CORELAŢII SI REGRESII

LEGĂTURA ÎNTRE MAI MULTE VARIABILE

Tipuri de “relaţii“

- două sau mai multe variabile cantitative
- două variabile de ordine
- două variabile calitative
- o variabilă cantitativă cu o variabilă calitativă

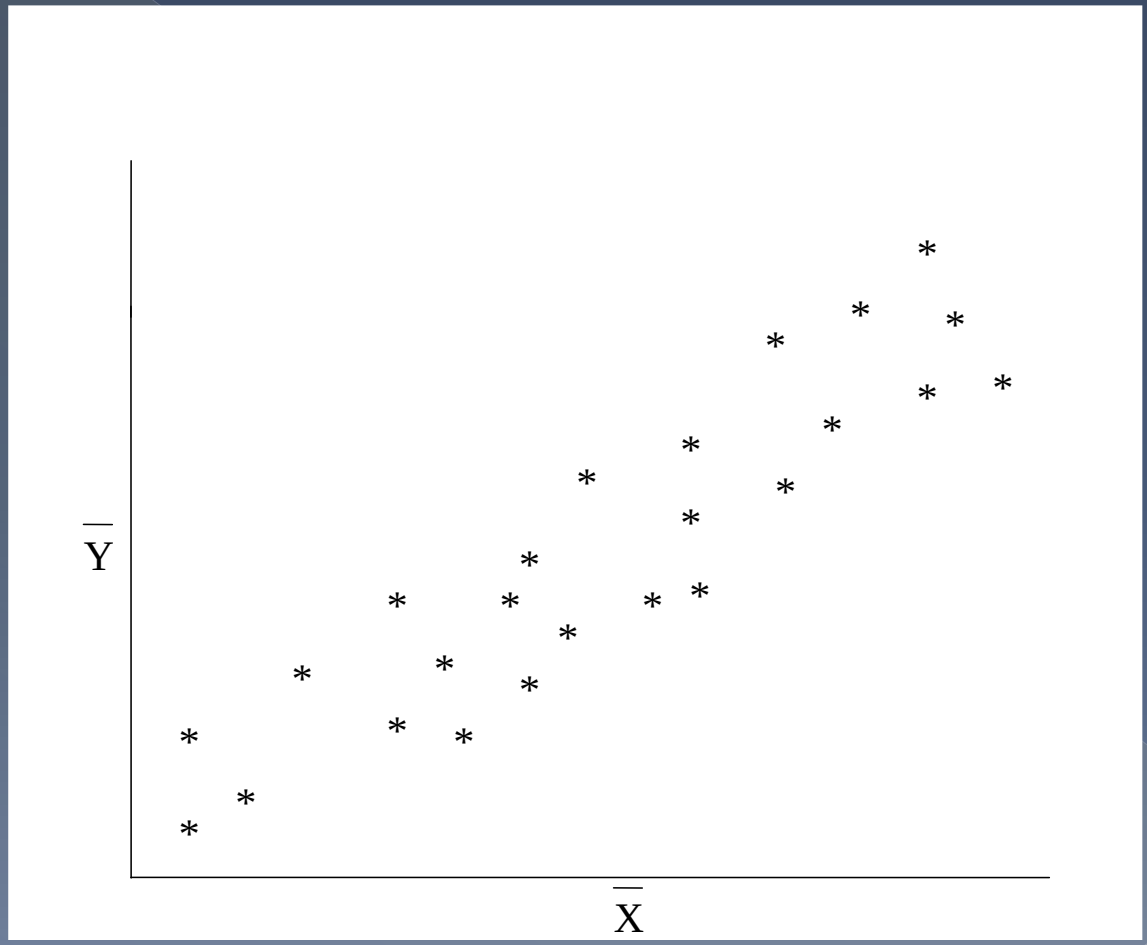
Variabile cantitative

Varsta X: $X_1, X_2, ..., X_n$

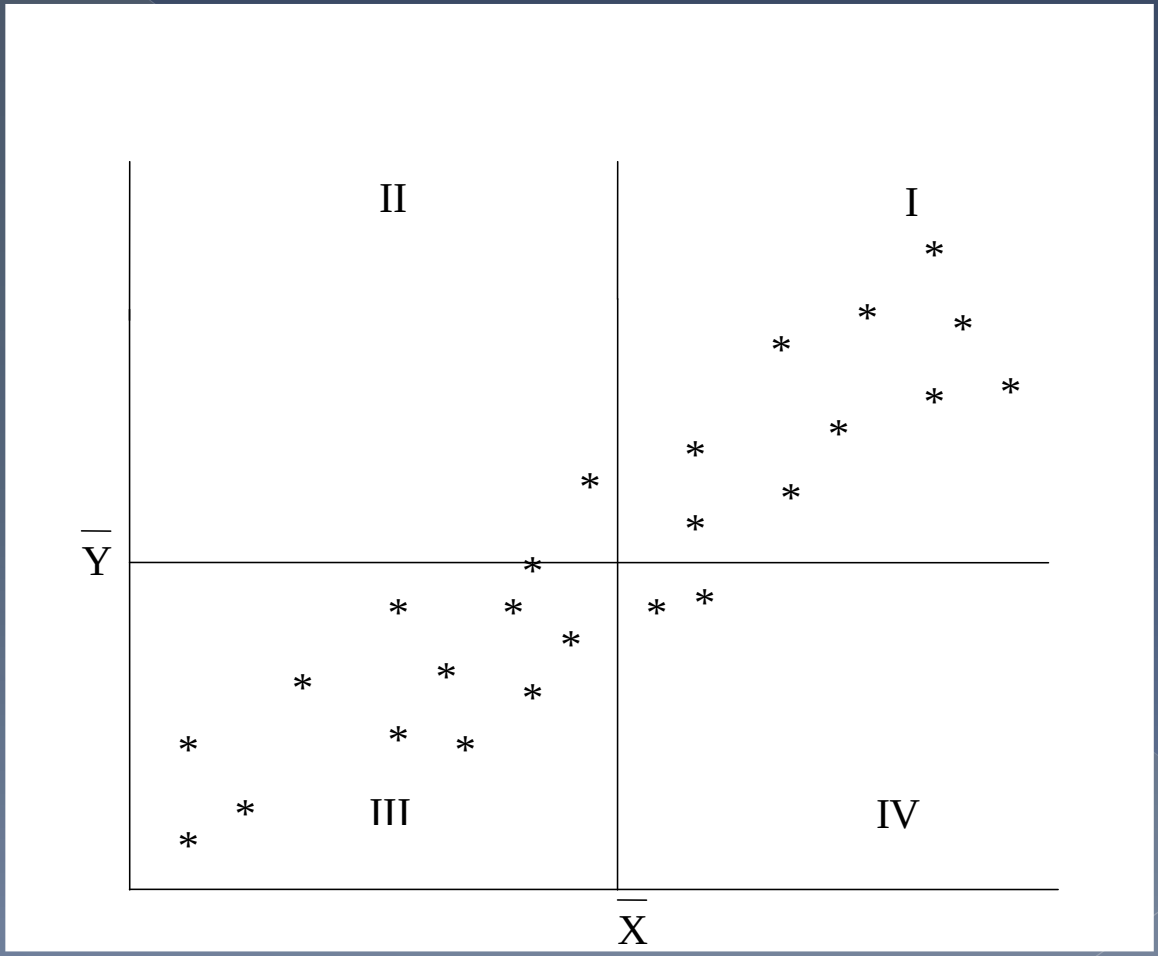
TAS Y: $Y_1, Y_2, ..., Y_n$.

- *1. Să se stabilească dacă există o legătură între variabilele X şi Y (cantitative continue) şi să se determine o modalitate de a măsura intensitatea acestei legături.*
  - *Coeficientul de corelaţie*
- *2. Să se stabilească dacă Y depinde de X şi dacă da în ce formă se realizează această dependenţă.*
  - *Funcţia de regresie*

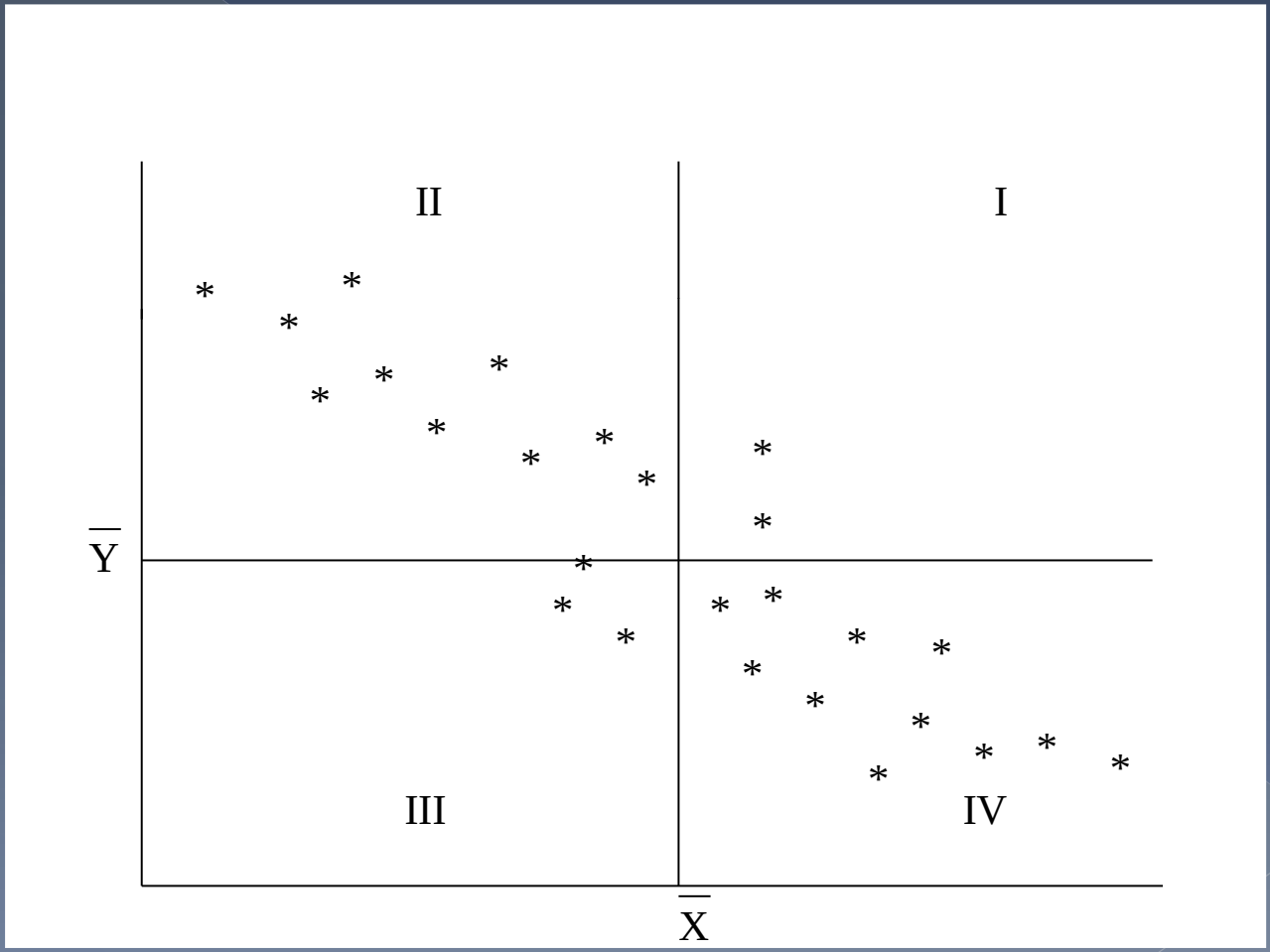
Statistici descriptive in două dimensiuni. Diagrama de dispersie



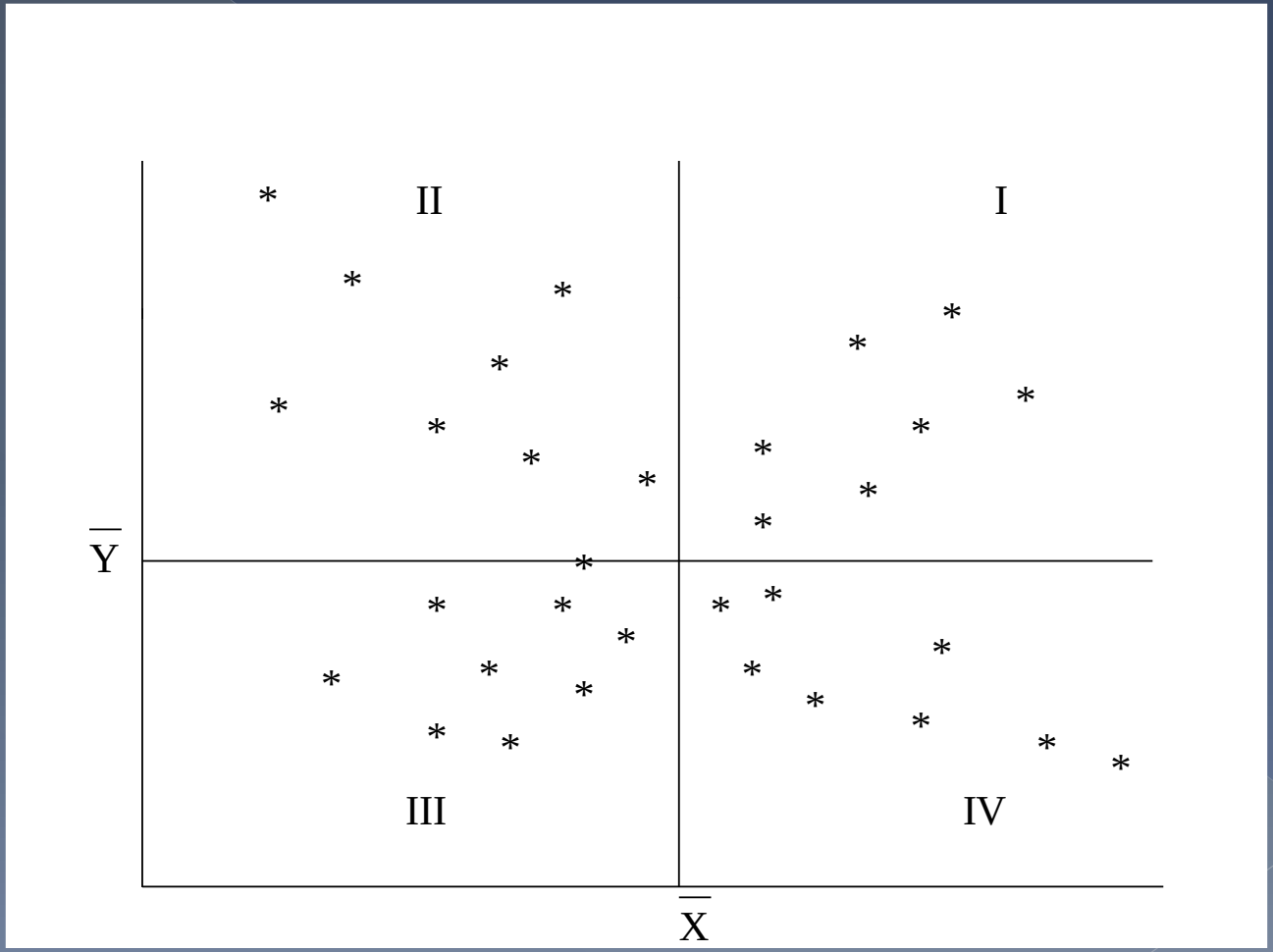
Statistici descriptive in două dimensiuni. Diagrama de dispersie



Statistici descriptive in două dimensiuni. Diagrama de dispersie



Statistici descriptive in două dimensiuni. Diagrama de dispersie



Indici de corelație. Suma produselor ecart

- **Descrierea "intensitatăţii" relaţiei dintre variabilele X şi Y:**
  - **(X_i,Y_i) -în cadranele I sau III:** $(X_i - \overline{X})(Y_i - \overline{Y}) \geq 0$
  - **(X_i,Y_i) -în cadranele II sau IV:** $(X_i - \overline{X})(Y_i - \overline{Y}) \leq 0$

$$SPE = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Indici de corelație. Covarianța

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

Avantaje: mărime independentă faţă de volumul seriei statistice.

Indici de corelație. Coeficientul de corelație

Pentru a obține un indicator independent şi de unitățile de măsură ale celor două variabile se utilizează coeficientul de corelație sau coeficientul Bravais-Pearson:

$$r = \frac{COV(X,Y)}{Sx \cdot Sy}$$

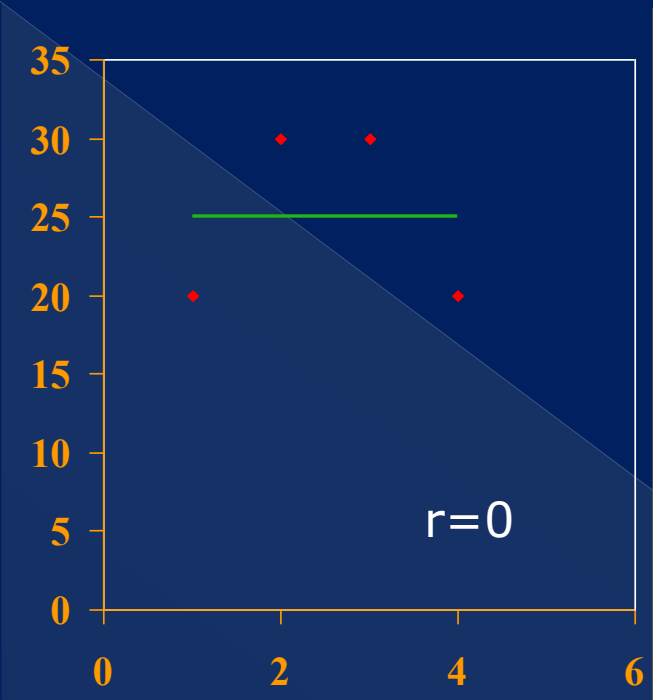
S_X şi S_Y reprezintă abaterile standard pentru seriile X şi respectiv Y:

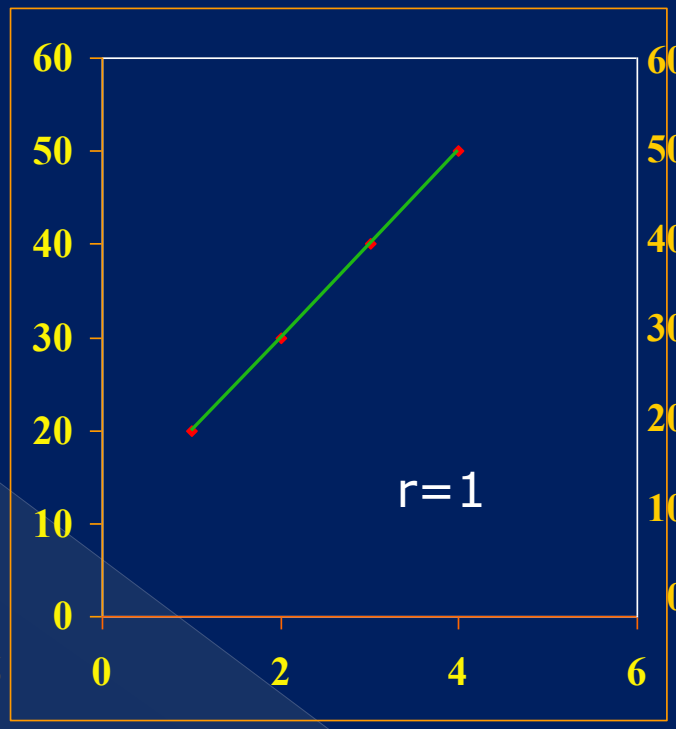
$$S = \sqrt{s^2}$$

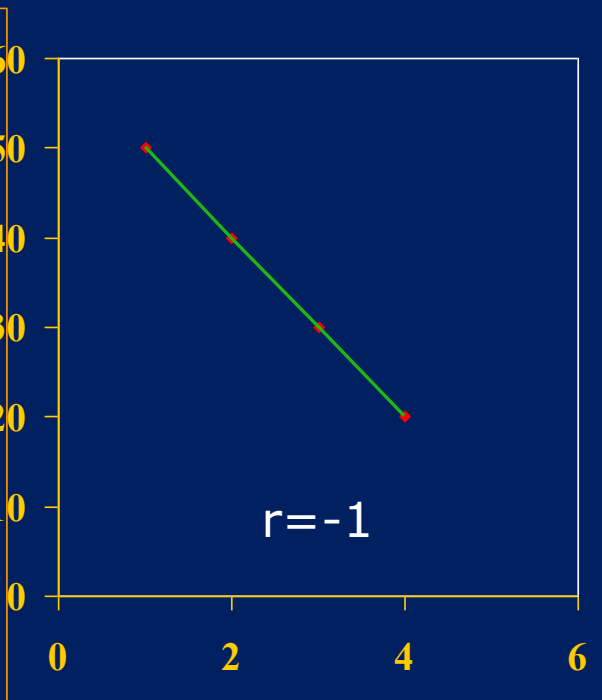
$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

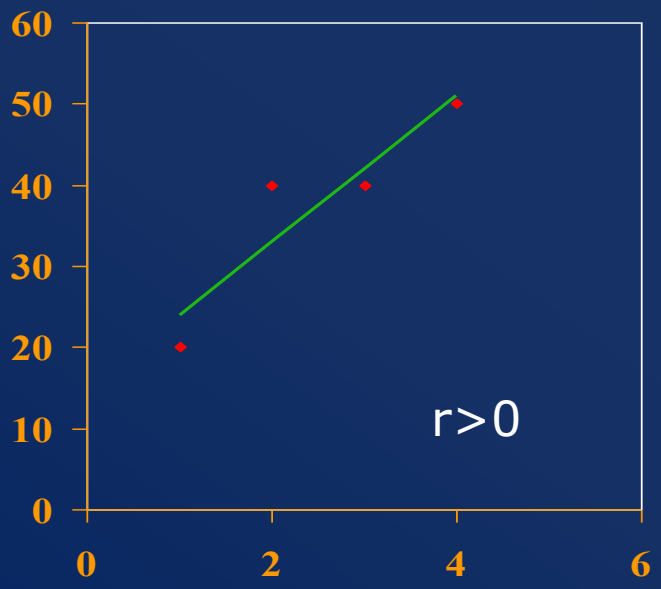
Coeficientul de corelație -proprietăți

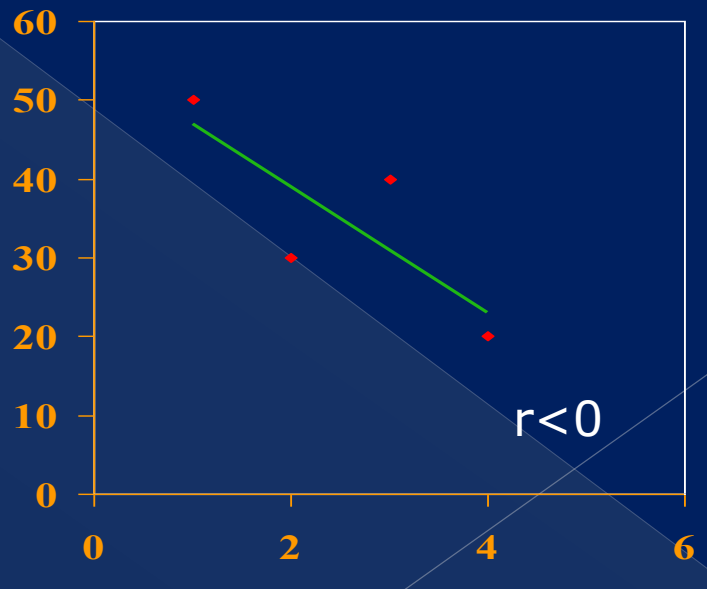
- $r \in [-1, 1]$
- măsoară intensitatea relației dintre variabilele X şi Y
-
- dacă $0 < r < 1$, norul de puncte poate fi înlocuit (ajustat) printr-o dreaptă de pantă pozitivă .
- Dacă $-1 < r < 0$ atunci norul de puncte poate fi aproximat cu o dreaptă de pantă negativă. Dispersia punctelor față de dreaptă va fi cu atât mai mică cu cât r este mai apropiat de -1.


35
30
25
20
15
10
5
0
0
2
4
6
r=0


60
50
40
30
20
10
0
0
2
4
6
r=1


60
50
40
30
20
10
0
0
2
4
6
r=-1


60
50
40
30
20
10
0
0
2
4
6
r>0


60
50
40
30
20
10
0
0
2
4
6
r<0

Testul de semnificație pentru coeficientul de corelație Pearson

- Semnificația coeficientului de corelație Pearson poate fi evaluată dacă valoarea observată a apărut datorită întâmplării (dacă este semnificativ diferită de zero).

Coeficintul de corelaţie -interpretare

Regulile empirice ale lui Colton (1974)

Valoarea r	p > 0,05	p < 0,05
in (-0.25 ; 0,25)	corelație slabă sau nulă	corelație slabă sau nulă
in [0.25 ; 0.50) sau in (-0.50 ; -0,25]	Nu are semnificatie statistica	Grad de asociere acceptabil
in [0.50 ; 0.75) sau in (-0.75 ; -0,50]	Nu are semnificatie statistica	O corelație moderată spre bună
>0.75 sau < -0,75	Nu are semnificatie statistica	O foarte bună asociere sau corelație
>1 sau <-1	Eroare	Eroare

Interpretarea r,p

Atentie:

- **Variabilele corelate trebuie sa fie cantitative**
- **Intre variabilele pentru care se calculeaza corelatia trebuie sa existe o relatie de cauzalitate**

Coeficientul de corelație-interpretare Alura norului de puncte

- r >0
  - O creştere a lui X determină o creştere a lui Y (direct proporționale).
- r < 0
  - O creştere a lui X determină o diminuare a lui Y (invers proporționale)

Indici de corelație. Coeficientul de determinare

- **$d = r^2$**
- Reprezintă partea din variația totală a lui Y explicată prin relația liniară existentă între X şi Y.

- Cazuri particulare:
  - d=1:
  - d=0:

  Dacă d este exprimat în procente: d reprezintă procentul în care variația lui Y este dată prin relația liniară între cele două variabile.

Două variabile ordinale (sau o variabilă ordinală şi una cantitativă) Coeficientul de corelație al lui Spearman

Se procedeaza astfel:

i. Se inlocuieste seria bivariata ($x_1,\ldots,x_n;y_1,\ldots,y_n$) cu seria rangurilor ($R_{x1},\ldots R_{xn};R_{y1},\ldots,R_{yn}$), valorilor x_i si y_i dupa ordonarea lor in ordine crescatoare (pentru valorile egale se ia media aritmetica a rangurilor).

ii. Pentru determinarea coeficientului r_s al lui Spearman se calculeaza coeficientul de corelatie (Pearson) pentru seria rangurilor.

Analiza chestionarelor Reproductibilitatea

- **Indicele Kappa** de concordanța între observatori
- Coeficientul de corelație interclase
- Coeficientul **Cronbach alfa** de consistență internă

Analiza chestionarelor Validitatea

- Validitatea de continut (ceea ce testam (intrebarile) este in acord cu cunostintele care sunt testate)
- Validitatea de criteriu (daca e in acord cu alte masuratori cu care ar trebui sa fie in relatie)
- Validitatea de construct (masuratoarea e in acord cu caracteristica masurata si nu cu alte caracteristici)

Statistici descriptive in două dimensiuni. Drepte de regresie

- Dreapta de regresie Y(X):

  y = a + b x

Dreapta de regresie Y(X)

$$\min_{a,b\in R} \sum_{i=1}^{n} (a + bX_i - Y_i)^2$$

Valorile lui a şi b pentru care este atins minimul sumei sunt date prin formulele:

$$b = \frac{COV(X,Y)}{S_X}.$$

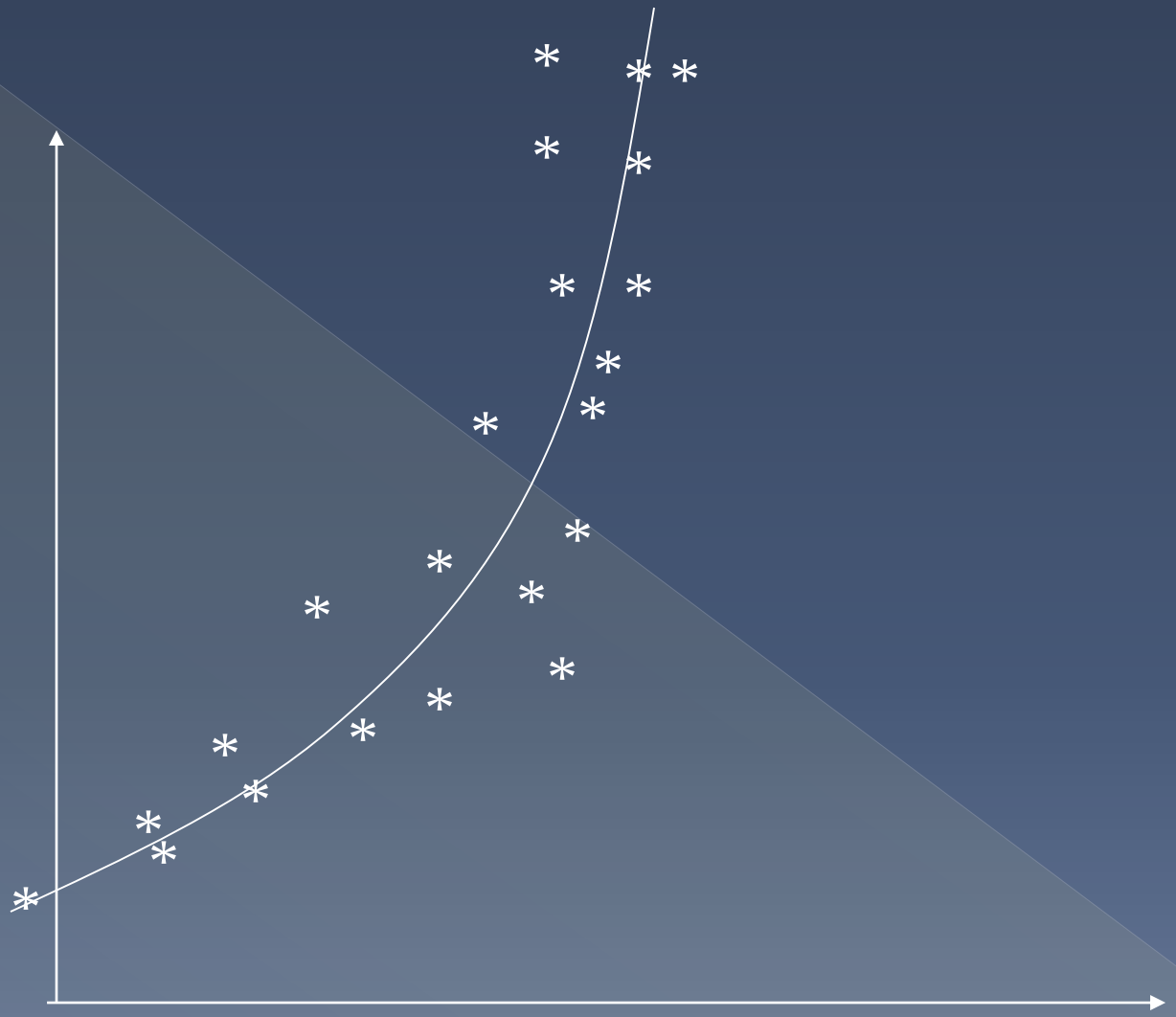
$$a = \overline{Y} - b \cdot \overline{X}$$

Utilizarea funcțiilor de regresie

- **Extrapolare şi interpolare**
  - Când se determină valoarea funcţiei (adică a lui Y), pentru un X cuprins intervalul [X_{min}, X_{max}], atunci se efectuează o operaţie de interpolare, iar când X se află în afara intervalului se spune că este vorba de o extrapolare.
- **Prezicerea lui Y pentru un X dat**
- **Simulari**

Funcția de regresie. Schimbări de variabile

- In unele cazuri se constată că relaţia liniară pare a nu fi adecvată pentru descrierea dependenţei dintre variabilele X şi Y, sau că scalele utilizate nu sunt cele mai potrivite.



Regresii multidimensionale

Fiind date variabilele:

X_i: $X_{i1},...,X_{in}$, i=1,2,...,m

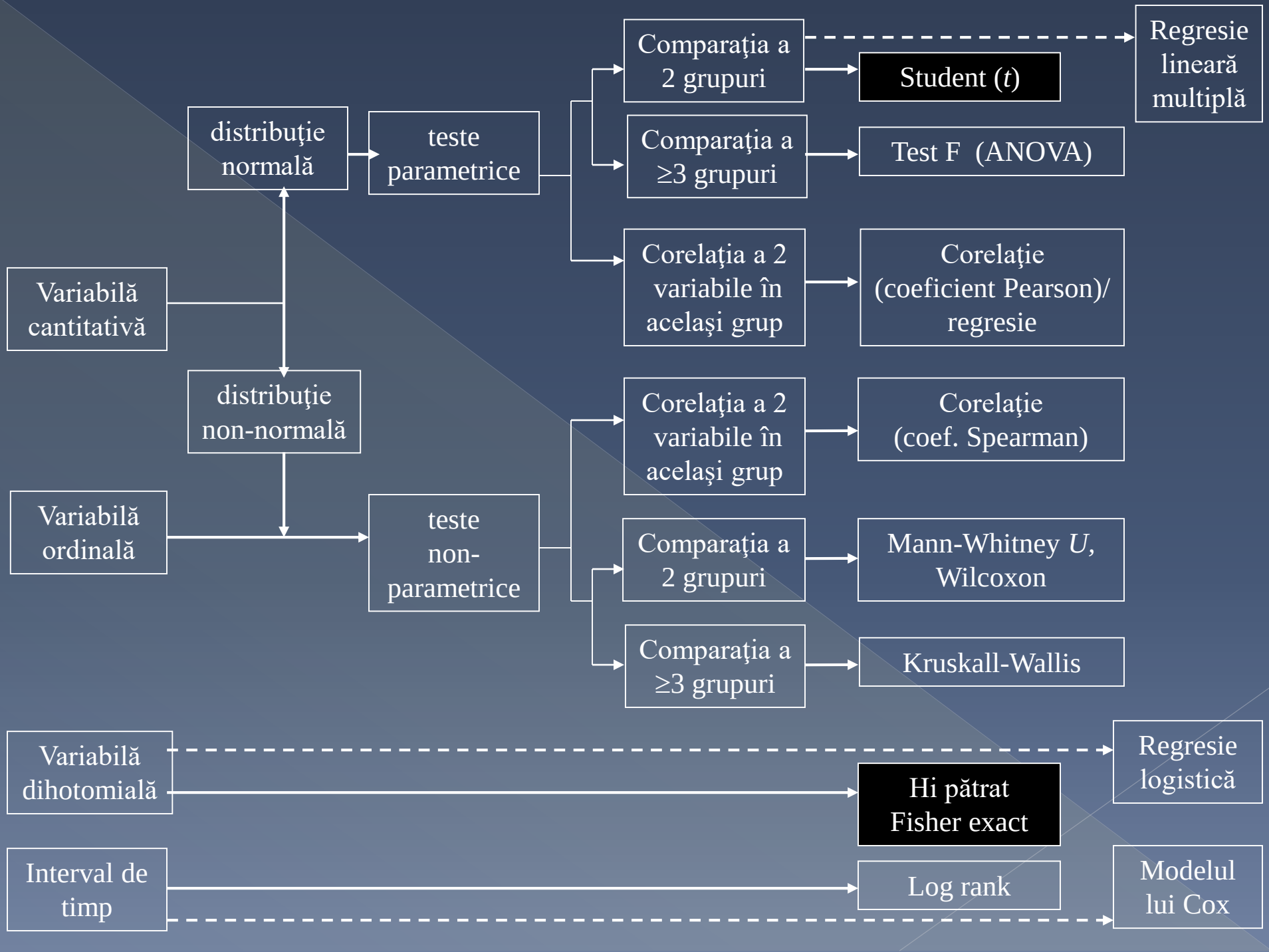
Y: $Y_1,..., Y_n$

se caută o relaţie de forma:

$$Y = a + b_1X_1+...+b_mX_m,$$

unde coeficienţii a si b_i (i=1,...,m) se determină astfel încât să minimizeze expresia:

$$\sum_{i=1}^{n}(Y_i - (a + b_1X_{1i}+...+b_mX_{mi}))^2 .$$


Variabilă cantitativă
distribuţie normală
distribuţie non-normală
Variabilă ordinală
teste parametrice
teste non-parametrice
Comparaţia a 2 grupuri
Comparaţia a ≥3 grupuri
Corelaţia a 2 variabile în acelaşi grup
Corelaţia a 2 variabile în acelaşi grup
Comparaţia a 2 grupuri
Comparaţia a ≥3 grupuri
Student (t)
Test F (ANOVA)
Corelaţie (coeficient Pearson)/ regresie
Corelaţie (coef. Spearman)
Mann-Whitney U, Wilcoxon
Kruskall-Wallis
Regresie lineară multiplă
Variabilă dihotomială
Interval de timp
Hi pătrat Fisher exact
Log rank
Regresie logistică
Modelul lui Cox