UMF
UNIVERSITATEA DE MEDICINĂ ȘI FARMACIE
IULIU HAȚIEGANU
CLUJ-NAPOCA

Testarea ipotezelor statistice
Analiza de corelație și regresie

Cuprins

» Analiza corelației dintre **2 VARIABILE CANTITATIVE**:
  - Corelații liniare versus corelații monotone
  - Regresia liniară simplă

Corelaţii liniare

Formularea problemei:

Studiul **relației / legăturii / corelației** dintre 2 VARIABILE CANTITATIVE:

- Greutatea (kg) și Tensiunea arterială sistolică (mmHg)?
- Nivelul de inteligență (IQ) și mărimea creierului?

Greutate = X: $X_1, X_2,,,,, X_n$
TAS = Y: $Y_1, Y_2,,,,, Y_n,$

0. Evaluarea <u>grafică</u> a relației:

- Nor de puncte (Scatter plot)

1. In<u>tensitatea</u> / forța/magnitudinea relației

- Coeficient de correlatie Pearson sau Spearman

2. Testarea relației dintre variabilele X si Y pe populație

- Test statistic pentru coeficientul de corelație

3. <u>Predicție</u> : predicția valorilor unei variabile în funcție de valorile celeilalte

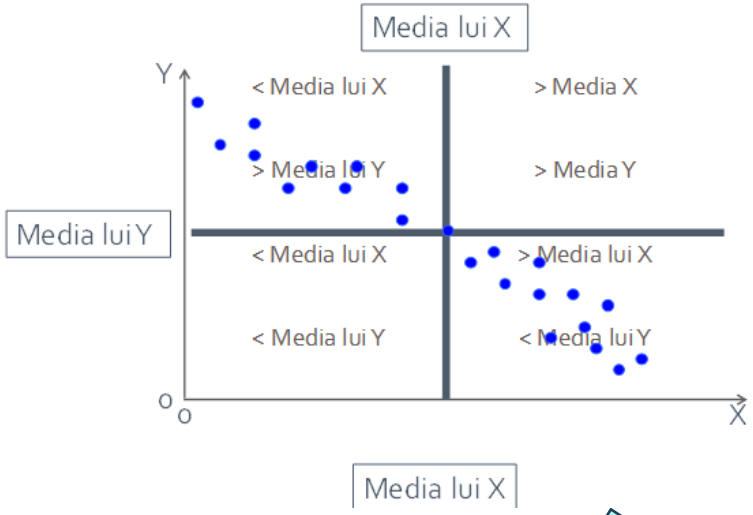
- Regresia liniară

Relaţia dintre variabile: coeficienţi de corelaţie

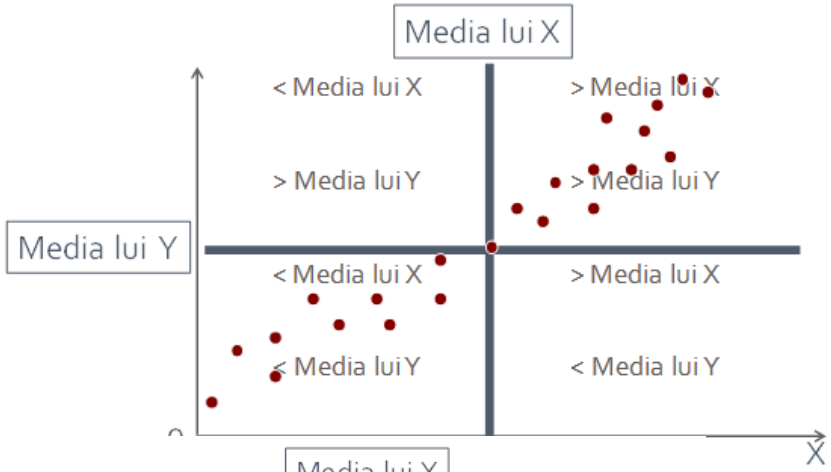
- 2 coeficienţi de corelaţie:
- Pearson și Spearman

	Pearson	Spearman
Simbol	R / r	ρ (rho)
Distribuţia	Fiecare dintre variabile are o distribuție normală de probabilitate	Orice distribuţie
Relaţie	Liniară	Monotonă

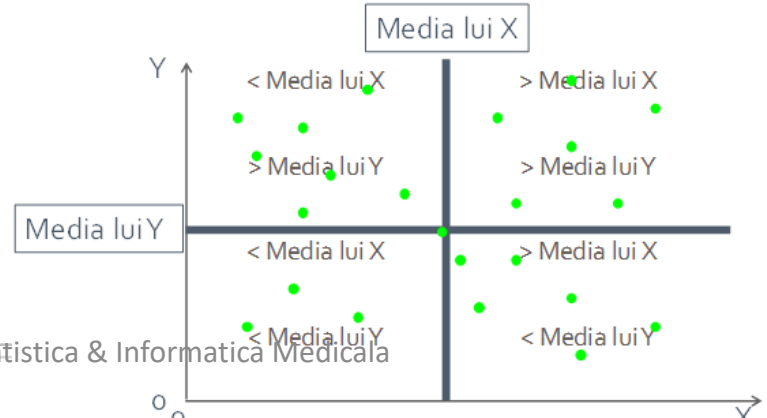
- Reprezentarea grafică: nor de puncte



corelatie liniară negativă



corelatie liniară pozitivă



lipsa corelației liniare

Calcul coeficient de corelație Pearson

Formula de calcul

Covarianța de eșantionare COV (X,Y): $$COV(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Coeficientul de corelație liniară Pearson: $$r = \frac{COV(X,Y)}{S_X \cdot S_Y}$$

Coeficient de determinare $d = r^2$

X_i , Y_i = valorile celor două variabile X si Y, $\overline{X}$ si $\overline{Y}$ sunt mediile celor două variabile; n= talie esantion; S_x si S_y = deviațiile standard de eșantionare
r = coefficient de corelatie Pearson, d = coefficient de determinare

Coeficient de corelație Pearson - interpretări

Coeficientul de corelatie Pearson (notație r): direcția și intensitatea relației dintre două variabile cantitative;

Interpretarea direcției/sensului/tendinței:

- dacă r> 0 ⇒ tendință crescătoare/ pantă ascendentă/ legatură de directă proporționalitate/ corelație pozitivă;
- dacă r< 0 ⇒ tendință descrescătoare/ pantă descendentă/ legatură de inversă proporționalitate/ corelație negativă;
- dacă r≈0 ⇒ nici o corelație liniară

Interpretarea intensității relației dintre variabile

- cu cât **r** este mai mare (în valoare absolută) cu atât relația dintre variabile este mai puternică;
- cu cât **r** este mai aproape de 0, cu atât relația este mai slabă/ puțin importantă,

Coeficient de corelație Pearson - interpretări

- **Regulile lui Colton:**
  - dacă r apartine intervalului [-0,25; 0,25] ⇒ Nu există nici o corelație liniară între cele două variabile sau există o corelatie liniară neglijabilă
  - dacă r apartine intervalului [-0,50; -0,25) sau (0,25; 0,50] ⇒ corelație liniară slabă spre acceptabilă
  - dacă r apartine intervalului [– 0,75; -0,50) sau (0,50; 0,75] ⇒ corelație liniară moderată spre bună
  - dacă r apartine intervalului [– 1; -0,75) sau (0,75; 1] ⇒ corelație linară foarte bună

Exemplu de calcul

- Pe un eșantion de 10 pacienți adulți s-au măsurat variabilele Greutate (kg) și HDL colesterol (mg/dL)-vezi tabelul de mai jos, Se stie că valoarea covarianței dintre cele două variabile este COV (Greutate, HDL)=14,32 iar deviațiile standard de eșantionare sunt S_{HDL}=1,07 mg/dL; $S_{Greutate}$=14,6 kg, Calculați coeficientul de corelație liniară.

HDL (mg/dL)	6,8	5,3	4,3	5,0	7,1	5,5	3,8	4,6	4,0	6,0
Greutatea (kg)	90	75	70	73	110	67	60	65	59	80

Soluție:

Interpretare :

r >0 => tendință crescatoare / relatie de directă proporționalitate,

r>0,75 => intensitatea corelației liniare este foarte bună (regulile lui Colton)

Test statistic de semnificație a coeficientului de corelație Pearson

- **Scopul testului**: testează dacă între cele două variabile cantitative există o relație de corelație (sau asociere) semnificativă statistic
- **Condiții de aplicabilitate**:
    - Observații independente în eșantion
    - Variabile cantitative
    - Cele două variabile sunt normal distribuite
    - Fără valori aberante (foarte îndepartate de norul de puncte)

Test statistic de semnificație a coeficientului de corelație Pearson

- **Ipoteza nulă (H0)**
  - **Nu există o corelație liniară semnificativă statistic între variabilele X și Y pe populația de interes**
- **Ipoteza alternativă (H1):**
  - **Există o corelație liniară semnificativă statistic între variabilele X și Y pe populația de interes**
- **Decizia testului cu ajutorul valorii p:**

✓ dacă **$p < \alpha$ (0,05)** => se respinge H_0 → suntem în favoarea lui H_1 → cu un risc de eroare de 5%, există o corelație liniară între variabilele X și Y pe populația de interes

✓ dacă **$p \geq \alpha$ (=0,05)** => nu se respinge H_0

Test statistic de semnificație a coeficientului de corelație Pearson

Exemplul1:

Coeficientul de corelație pentru relația dintre Trigliceride (mg/dL) și Greutate(kg) pentru 50 de subiecți adulți cu diabet DZ2 este 0,72, iar valoarea p asociată este 0,001. Perechile de observații sunt independente, datele au avut distribuție normală, relația este liniară.

Interpretarea valorii p asociate coeficientului de corelație:

- p <0,05 → **există o corelație semnificativă statistic între variabilele** trigliceride și greutate pe populația de adulți cu DZ2 (Corelația dintre trigliceride și greutate este **semnificativă statistic)**

Interpretarea direcției și intensității corelației:

- r = 0,72> 0 → relația este de proporționalitate directă
- r = 0,72 este în [0,50 și 0,75) → intensitatea corelației este moderată până la bună

Test statistic de semnificatie a coeficientului de corelatie Pearson

Exemplul2:

Coeficientul de corelație pentru relația dintre Trigliceride (mg/dL) și Glicemie (mg/dL) pentru 24 de subiecți cu varsta între 20-30 ani este de 0,39, iar valoarea p asociată este de 0,35. Perechile de observații sunt independente, datele sunt normal distribuite, relația este liniară.

Interpretarea valorii p asociate coeficientului de corelație:

- p> 0,05 → nu am găsit o **corelație semnificativă statistic** între Trigliceride și Glicemie pe populația de adulți cu varsta între 20-30 ani (formulare echivalentă: nu avem suficiente dovezi statistice pentru a demonstra existența corelației liniare semnificative între trigliceride și glicemie pe populația de interes

Interpretarea direcției și intensității corelației:

- interpretarea este inutilă deoarece relația nu a atins semnificația statistică - deci rezultatul poate fi influențat foarte mult de hazard (șansă). Pe eșantion, relația pare a fi de directă proporționalitate (r = 0,39> 0), iar intensitatea corelației este scăzută / acceptabilă (r = 0,39 - este în [0,25 și 0,50)).

Coeficientul de corelație Spearman

Scop: evaluarea **asocierii** dintre 2 variabile din punct de vedere al **direcției și magnitudinii** relației dintre acestea

Conditii de aplicare:

- Observații independente în eșantion;
- Cele 2 variabile sunt variabile ordinale/cantitative (normal distribuite sau fără distribuție normală)

Utilitate: evaluarea relației dintre

- două variabile ordinale
- o variabilă ordinală și una cantitativă
- două variabile cantitative care nu sunt normal distribuite

Coeficient de corelatie Spearman - interpretari

Interpretarea direcției/sensului/tendinței:

- dacă $\rho > 0 \rightarrow$ corelație pozitivă
- dacă $\rho < 0 \rightarrow$ corelație negativă
- dacă $\rho \approx 0 \rightarrow$ nici o corelație

Interpretarea intensității relației

- cu cât ρ este mai mare (în valoare absolută) cu atât relația dintre variabile este mai puternică
- cu cât ρ este mai aproape de 0, cu atât relația este mai slabă/ puțin importantă

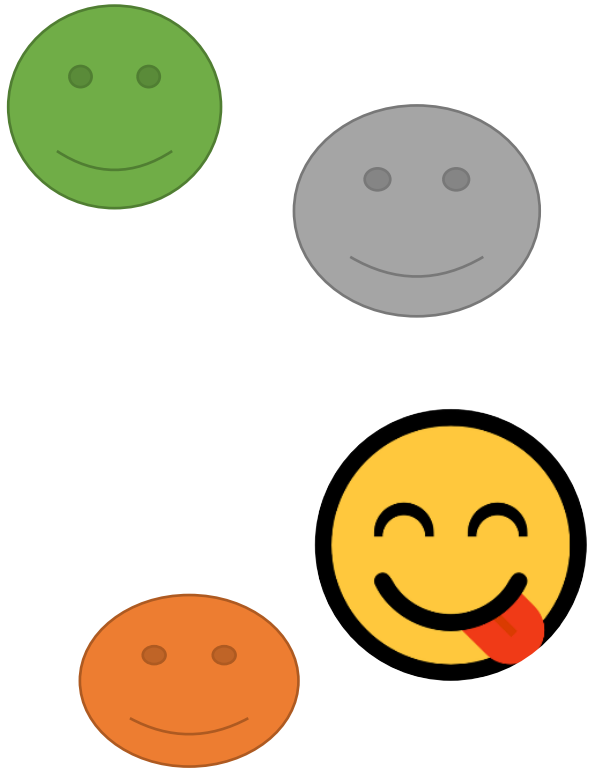
Test statistic de semnificație a coeficientului de corelație Spearman

- **Ipoteza nulă (Ho)**
  - **Nu există o corelație semnificativă statistic între variabilele X si Y pe populația de interes**
- **Ipoteza alternativă (H1):**
  - **Există o corelatie semnificativă statistic între variabilele X și Y pe populația de interes**
- **Decizia testului cu ajutorul valorii p:**

✓ dacă **p < α (=0,05) => se respinge H_0 => suntem in favoarea lui H_1 => cu un risc de eroare de 5%, există o corelație semnificativă statistic între variabilele X si Y pe populația de interes**

✓ dacă **p≥ α (=0,05) => nu se respinge H_0**

WOOCLAP -QUIZZ



- accesați link-ul: **www.wooclap.com**
- introduceți codul:
- **CODE: AMGC9**

Regresia liniară

Tipuri de Regresii

Clasificare în funcție de:

- **tipul variabilei dependente**
  - **variabila cantitativă → regresie liniară**
  - Variabila calitativa dihotomială – regresie logistică
- **Numărul de variabile dependente:**
  - **regresie univariată (o singură variabilă dependentă)**
  - Regresie multivariată (>=2 variabile dependente)
- **Numărul de variabile independente :**
  - regresie **simplă** (o variabilă independentă)
  - regresie **multiplă** (>=2 variabile independente)

Regresia liniară simplă

Scop: evaluarea relației liniare dintre o variabilă independentă și o variabilă dependentă cantitativă

Obiective:

- **predicția valorilor** unei variabile cantitative ca funcție a valorilor variabilei independente
- **evaluează**
    - **Existența** legăturii/**dependenței**/relației dintre 2 variabile
    - **Direcția** legăturii/relatiei dintre 2 variabile
    - **Importanta/magnitudinea** legăturii/relației dintre 2 variabile

Regresia liniară simplă

- **Ecuația dreptei de regresie:** $Y = b_0 + b_1 X$

  b_0 = ordonata în origine: valoarea lui Y când X=0

  b_1 = panta dreptei de regresie,

  Interpretarea lui b_1 = coeficientul variabilei X

  - **Semnul lui b_1**
    - dacă $\mathbf{b_1} > 0$ tendința crescătoare / pantă ascendentă / pantă pozitivă / legătură de directă proporționalitate
    - dacă $\mathbf{b_1} < 0$ tendința descrescătoare / pantă descendentă / pantă negativă / legătură de inversă proporționalitate
    - dacă $\mathbf{b_1} \approx 0$ nici o tendință liniară
  - **Magnitudinea lui b_1**
    - **La o creștere cu 1 unitate a variabilei independente X , valoarea variabilei dependente Y crește (sau scade) în medie cu b_1 unități**

Exemplu: corelație liniară

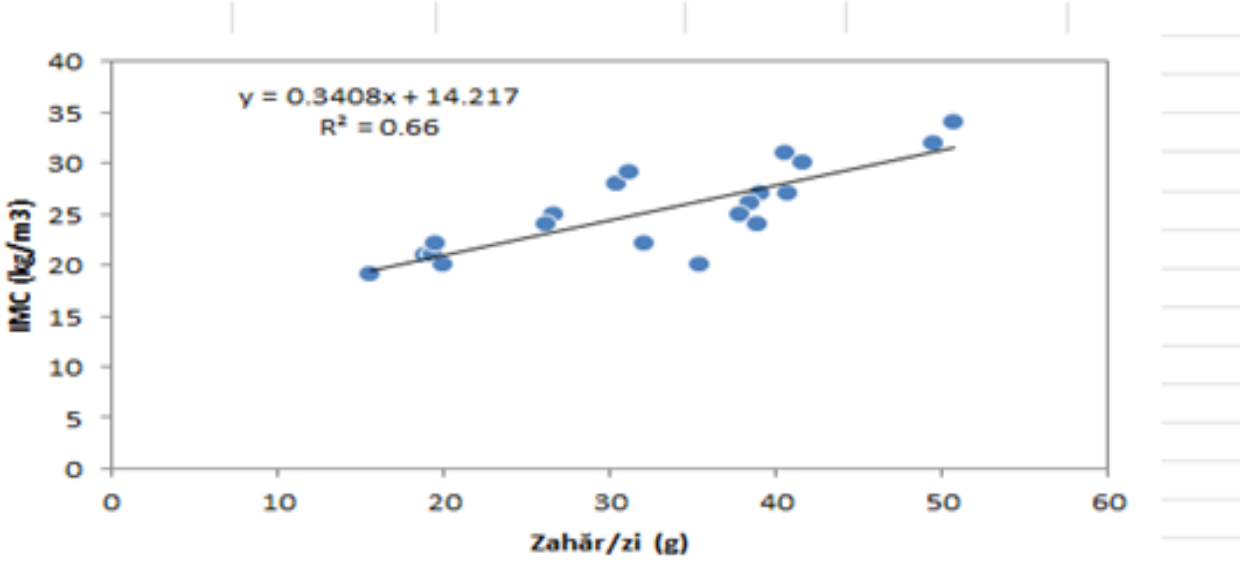
- **Un eșantion aleatoriu de 20 studenți AMG 1, UMF Cluj, a fost investigat pentru a identifica corelația liniară dintre indicele de masă corporală și consumul de zahăr pe zi (grame/zi).**
- Variabilele IMC(kg/m2) și Consumul zilnic de zahar (gr.) au avut distribuție normală
- Analiza corelației: $r = 0,812$ ($p<0,001$)

Exemplu: regresie liniară

- **Un eșantion aleatoriu de 20 studenți AMG 1, UMF Cluj, a fost investigat pentru a identifica dependența dintre indicele de masă corporală (IMC, kg/m2- variabilă dependentă) și consumul de zahăr pe zi ca variabilă independentă.**
- Rezultate:
- Analiza de regresie:
- **$IMC(kg/m^2) = 14,22+0,34\times Zahăr/zi(g)$**

Exemplu: regresie liniară

- IMC(kg/m^3) = 0,34×Zahăr/zi(g)+14,22
- R^2=0,66
- Coeficientul regresiei: la o creștere a consumului zilnic de zahăr cu 1 gram, IMC crește în medie cu 0,34 kg/m^2 . In plus, 66% (R^2=0,66) din variația IMC-ului este explicată de relația liniară a acestuia cu consumul de zahar,



De reținut!

- Testarea **corelatiei** dintre doua variabile se realizeaza prin:
  - Coeficientul de corelatie Pearson: 2 variabile cantitative normal distribuite
  - Corelația Spearman: 2 variabile de orice tip (cantitative care nu urmează o distribuție normală sau 2 variabile calitative ordinale)
  - Testarea **dependentei liniare** dintre 2 variabile se realizează prin regresia liniară simplă

Exemple de probleme

Q1. S-a realizat un studiu pe 2000 subiecţi cu vârstă de peste 55 ani pentru a identifica relaţia liniara varsta (ani) şi tensiunea arteriala sistolica(TAS, mmHg), S-a obţinut un coeficient de corelaţie liniara (r) între varsta şi TAS de 0,55.

Care din următoarele afirmaţii sunt corecte?

A. Există o relaţie de dependenţă negativă între vârsta şi TAS

B. Corelaţia este pozitiva; TAS tinde să fie mai mare pentru subiecţii cu varsta mai mare, intensitatea corelatiei fiind moderata spre buna

C. Nu există relaţie între vârsta şi TAS

D. Corelaţia este negativă; TAS tinde să fie mai mare pentru subiecţii cu varsta mai mica, intensitatea corelatiei fiind moderata spre buna

E. Coeficientul de corelatie folosit este coeficientul de corelatie al lui Pearson

R1: B, E

Exemple de probleme

Q2. S-a realizat un studiu pe 2000 subiecţi cu vârstă de peste 55 ani pentru a identifica relaţia liniara varsta (ani) şi tensiunea arteriala sistolica(TAS, mmHg), Covarianța dintre varsta si TAS a fost egala cu 10,24 iar deviatiile standard de esantionare au fost egale cu 5 ani pentru varsta respectiv 4 mmHg pentru TAS, Presupunem ca cele doua variabile au o distributie normala.

Care din următoarele afirmaţii sunt corecte?

A. Coeficientul de corelatie Pearson a fost egal cu r=0,512

B. Nu este posibila determinarea nici unui coeficient de corelatie

C. Nu există relaţie între vârsta şi TAS

D. In esantionul de studiu, ambele variabile variaza in acelasi sens

E. In esantionul de studiu, ambele variabile variaza in sens contrar

R2: A, D

Exemple de probleme

Q3. S-a realizat un studiu pe 2000 subiecţi cu vârstă de peste 55 ani pentru a identifica relaţia liniara varsta (ani) şi tensiunea arteriala sistolica(TAS, mmHg), S-a obţinut un coeficient de corelaţie liniara între varsta şi TAS de 0,55 si o valoare p asociata p<0,001.

Care din următoarele afirmaţii sunt corecte?

A. Există o corelaţie liniara semnificativa statistic între vârsta şi TAS pentru ca p<0,05

B. In esantionul de studiu există o corelaţie liniara pozitiva semnificativa statistic între vârsta şi TAS pentru ca p<0,05

C. Ipoteza nula a testului statistic a carui valoare p a fost <0,001 afirma ca: Nu există corelaţie liniara semnificativa între vârsta şi TAS

D. Ipoteza alternativa a testului statistic a carui valoare p a fost <0,001 afirma ca: Există corelaţie liniara semnificativa între vârsta şi TAS

E. In esantionul de studiu, corelatia liniara dintre variabile a fost pozitiva iar intensitatea acesteia a fost moderata spre buna

R3: A, C, D, E

Exemple de probleme

Q4. S-a realizat un studiu pe 2000 subiecţi cu vârstă de peste 55 ani pentru a identifica relaţia liniara vârstă (ani) şi tensiunea arteriala sistolica (TAS, mmHg). S-a obţinut un model de regresie liniara între varsta (ani) şi TAS de forma: TAS=100,13 + 0,365 × varsta. Care din următoarele afirmaţii sunt corecte?

A. Există o relaţie de dependenţă pozitiva între vârsta şi TAS

B. Exista o relatie de directa proportionalitate intre vârsta şi TAS pentru ca panta regresiei este pozitiva

C. In esantionul de studiu, la o crestere a varstei pacientului cu 1an, TAS creste in medie cu 0,365 mmHg,

D. Formula **TAS=100,13 + 0,365 × varsta** reprezinta rezultatul unei regresii univariate multiple

E. Pe esantionul de studiu, nu există relaţie între vârsta şi TAS

R4: A, B, C

MULȚUMESC PENTRU ATENȚIE!