL MATEMATIC



VARIABLE CORELATE

- Corelația dintre înaltimea și masa corporală a unor indivizi (G = greutate, h = inaltime)



CORELATIA LINEARA

- COEFICIENT DE CORELATIE (PEARSON)
  - DEFINIȚIE: CARACTERIZEAZĂ“INTENSITATEA” ASOCIAȚIEI DINTRE 2 VARIABILE
  - FORMULA DE CALCUL:

$$r = r_{xy} = \frac{s_{xy}}{S_x S_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

unde

S_x^2 si S_y^2 reprezintă varianța lui x, respectiv y:

$$S_x^2 = \frac{\sum (X_i - \bar{X})^2}{N} \quad , \; S_y^2 = \frac{\sum (Y_i - \bar{Y})^2}{N}$$

iar S_{xy} se numeşte convarianța între x şi y şi este dat de:

$$S_{xy} = \frac{\sum (X_i - \bar{X})\,(Y_i - \bar{Y})}{N}$$

CORELAȚIA LINIARĂ

- PROPRIETĂȚI:
  - $r \in [-1\,,1]$
  - $r > 0$ ⇒ CORELAȚIE DIRECTĂ (POZITIVĂ)
  - $r < 0$ ⇒ CORELAȚIE INDIRECTĂ (INVERSA, NEGATIVĂ)
  - $r = 1 \Rightarrow$ CORELAȚIE DIRECTĂ PERFECTĂ
  - $r = -1 \Rightarrow$ CORELAȚIE INVERSĂ PERFECTĂ
  - $r = 0 \Rightarrow$ VARIABILE NECORELATE

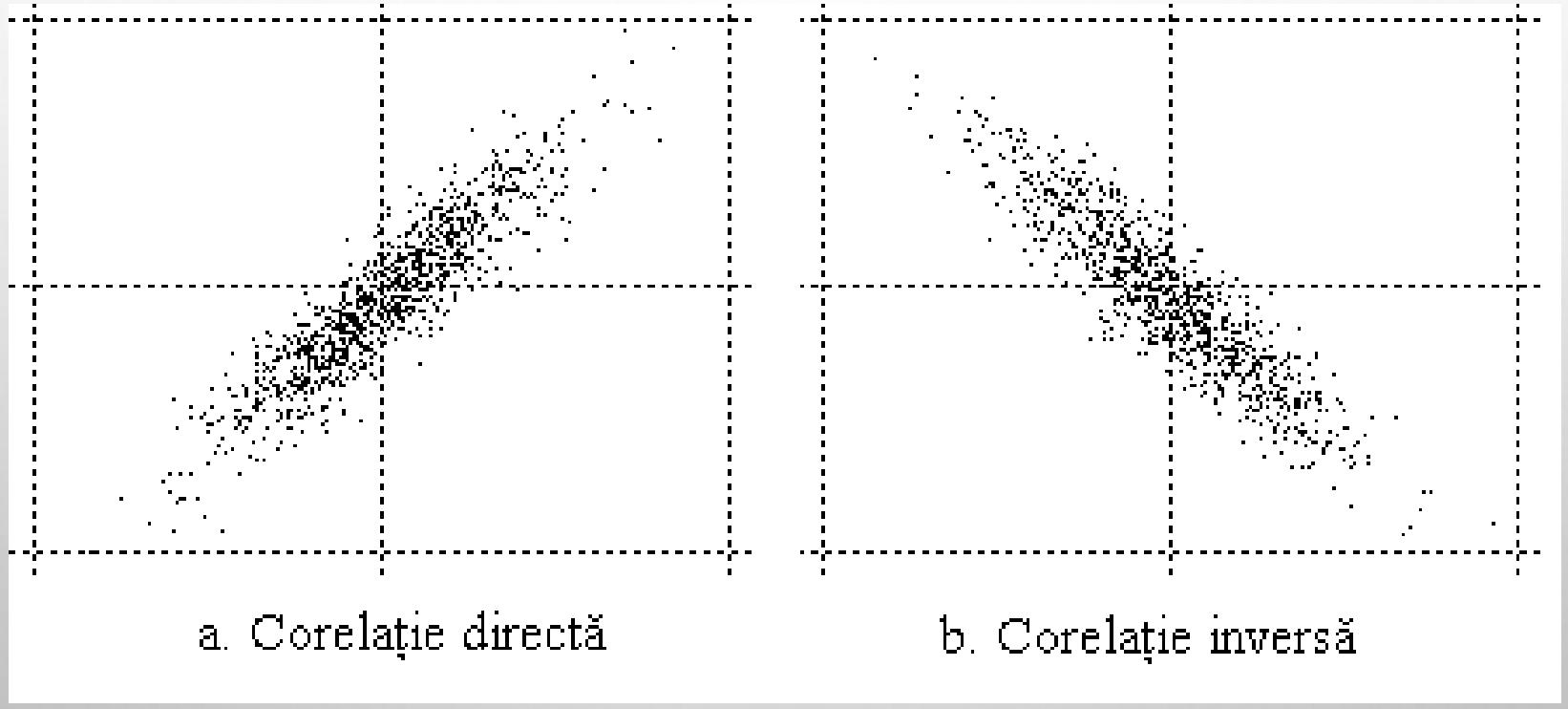
COEFICIENTUL

LUI PEARSON

ÎN 1974 COLTON SUGEREAZĂ URMĂTOARELE REGULI EMPIRICE PRIVIND INTERPRETAREA COEFICIENTULUI DE CORELAȚIE

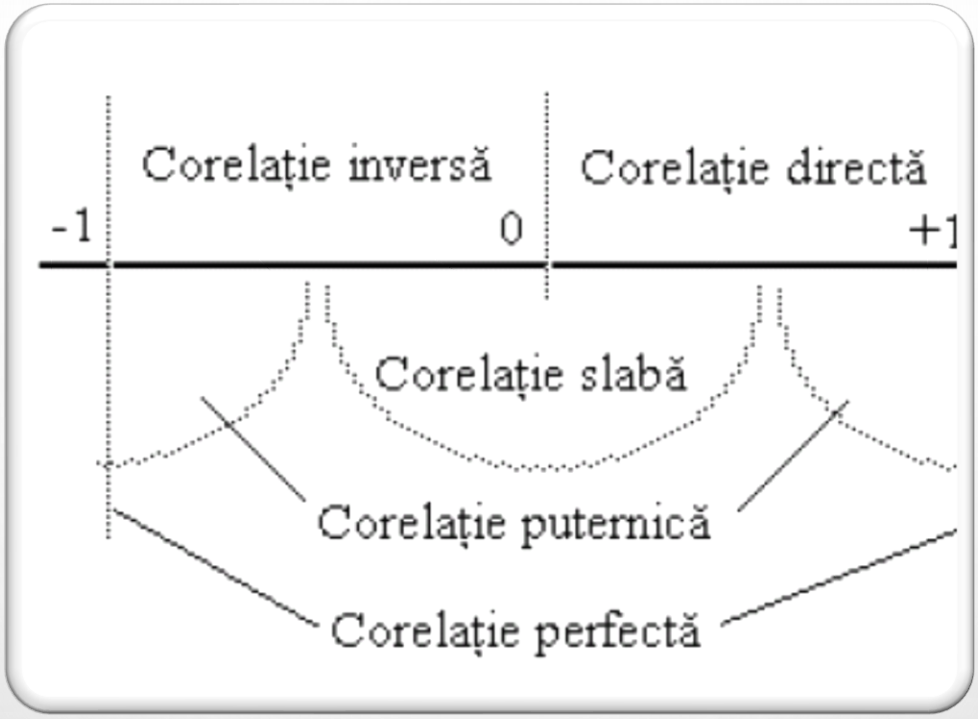
- $r \in [-0.25\ , 0.25]$ ⇒CORELAȚIE SLABĂ SAU NULĂ, ATUNCI CÂND **R** SE APROPIE DE 0
- $0.25 < |r| < 0.5$ ⇒ CORELAȚIE SLABĂ SPRE MEDIE
- $0.5 < |r| < 0.75$ ⇒ CORELAȚIE MEDIE SPRE BUNĂ
- $|r| > 0.75$ ⇒ CORELAȚIE FOARTE BUNĂ SPRE PUTERNICĂ, ATUNCI CÂND **R** SE APROPIE DE ± 1

CORELAȚII DIRECTE ȘI INVERSE



a. Corelație directă

b. Corelație inversă



APRECIEREA “INTENSITĂȚII” CORELAȚIEI LINIARE DUPA VALORILE LUI “R”

SEMNIFICAȚIA COEFICIENTULUI DE CORELAȚIE

Valorile lui r depind de

- gradul de împrăștiere a datelor
- numărul de puncte (N)
- OBS: Dacă N -> este mult prea mic putem obține concluzii greșite

Ipotezele statistice

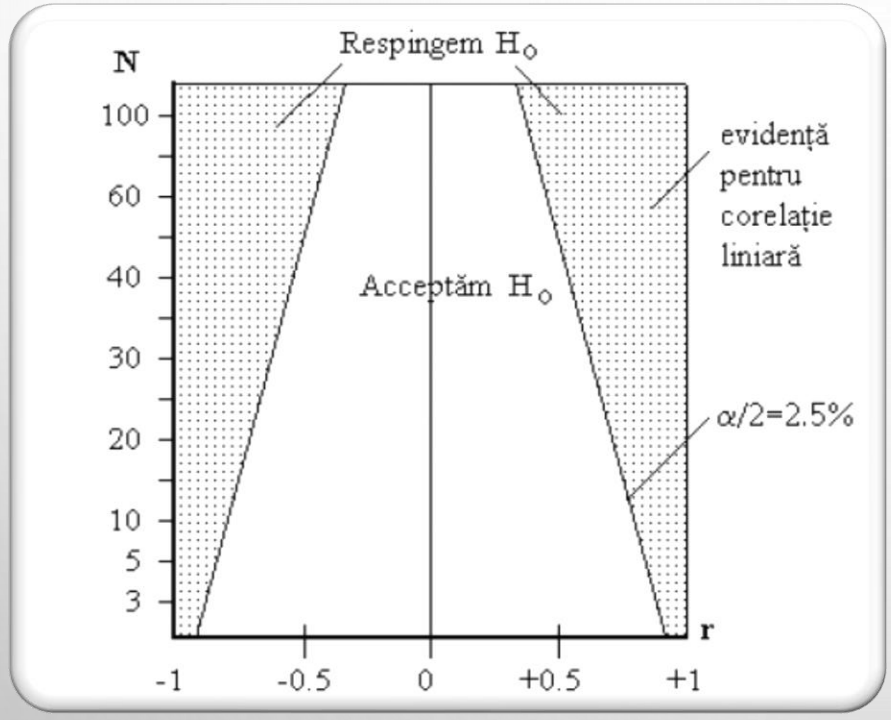
- H0: $\rho=0$ (cele 2 variabile sunt independente)
- Ha: $\rho\neq0$ (cele 2 variabile sunt corelate)

SEMNIFICAȚIA COEFICIENTULUI DE CORELATIE

- IPOTEZELE STATISTICE: H0 ȘI HA
- TEST APLICAT: SE TESTEAZĂ DACĂ APARIȚIA CORELAȚIEI ESTE ÎNTÂMPLĂTOARE SAU E REPRODUCTIBILĂ ÎN POPULAȚIE.
  - TESTUL **T** (STUDENT) -> ÎN EXCEL CU FUNCȚIA TDIST
- FUNDAMENTARE TEORETICA

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

  - ARE O REPARTITIE STUDENT CU Y=N-2 GRADE DE LIBERTATE



SEMNIFICATIA COEFICIENTULUI DE CORELATIE

COEFICIENTUL DE DETERMINARE R

$R^2 = r * r = r^2$

Definiție:

- măsoara proporția din variația uneia dintre variabile ce poate fi atribuită sau explicată de variația celeilalte variabile

CORELAȚII PENTRU VARIABLE ORDINALE

CORELATIA RANGURILOR

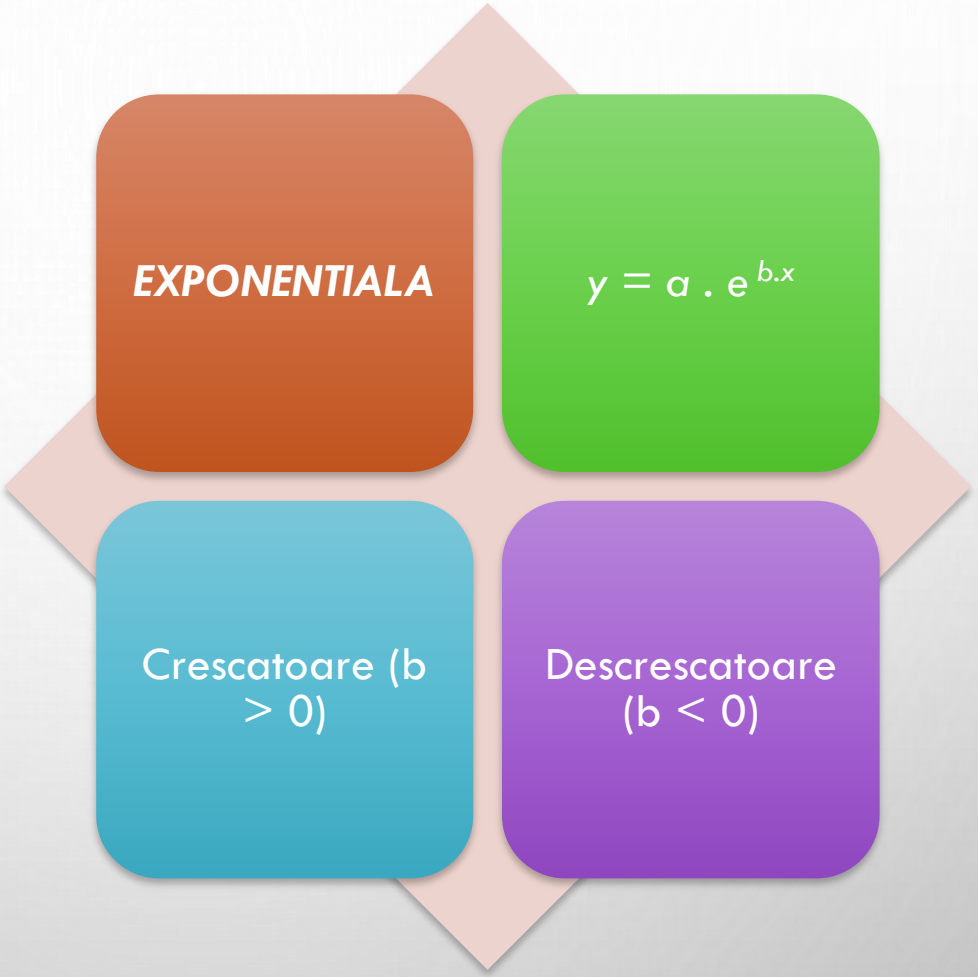
SPEARMAN "R"

- *Compararea a doua clasificari*

COEFICIENTUL DE CORELATIE KENDALL

Apl. pentru variable ordinale

CORELATII NELINEARE



CORELATII NELINEARE

LOGARITMICA:

$y = a + b \cdot \log x$

PUTERE:

$y = a \cdot x^{b}$

HIPERBOLICA:

$(x - a) \cdot (y - b) = k$

e) LOGISTICA:

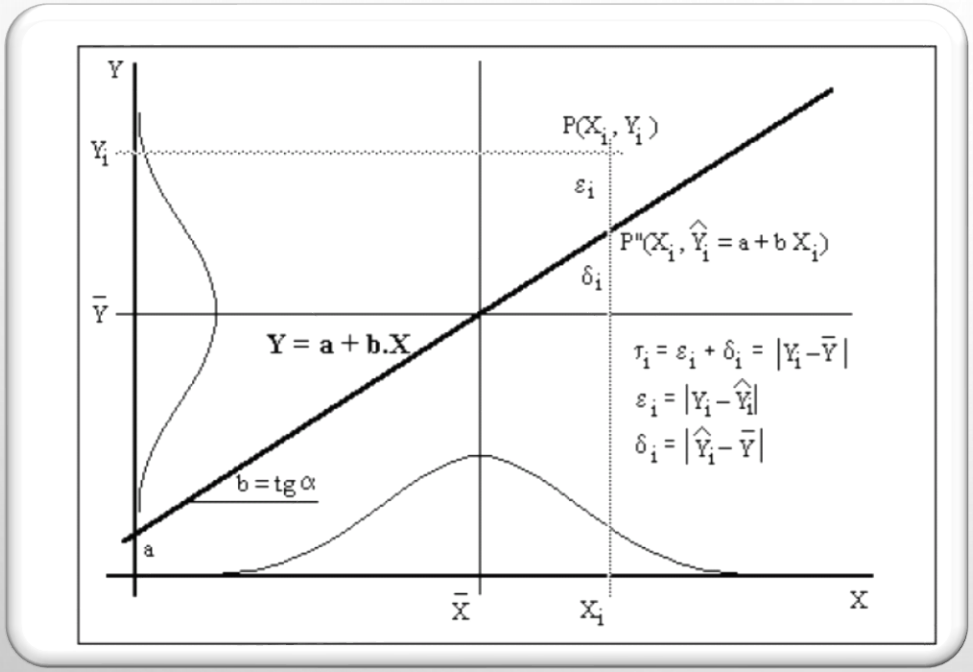
$y = a \cdot x / (k + x)$

DREAPTA DE REGRESIE:

- ***DEFINITIE:***

  DREAPTA CARE TRECE *“CEL MAI BINE”* PRINTRE PUNCTE

- ***PARAMETRII DREPTEI: Y = A + B X***
  - A = ORDONATA LA ORIGINE (INTERCEPT)
  - B = PANTA (SLOPE)



DREAPTA DE REGRESIE: (REPREZENTARE GRAFICA)

DREAPTA DE REGRESIE:

- **TEHNICA DE FITARE – METODA CELOR MAI MICI PATRATE (GAUSS)**
  - **PENTRU DETERMINAREA COEFICIENTILOR A SI B DIN ECUATIA DREPTEI VOM CONSIDERA DREAPTA CARE TRECE PRINTRE PUNCTELE EXPERIMENTALE CA AVAND SUMA PATRATELOR ABATERILOR MINIMA**

$$SSE = \sum \varepsilon_i^2 = min.$$

Aplicarea testului:

În excel: Din meniul Data -> Data Analysis -> Regression

În Epi Info: Din meniul Analyze Data-> Linear Regression

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**

Întrebări?

moleriu.lavinia@umft.ro