

## Probabilități și Statistică - Curs 13

30 Mai, 2017

## 1 Corelația liniară

Corelația - un exemplu

Coeficientul de corelație

Linia deviației standard (SD)

Sumar

## 2 Regresia Liniară

Linia de regresie

Linia de regresie - exemple

## 3 Bibliografie

- *Corelația* este o metodă destinată studiului relației dintre două variabile, fiind parte a *statisticii bivariate*.
- Primul statistician care a făcut un progres notabil în acest domeniu a fost Francis Galton care a încercat să studieze gradul în care copiii se aseamănă părinților lor.
- Tendința de a studia influențele eredității prin intermediul instrumentelor statistice și matematice a fost caracteristică epocii victoriene.
- Parte a unui studiu întreprins de Karl Pearson (un discipol al lui Galton), a fost măsurată înălțimea a 1078 tați și fii ajunși la maturitate.
- Relația dintre două variabile (înălțimea tatălui și a fiului) poate fi exprimată vizual într-o diagramă bidimensională.

- Noi vom întreprinde un studiu legat de relația dintre punctajele obținute de studenții FII la examenul de Probabilități (din săptămâna 8-a) și cele de la examenul de Statistică (din săptămâna a 15-a).
- Punctajele a 335 de studenți din anul școlar 2014/2015 și a 355 de studenți din 2015/2016 care au susținut ambele examene sunt reprezentate în următoarele două diagrame.
- Fiecare punct reprezintă o pereche de punctaje: pe coordonata  $x$  cel de la Probabilități, iar pe  $y$  punctajul de la Statistică.
- Am reprezentat de asemeni prima bisectoare; această dreaptă corespunde studenților care au avut același punctaj la ambele examene.
- Dacă punctajul unui student la Probabilități este apropiat de cel de la Statistică, atunci punctul corespunzător este situat aproape de prima bisectoare.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

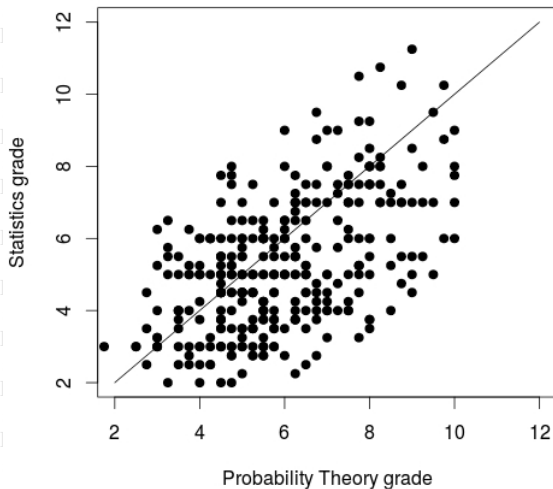
Probabilități

Probabil

Probabilități

Probabil

2014/2015



i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

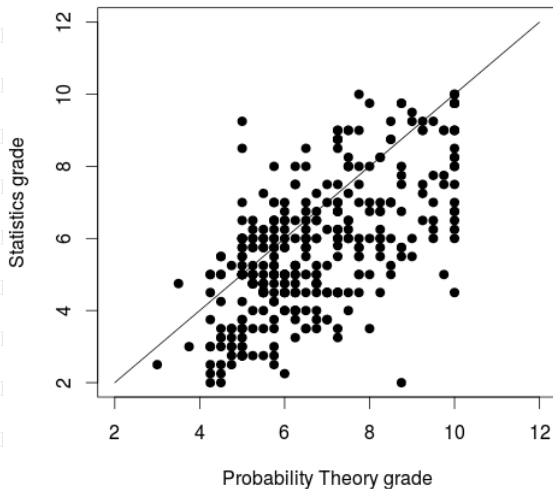
Probabil

Probabilități

Probabil

Probabilități

2015/2016



i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

- Punctele aflate sub prima bisectoare corespund studenților care au avut un punctaj mai bun la Probabilități: aceasta este zona unde se află cele mai multe puncte.
- Există o împrăștiere destul de mare în jurul primei bisectoare ceea ce sugerează o legătură nu foarte puternică între cele două variabile.
- Dacă ar exista o legătura foarte puternică între cele două variabile, atunci cunoscând una dintre ele o putem afla și pe cealaltă.
- Atunci când asocierea este slabă, informația despre una dintre variabile nu ajută prea mult la aflarea celei de-a doua.
- Vom încerca să facem o predicție despre punctajul de la Statistică din cel de la Probabilități.

- Observăm că diagrama are forma unui "nor" eliptic.
- Cum putem exprima relația dintre ce două variabile?
- Primul pas este să marcăm punctul care are drept coordonate media valorilor  $x$  și cea a valorilor  $y$ : acesta este *punctul mediilor*, aflat în centrul "norului".
- Al doilea pas ar fi să măsurăm împrăștierea norului dintr-o parte în cealaltă. Aceasta se poate face utilizând deviațiile standard ale celor două eșantioane.
- Majoritatea punctelor se vor afla într-un interval de 2 deviații standard (pe verticală și pe orizontală).
- Aceste statistici însă nu arată în întregime puterea *asocierii* dintre două eșantioane.



Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

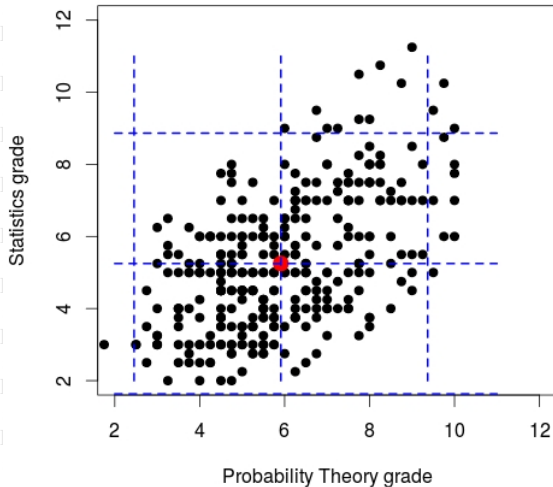
Probabilități

Probabil

Probabilități

Probabil

2014/2015



i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

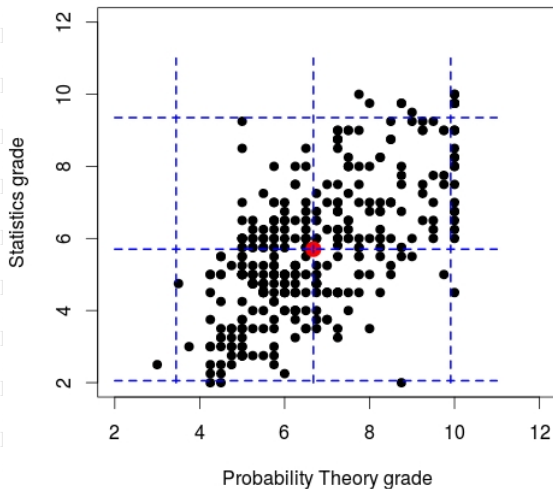
Probabil

Probabilități

Probabil

Probabilități

2015/2016



i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

- Cea mai utilizată statistică pentru măsurarea dependenței dintre două variabile este **coeficientul de corelație** sau **coeficientul de corelație Pearson**.
- Pentru două variabile aleatoare  $X$  și  $Y$  coeficientul de corelație este

$$\rho[X, Y] = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y},$$

unde  $\sigma_X^2$  și  $\sigma_Y^2$  sunt dispersiile celor două variabile.

- Pentru două eșantioane aleatoare  $x = \{x_1, x_2, \dots, x_n\}$  și  $y = \{y_1, y_2, \dots, y_n\}$  este

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\left[ \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right] \left[ \sum_{i=1}^n (y_i - \bar{y}_n)^2 \right]}},$$

unde  $\bar{x}_n$  și  $\bar{y}_n$  sunt mediile eșantioanelor.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

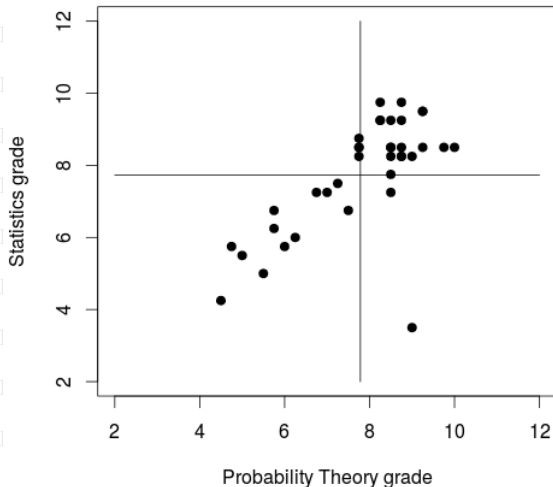
Probabilități

Probabil

Probabilități

Probabil

**Pozitive correlation  $r = 0.6964$**



i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

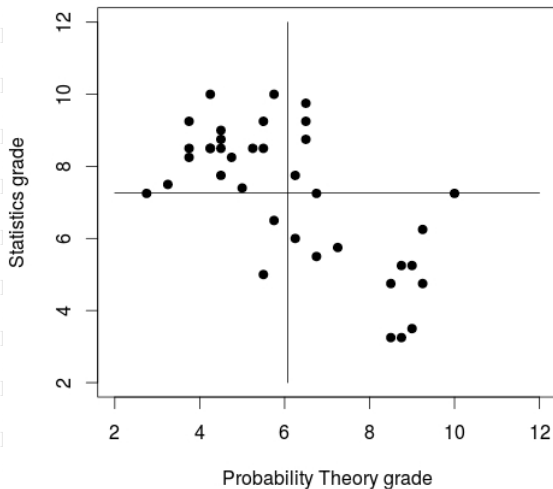
Probabilități

Probabil

Probabilități

Probabil

## Negative correlation -0.6708



i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

i și Statistică

tatistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

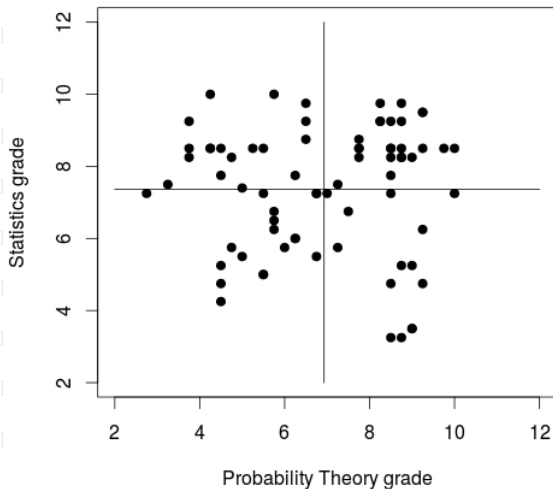
Probabilități

Probabil

Probabilități

Probabil

**Zero correlation 0.0059**



- Cum este folosit coeficientul de corelație ca măsură a asocierii?
- Cele două drepte trasate prin punctul mediilor împart diagrama în patru cadrane:
  - în cadranul din stânga jos amândouă variabilele sunt mai mici decât mediile lor:  $(x_i - \bar{x}_n)(y_i - \bar{y}_n) > 0$ ;
  - în cadranul din dreapta sus amândouă variabilele sunt mai mari decât mediile lor: produsul va fi de asemenea pozitiv;
  - în cadranul din dreapta jos variabila  $x$  este mai mare decât media și variabila  $y$  este mai mică decât media:  
 $(x_i - \bar{x}_n)(y_i - \bar{y}_n) < 0$ ;
  - în ultimul cadran variabila  $y$  este mai mare decât media și variabila  $x$  este mai mică decât media: produsul va fi de asemenea negativ.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

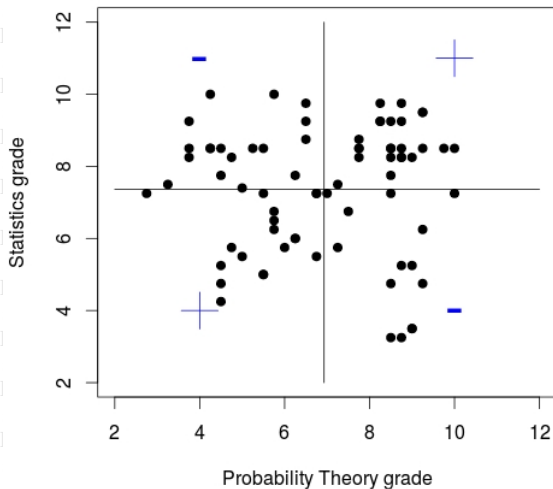
i și Statistică

tistică

i și Statistică

tistică

## Signs of the products





- Media tuturor acestor produse este coeficientul de corelație; dacă  $r$  este negativ, vor predomina punctele din cadranele negative; dacă  $r$  este pozitiv, vor predomina punctele din cadranele pozitive.
- Coeficientul de corelație nu este afectat de
  - interschimbarea variabilelor;
  - adăugarea unei aceleiași constante la valorile dintr-un eșantion;
  - înmulțirea cu o aceeași constantă pozitivă a valorilor dintr-un eșantion.
- Coeficientul de corelație ia valori cuprinse între  $-1$  și  $1$ ; valorile ale coeficientului de corelație apropiate de zero sugerează o foarte slabă asociere.

- Punctele dintr-o diagramă bidimensională se grupează în general în jurul *liniei deviației standard (linia SD)*.
- Linia SD trece prin punctul mediilor și prin toate punctele care se află la un număr egal de deviații standard față de medie.
- Altfel spus are o pantă egală în modul cu raportul deviațiilor standard ale celor două eșantioane:  $m = s_Y/s_X$  pentru corelație pozitivă și  $m = -s_Y/s_X$  pentru corelație negativă.
- Ecuația acestei drepte (linia SD) este

$$y - \bar{y}_n = m(x - \bar{x}_n),$$

unde  $\bar{x}_n$  și  $\bar{y}_n$  sunt mediile.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

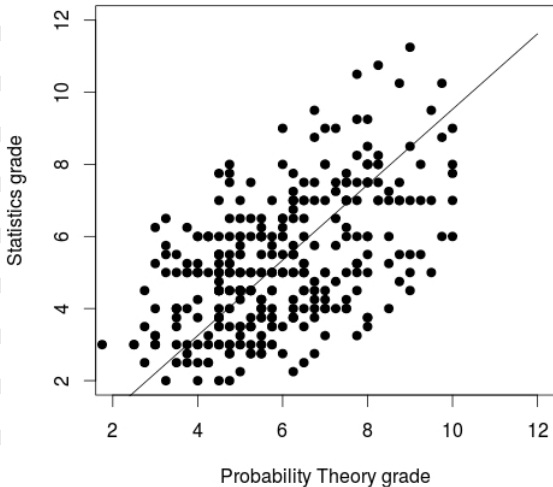
Probabilități

Probabil

Probabilități

Probabil

Probabilități

**2014/2015 SD line**

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

Probabil

Probabilități

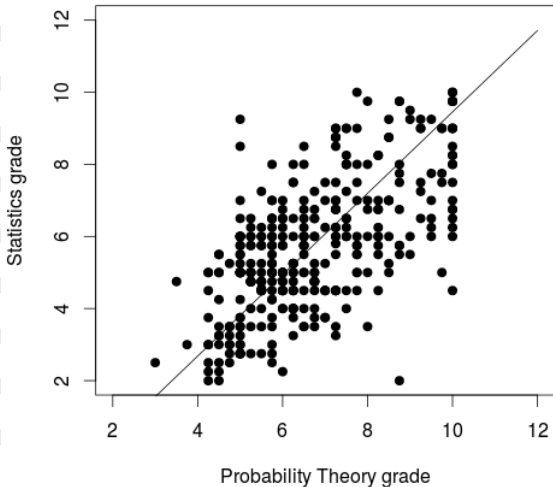
Probabil

Probabilități

Probabil

Probabilități

Probabil

**2015/2016 SD line**

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

i și Statistică

tistică

- Când diagrama este strânsă în jurul liniei SD există o *asociere liniară* puternică între variabile.
- Diagrama poate fi exprimată prin următoarele cinci statistici
  - media și deviația standard ale eșantionului valorilor  $x$ ;
  - media și deviația standard ale eșantionului valorilor  $y$ ;
  - coeficientul de corelație.
- *Asocierea pozitivă* este indicată de semnul pozitiv coeficientului de corelație sau de panta "norului" care urcă.
- *Asocierea negativă* este indicată de semnul negativ coeficientului de corelație sau de panta "norului" care coboară.
- Coeficientul de corelație ia valori între  $-1$  (când toate punctele se găsesc pe o dreaptă care coboară) și  $+1$  (când toate punctele se găsesc pe o dreaptă care urcă).

- Asocierea pozitivă (negativă) perfectă,  $r = +1$  ( $r = -1$ ), corespunde situației când între cele două variabile există o dependență liniară cu panta pozitivă (negativă) slope:

$$Y = mX + n,$$

$m > 0$  pentru asocierea pozitivă și  $m < 0$  pentru asocierea negativă

- Dacă  $|r|$  este aproape de 1, atunci un punct tipic al diagramei se găsește doar la o mică distanță de o deviație standard  $y$  (respectiv  $x$ ) deasupra sau dedesubtul (respectiv la stânga sau la dreapta) față de linia SD.
- Relația dintre coeficientul de corelație și distanța tipică față de linia SD se poate exprima matematic.
- Împrăștierea de-a lungul liniei SD line este cca.  $\sqrt{2(1 - |r|)} \cdot s_Y$  pe verticală, iar pe orizontală este  $\sqrt{2(1 - |r|)} \cdot s_X$ .

- Coeficientul de corelație este o statistică utilă pentru diagramele care au forma unei elipse, pentru alte forme ale diagramei corelația poate fi înșelătoare.
- Acest comportament poate fi cauzat de valori aberante, sau de alte asocieri de tip neliniar.
- **Coeficientul de corelație măsoară asocierea liniară, nu asocierea în general.**
- Revenim la exemplele noastre: coeficientul de corelație al anului 2014/2015 este 0.5418, iar cel al anului 2015/2016 este 0.6313.
- Există foarte puține valori de tip aberant, deci coeficientul de corelație este o măsură bună a unei prezumtive asocieri liniare.

- Împrăștierea în jurul liniei SD pentru anul 2014/2015 este 165% vertical și 173% orizontal. Împrăștierea în jurul liniei SD pentru anul 2015/2016 este 138% vertical, respectiv 156% orizontal.
- Evident există o corelație pozitivă în cei doi ani școlari, dar cu o împrăștiere mare în jurul liniei SD.
- Observăm o corelație pozitivă mai pronunțată în cel de-al doilea an și cu o împrăștiere mai mică în jurul liniei SD.
- Pantele SD sugerează un trend similar pentru punctajele la cele două examene.
- Putem considera că există o asocieră liniară între cele două tipuri de punctaje, deși nu foarte puternică.



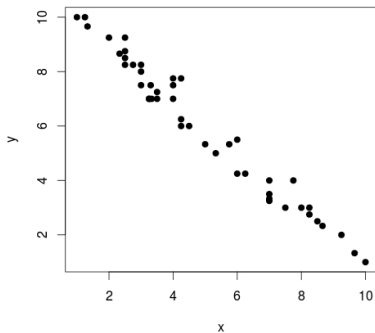
1. Se presupune că bărbații se căsătoresc cu femei care sunt cu exact 8% mai puțin înalte. Cum ar trebui să fie corelația dintre înălțimile lor?
2. Pentru un eșantion reprezentativ de autoturisme, cum ar trebui să fie corelația dintre vârsta mașinii și consumul de combustibil, pozitivă sau negativă?
3. Imaginile de mai jos conțin patru diagrame bidimensionale asociate unor date ipotetice. Coeficienții de corelație, într-o ordine schimbată, sunt

$$1 \quad -0.9833 \quad 0.9829 \quad -0.0760$$

Indicați coeficientul care corespunde fiecărei diagrame.

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

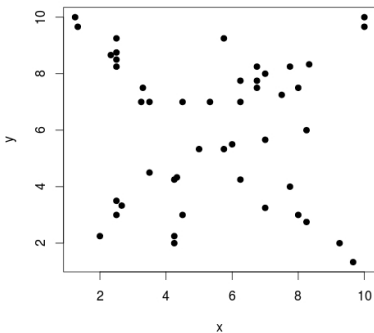
(a)



Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

(b)



Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

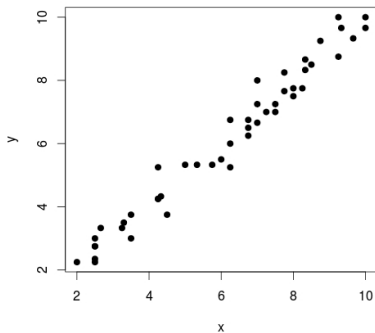
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

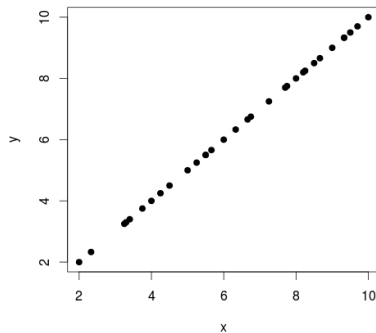
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

(c)



(d)

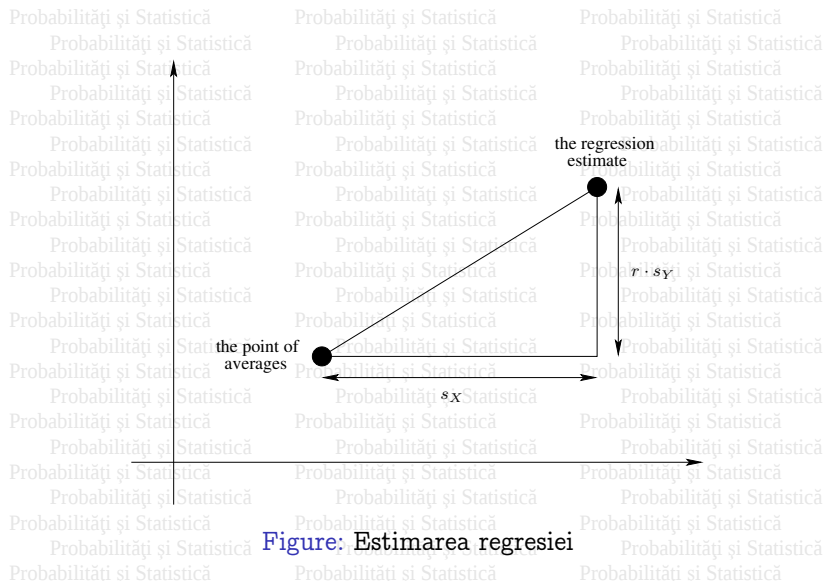


Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

Probabilități și Statistică  
Probabilități și Statistică  
Probabilități și Statistică

- Dacă metoda corelației are drept rol să detecteze asocierea liniară a două variabile, regresia încearcă să descrie cum una dintre variabile depinde de cealaltă.
- Există două *linii de regresie*: **linia de regresie a lui  $y$  față de  $x$**  care estimează valoarea medie a lui  $y$  corespunzătoare fiecărei valori a lui  $x$  și **linia de regresie a lui  $x$  față de  $y$** . Vom discuta doar despre primul tip regresie.
- Metoda regresiei poate fi descrisă astfel: *unei creșteri cu o deviație standard a lui  $x$  îi corespunde o creștere de  $r$  deviații standard a lui  $y$ , în medie.*
- Metoda regresiei liniare determină o dreaptă în care se potrivesc cel mai bine ("best fits") toate punctele diagramei trecând prin punctul mediilor.



- Formal, dacă această dreaptă este  $y = px + q$ , atunci  $p$  și  $q$  trebuie să minimizeze următoarea sumă a pătratelor distanțelor la toate punctele diagramei (cea mai bună potrivire corespunde metodei celor mai mici pătrate):

$$\sum_{i=1}^n (y_i - px_i - q)^2$$

- Soluția acestei probleme de minimizare este

$$p = rs_y/s_x, q = \bar{y}_n - \bar{x}_n rs_y/s_x.$$

- Cunoscând ecuația acestei drepte putem prezice valoarea uneia dintre variabile prin cealaltă.
- Linia de regresie nu trebuie să fie folosită dacă nu există o asocieră liniară între variabile: dacă există o asocieră neliniară linia de regresie va rata predicția corectă.

- Revenim la punctajele de la Probabilități/Statistică

- Pentru anul 2014/2015:

$$\bar{x}_n = 5.9116, \bar{y}_n = 5.2492, s_X = 1.7281, s_Y = 1.8082, r = 0.5418$$

Linia de regresie este  $Y = 0.5669X + 1.8978$ .

- Dacă alegem un student la întâmplare din acest an, de exemplu cu punctajul 5.25 la examenul de Probabilități, atunci putem prezice că nota lui la Statistică a fost 4.6004.

- Pentru anul 2015/2016:

$$\bar{x}_n = 6.6772, \bar{y}_n = 5.7029, s_X = 1.6163, s_Y = 1.8243, r = 0.6313$$

Linia de regresie este  $Y = 0.7125X + 0.9451$ .

- Dacă alegem un student la întâmplare din 2015/2016, cu punctajul 5.00 la examenul de Statistică, atunci putem prezice că punctajul lui la Probabilități a fost 5.6910.

1. O universitate face o analiză statistică a relației dintre scorul Math SAT (cu valori între 200 și 800) și scorul GPA (Grade Point Average, cu valori între 0 și 40, pentru studenții care termină primul an). Rezultatele sunt:

$$\text{scorul Math SAT mediu} = 550, s = 80$$

$$\text{scorul GPA mediu} = 2.6, s = 0.6, r = 0.4$$

Diagrama asociată are o forma de elipsă. Dacă un student este ales aleator și are scorul Math SAT 650, care a fost valoarea GPA?

2. Un profesor și-a standardizat examenele de la mijlocul semestrului și cel final astfel ca mediile lor să fie 50 cu deviația standard 10 (la ambele teste). Coeficientul de corelație dintre teste este 0.60. Știind că analiza corelației a evidențiat o asociere liniară, estimați scorul la cel de-al doilea test al unui student care a obținut sub 30 de puncte la primul test.



3. Într-un studiu relativ la stabilitatea IQ-ului, un grup mare de indivizi aleși aleator sunt testați mai întâi la 18 ani și apoi, din nou, la 35 de ani. S-au obținut următoarele rezultate

18 ani: IQ-ul mediu = 100,  $s = 15$

35 ani: IQ-ul mediu = 100,  $s = 15$ ,  $r = 0.80$

Estimați valoarea IQ-ului unui individ de 35 de ani care la vârsta de 18 ani avea un IQ egal cu 115. (Diagrama are o formă eliptică.)

4. Dintr-un studiu care a folosit 100 de familii:

înălțimea medie a soțului =  $68in$ ,  $s = 2.7in$

înălțimea medie a soției =  $63in$ ,  $s = 2.5in$ ,  $r = 0.25$

Preziceți înălțimea unei soții al cărei soț are (a) 72 in; (b) 64 in; (c) 68 in. (Diagrama are o formă eliptică.)

# Sfârșit

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică



Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.



Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică