Statistică - Cursul 13

Olariu E. Florentin

Mai, 2016

Cuprins

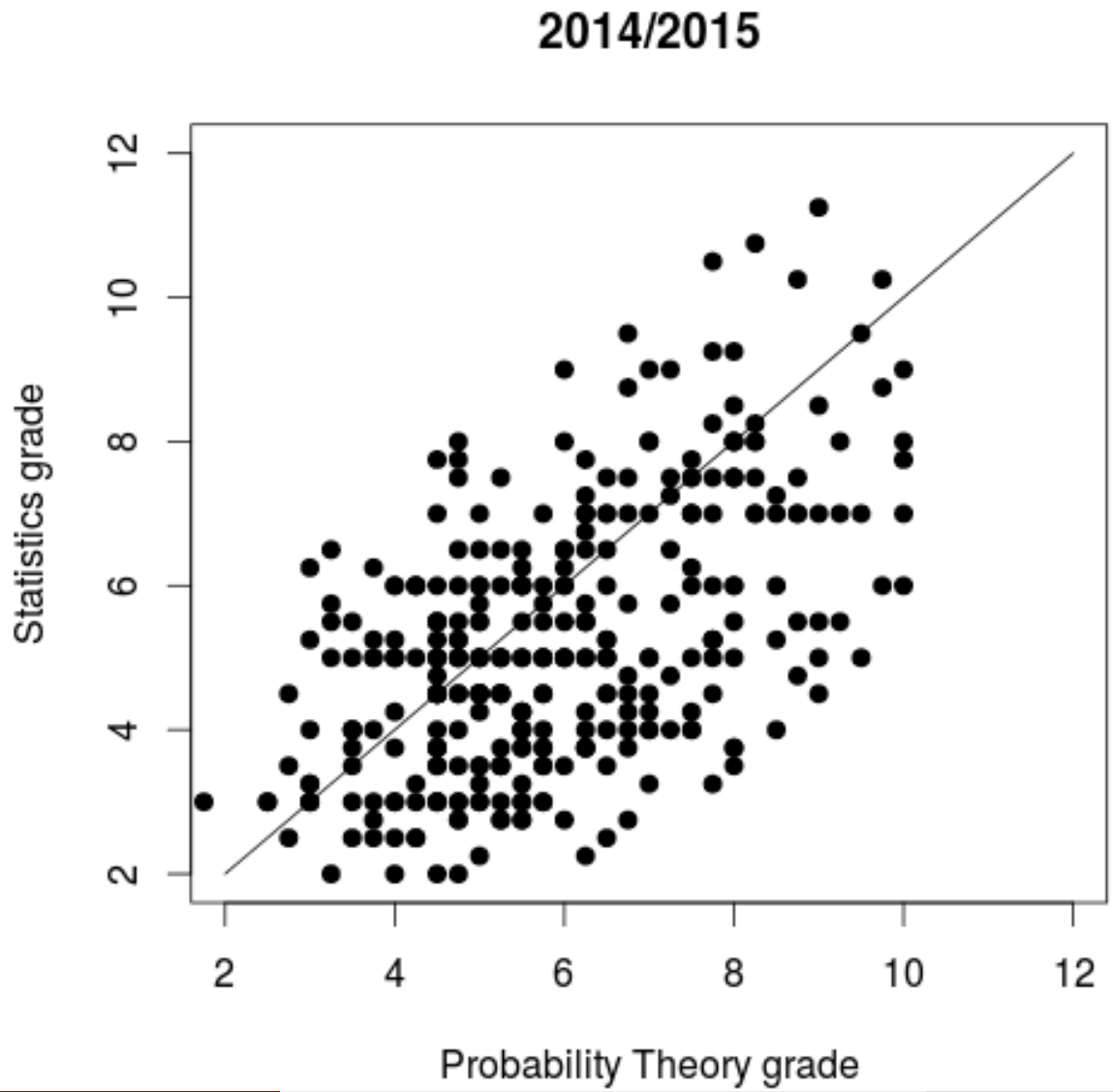
Corelaţia liniară

- *Corelaţia* este o metodă destinată studiului relaţiei dintre două variabile, fiind parte s *statisticii bivariate*.
- Primul statistician care a făcut un progres notabil în aceast domeniu a fost Francis Galton care a în cercat să studieze gradul în care copiii se aseamănă părinţilor lor.
- Tendinţa de a studia influenţele eredităţii prin intermediul instrumentelor statistice şi matematice a fost caracteristică epocii victoriene.
- Parte a unui studiu întreprins de Karl Pearson (un discipol al lui Galton), a fost măsurată înălţimea a 1078 taţi şi fii ajunşi la maturitate.
- Relaţia dintre două variabile (înălţimea tatălui şi a fiului) poate fi exprimată vizual într-o diagramă bidimensională.

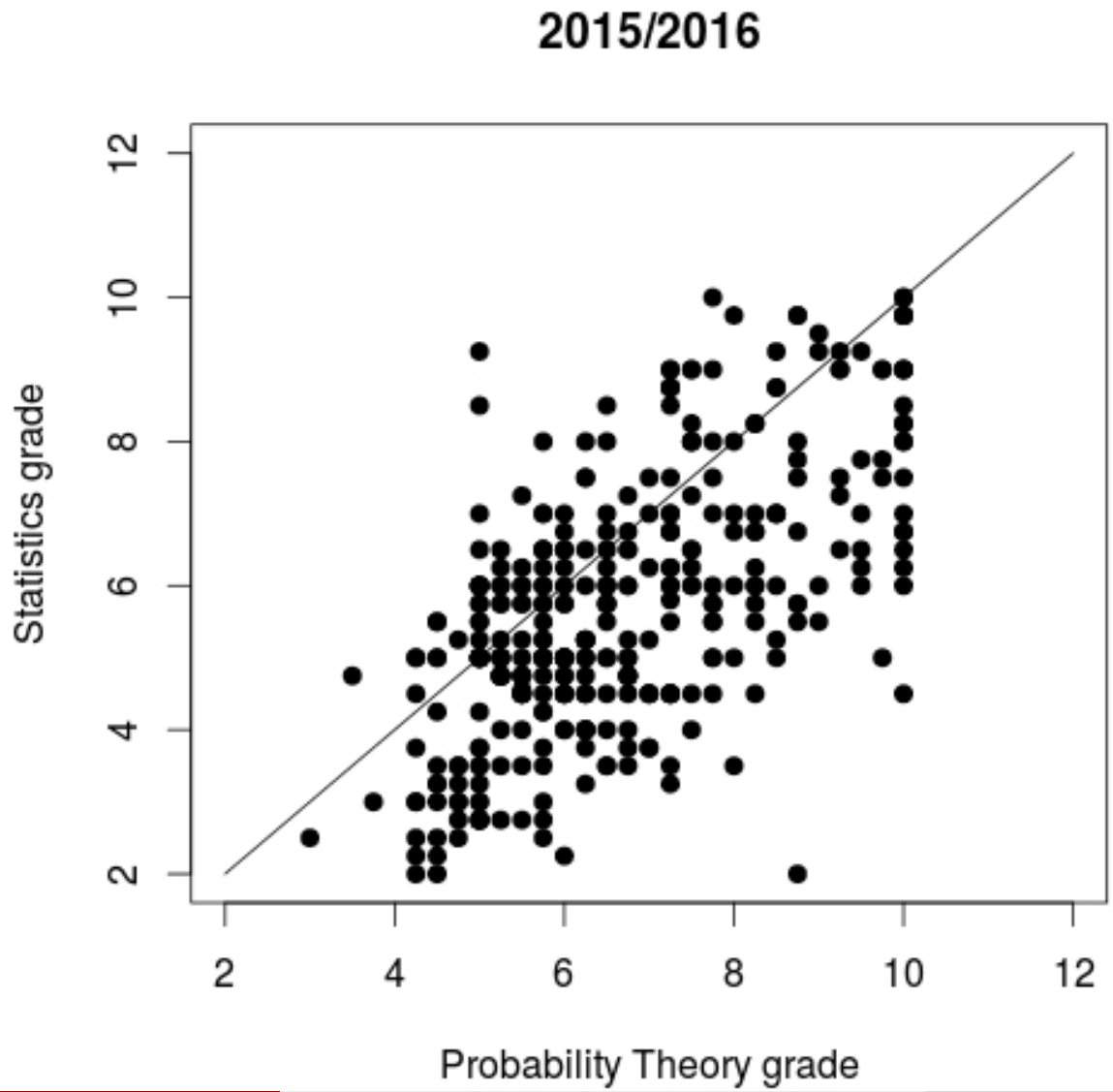
Corelaţia - un exemplu

- Noi vom întreprinde un studiu legat de relaţia dintre punctajele obţinute de studenţii FII la examenul de Probabilităţi (din săptămâna 8-a) şi cele de la examenul de Statistică (din săptămâna a 16-a).
- Punctajele a 335 de studenţi din anul şcolar 2014/2015 year şi a 355 de studenţi din 2015/2016 care au susţinut ambele examene sunt reprezentate în următoarele două diagrame.
- Fiecare punct reprezintă o pereche de punctaje: pe coordonata x cel de la Probabilităţi, iar pe y punctajul de la Statistică.
- Am reprezentat de asemeni prima bisectoare; această dreaptă corespunde studenţilor care au avut acelaşi punctaj la ambele examene.
- Dacă punctajul unui student la Probabilităţi este apropiat de cel de la Statistică, atunci punctul corespunzător este situat aproape de prima bisectoare.

Corelaţia - un exemplu



Corelaţia - un exemplu



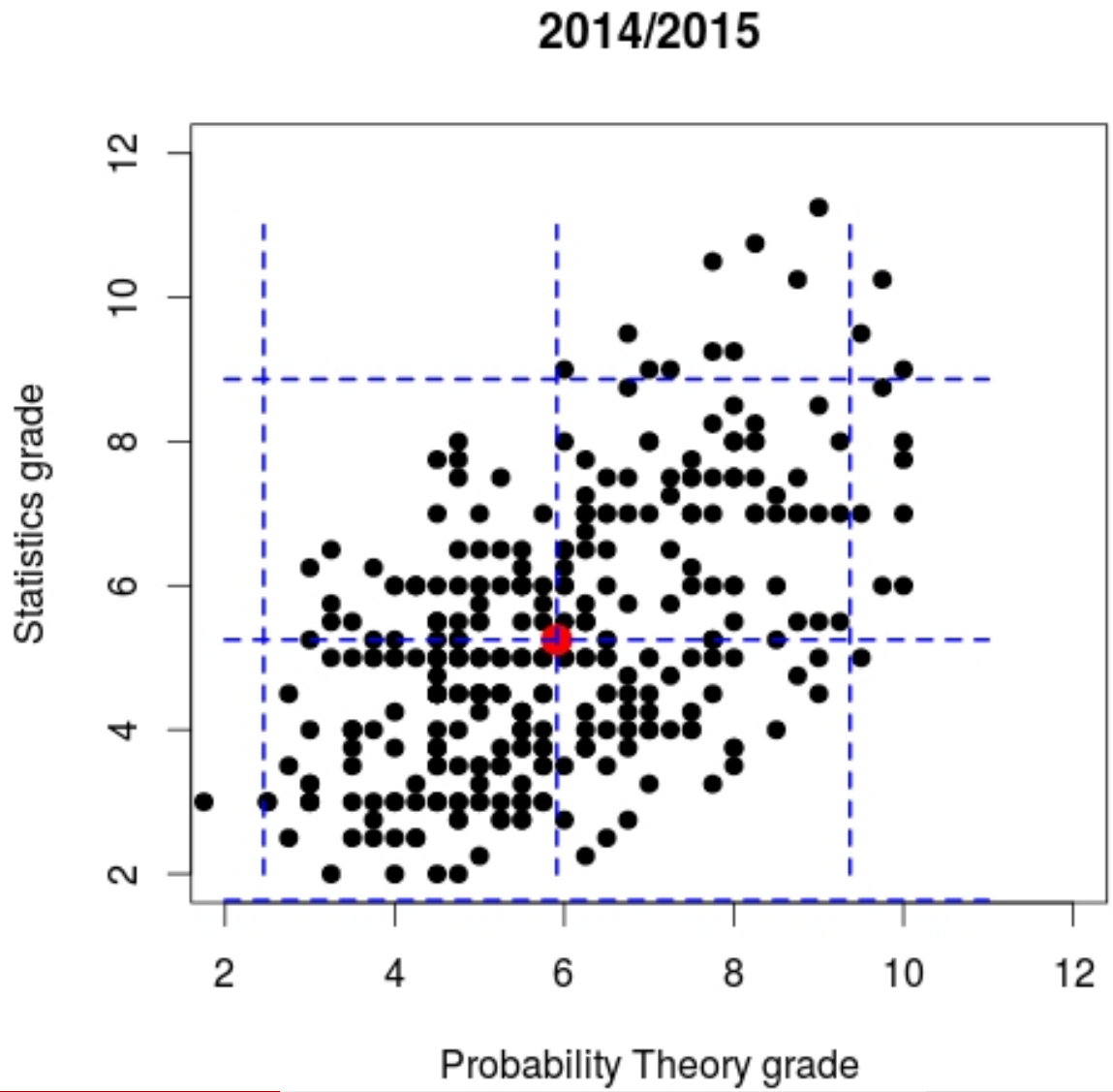
Corelaţia - un exemplu

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

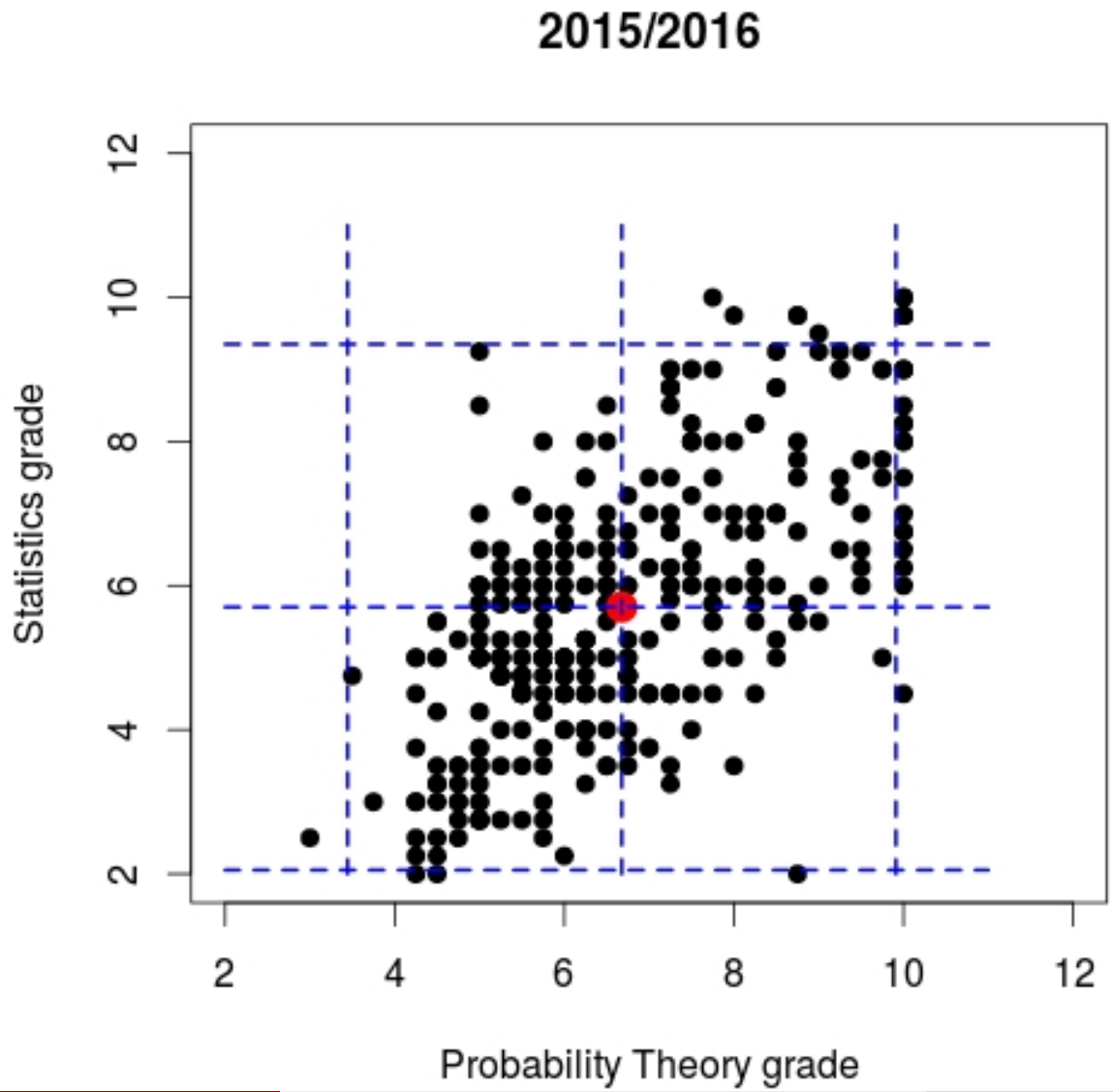
Corelaţia - un exemplu

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

Corelaţia - un exemplu



Corelaţia - un exemplu



Coeficientul de corelaţie

- Cea mai obişnuită statistică pentru măsurarea dependenţei dintre două variabile este **coeficientul de corelaţie** sau **coeficientul de corelaţie Pearson**.
- Pentru două variabile aleatoare X şi Y coeficientul de corelaţie este

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

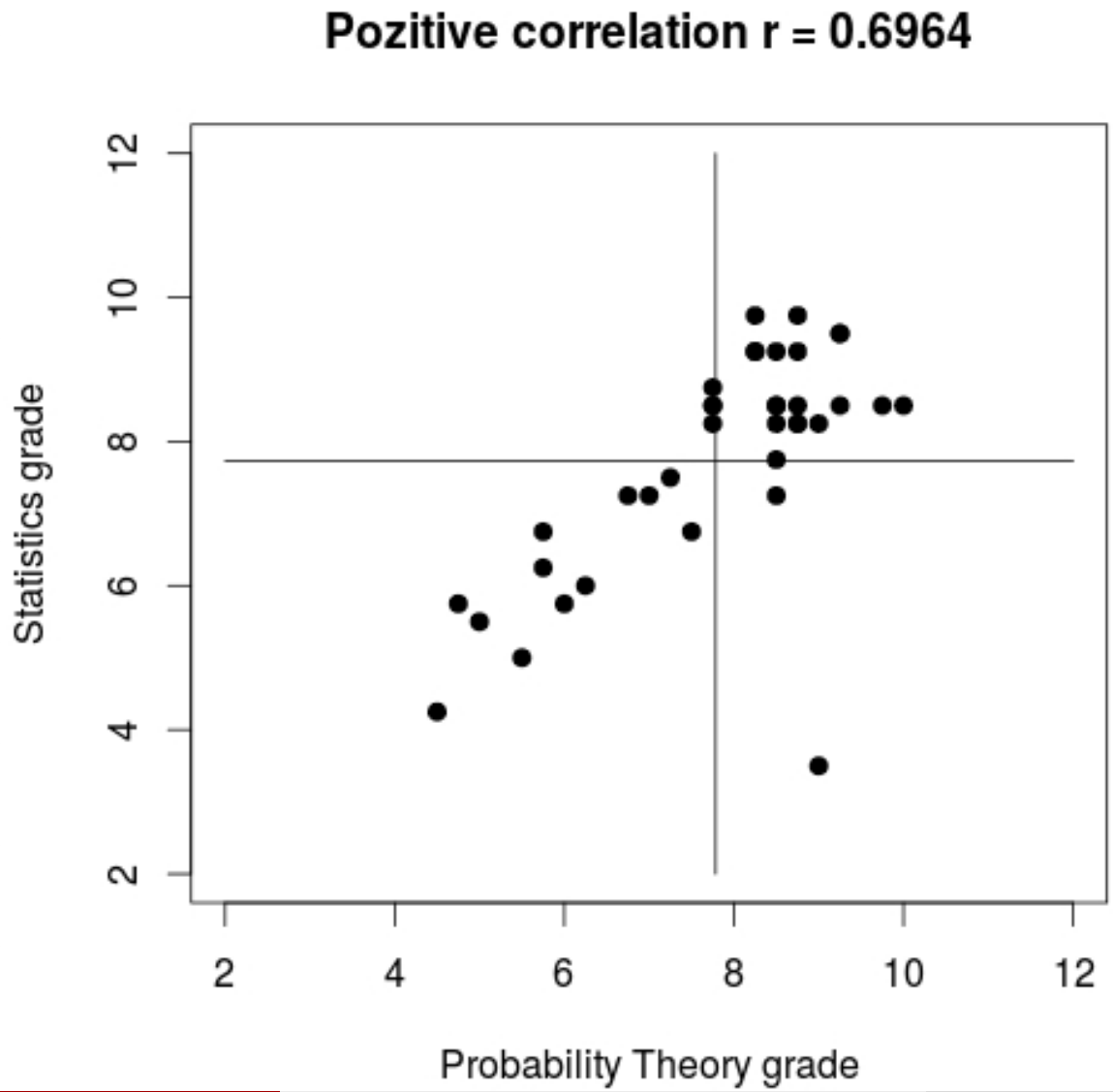
unde σ_X^2 şi σ_Y^2 sunt dispersiile celor două variabile.

- Pentru două eşantioane aleatoare $x = \{x_1, x_2, \ldots x_n\}$ şi $y = \{y_1, y_2, \ldots$ este

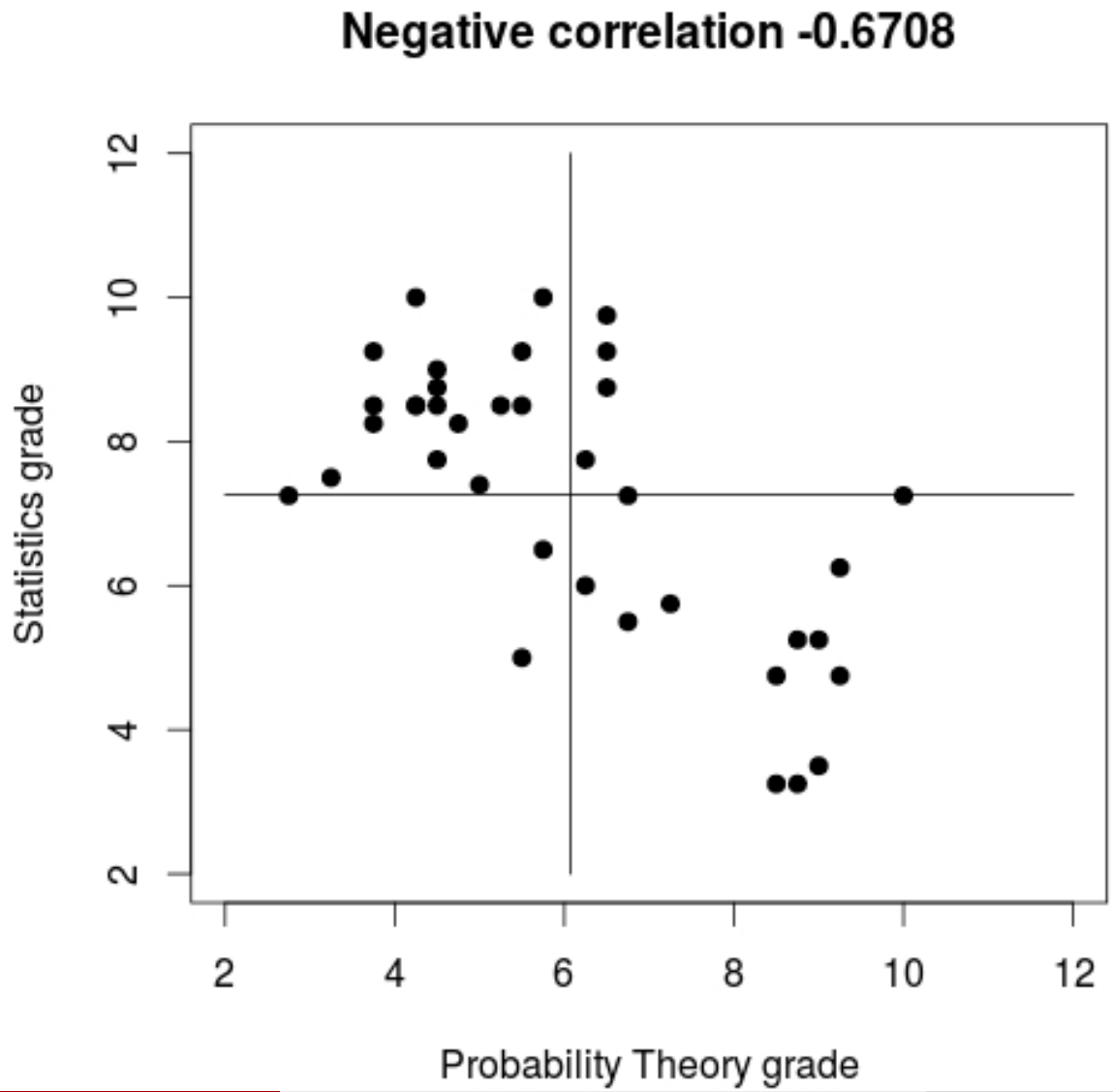
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \sum_{i=1}^{n}(y_i - \overline{y}_n)^2}},$$

unde $\overline{x}_n$ şi $\overline{y}_n$ sunt mediile eşantioanelor.

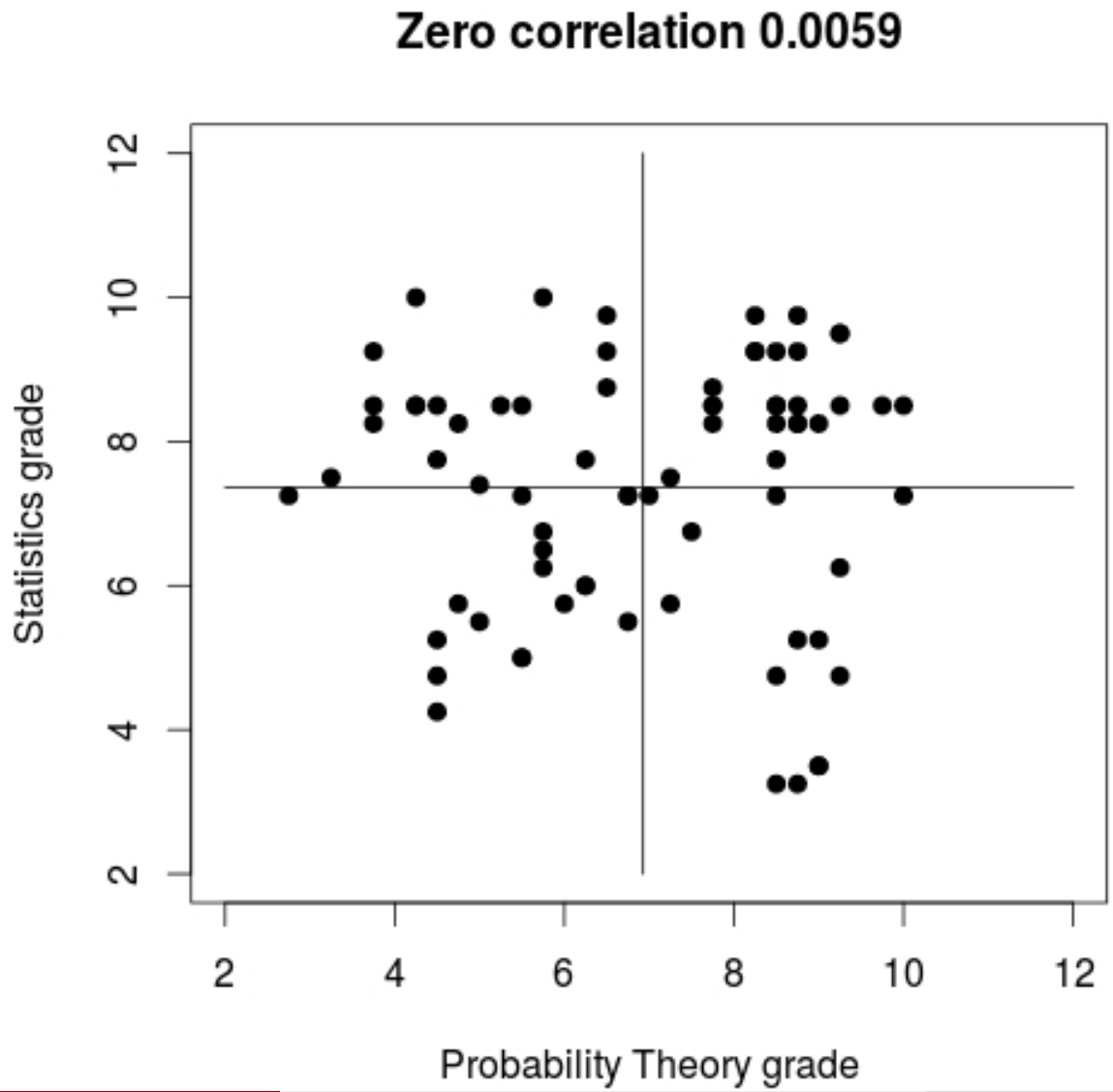
Coeficientul de corelaţie



Coeficientul de corelaţie



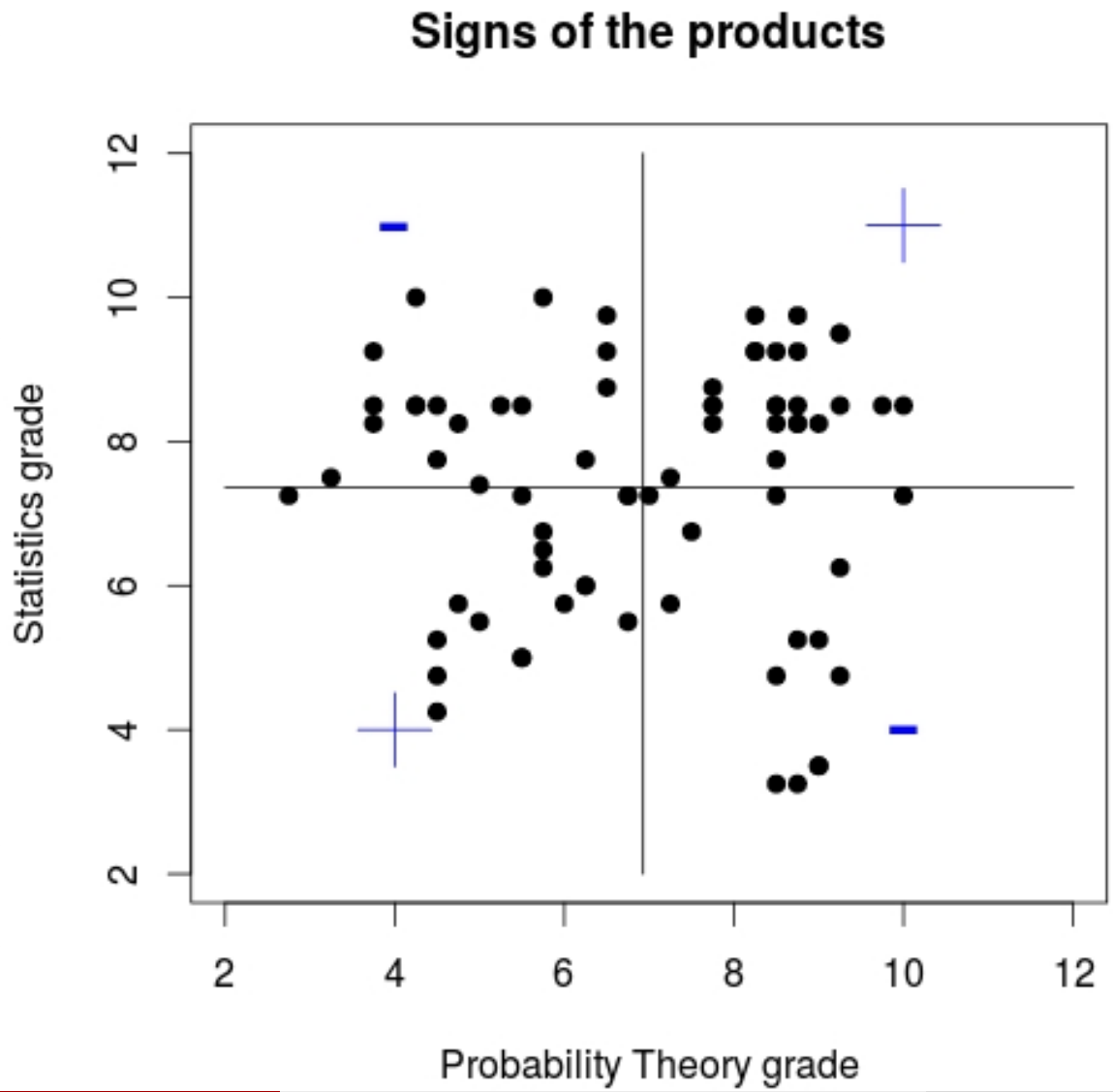
Coeficientul de corelaţie



Coeficientul de corelaţie

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

Coeficientul de corelaţie



Coeficientul de corelaţie

- Media tuturor acestor produse este coeficientul de corelaţie; dacă r este negativ, vor predomina punctele din cadranele negative; dacă r este pozitiv, vor predomina punctele din cadranele pozitive.
- Coeficientul de corelaţie nu este afectat de
  - interschimbarea variabilelor;
  - adăugarea unei aceleiaşi constante la valorile dintr-un eşantion;
  - înmulţirea cu o aceeaşi constantă pozitivăa valorilor dintr-un eşantion.
- Coeficientul de corelaţie ia valori cuprinse între -1 şi 1; valorile ale coeficientului de corelaţie apropiate de zero sugerează o foarte slabă asociere.

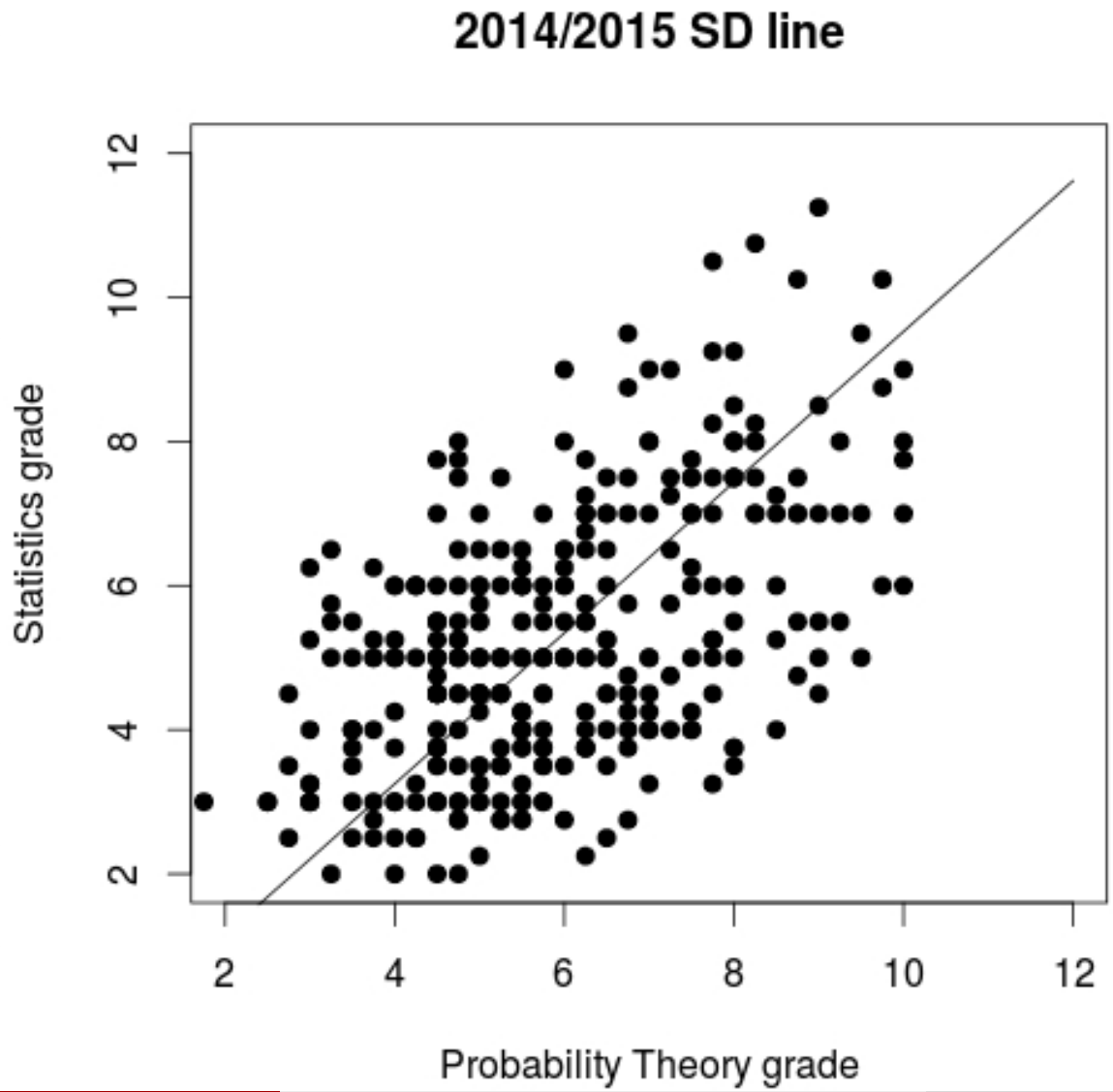
Linia SD

- Punctele dintr-o diagramă bidimensională se grupează în general în jurul *liniei deviaţiei standard* (*linia SD*).
- Linia SD trece prin punctul mediilor şi prin toate punctele acre se află la un număr egal de deviaţii standard faţă de medie.
- Altfel spus are o pantă egală în modul cu raportul deviaţiilor standard ale celor două eşantioane: $m = s_Y/s_X$ pentru corelaţie pozitivă şi $m = -s_Y/s_X$ pentru corelaţie negativă.
- Ecuaţia acestei drepte (linia SD) este

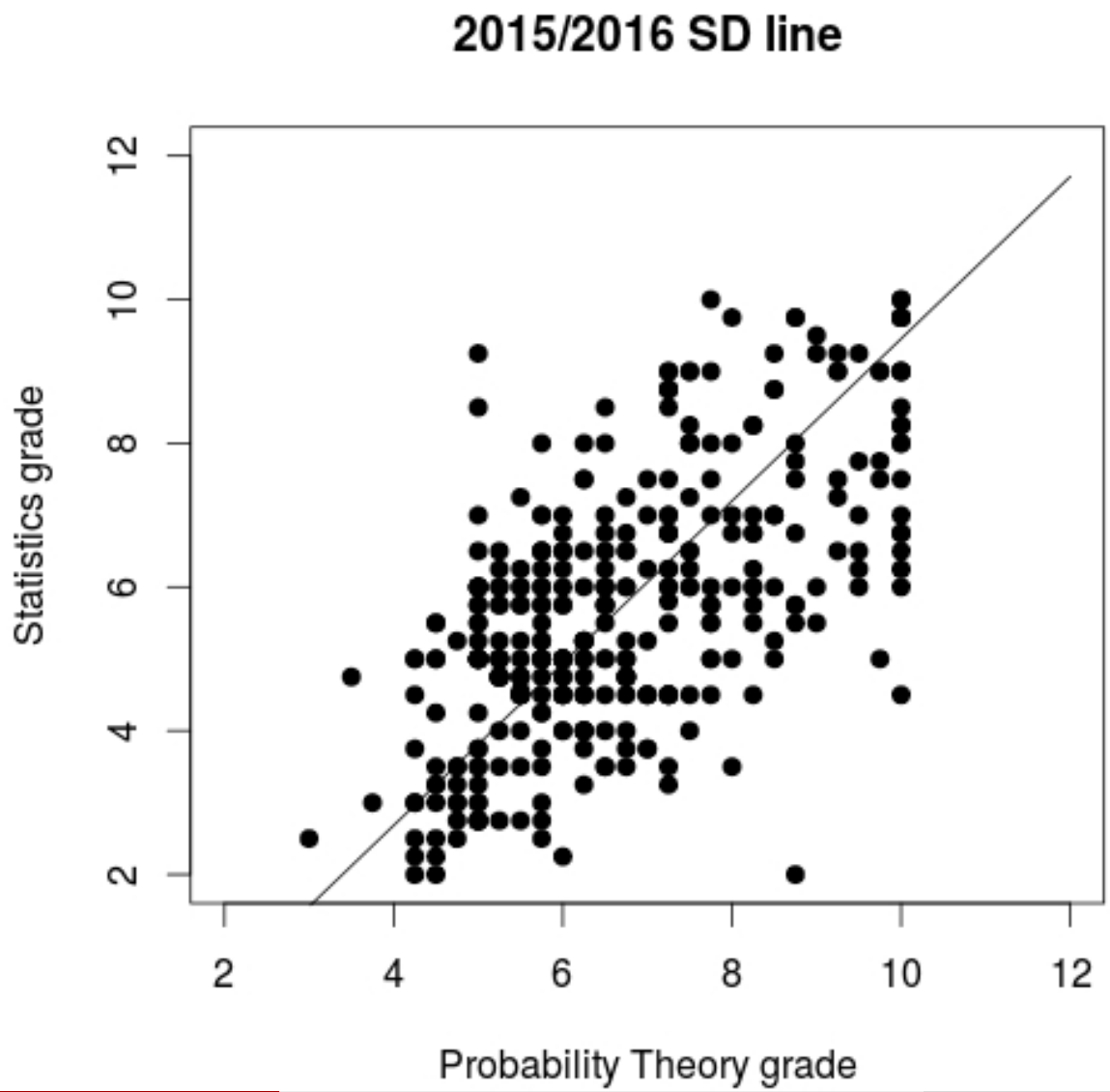
$$y - \overline{y}_n = m(x - \overline{x}_n),$$

unde $\overline{x}_n$ şi $\overline{y}_n$ sunt mediile.

Linia SD


2014/2015 SD line
Statistics grade
Probability Theory grade
2
4
6
8
10
12

Linia SD


2015/2016 SD line
Statistics grade
Probability Theory grade
2
4
6
8
10
12

Sumar

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

Sumar

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

$$Y = mX + n,$$

$m > 0$ pentru asocierea pozitivă şi $m < 0$ pentru asocierea negativă.

- Dacă $|r|$ este aproape de 1, atunci un punct tipic al diagramei se găseşte doar la o mică distanţă de o deviaţie standard y (respectiv x) deasupra sau dedesubtul (respectiv la stânga sau la dreapta) faţă de linia SD.
- Relaţia dintre coeficientul de corelaţie şi distanţa tipică faţă de linia SD se poate exprima matematic.
- Împrăştierea de-a lungul liniei SD line este aproximativ $\sqrt{2(1-|r)} \cdot s_Y$ pe verticală, iar pe orizontală este $\sqrt{2(1-|r)} \cdot s_X$.

Sumar

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

Sumar

- Împrăştierea în jurul liniei SD pentru anul 2014/2015 este 165% vertical şi 173% orizontal. Împrăştierea în jurul liniei SD pentru anul 2015/2016 este 138% vertical, respectib 156% orizontal.
- Evident există o corelaţie pozitivă în cei doi ani şcolari, dar cu o împrăştiere mare în jurul liniei SD.
- Observăm o corelaţie pozitivă mai pronunţată în cel de-al doilea an şi cu o împrăştiere mai mică în jurul liniei SD.
- Pantele SD sugerează un trend similar pentru punctajele la cele douăexamene.
- Putem considera că există o asociere liniară între cele două tipuri de punctaje, deşi nu foarte puternică.

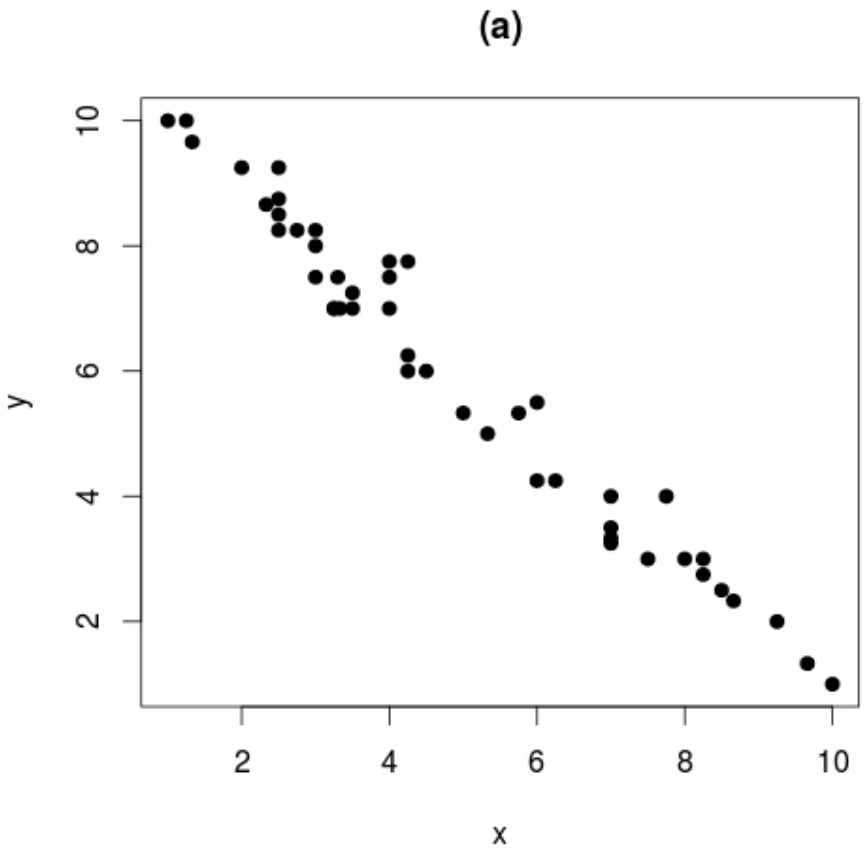
Corelaţie - Exerciţii

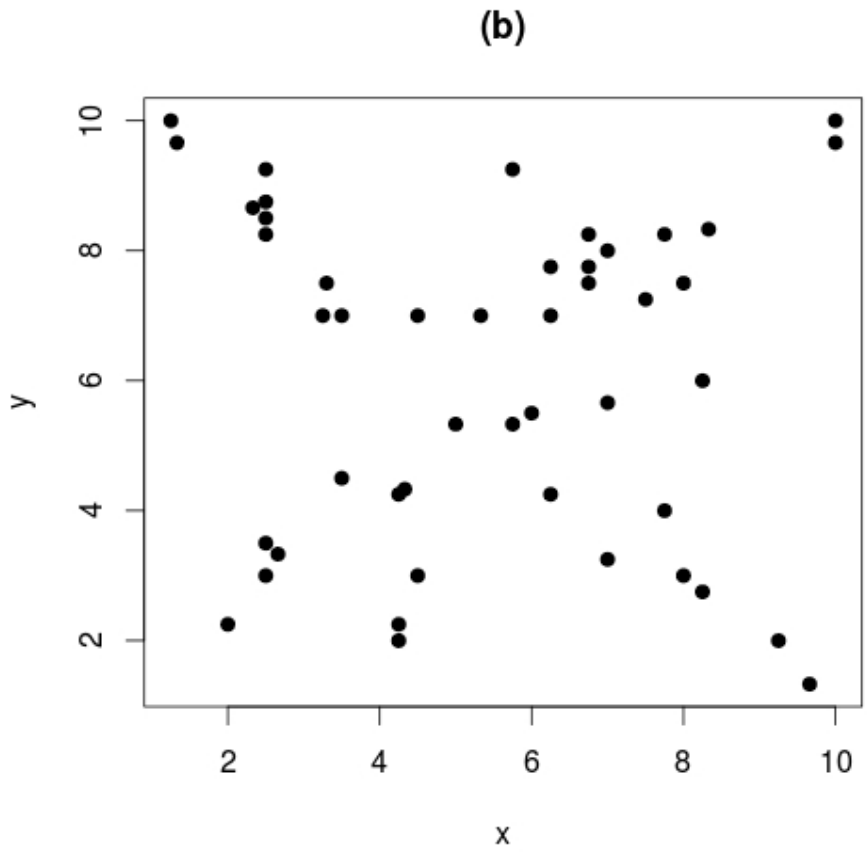
1. Se presupune că bărbaţii se căsătoresc cu femei care sunt cu exact 8% mai puţin înalte. Cum ar trebui să fie corelaţia dintre înălţimile lor?
2. Pentru un eşantion representativ de autoturisme, cum ar trebui să fie corelaţia dintre vârsta maşinii şi consumul de combustibil, pozitivă sau negativă?
3. Imaginile de mai jos conţin patru diagrame bidimensionale asociate unor date ipotetice. Coeficienţii de corelaţie, într-o ordine schimbată, sunt

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

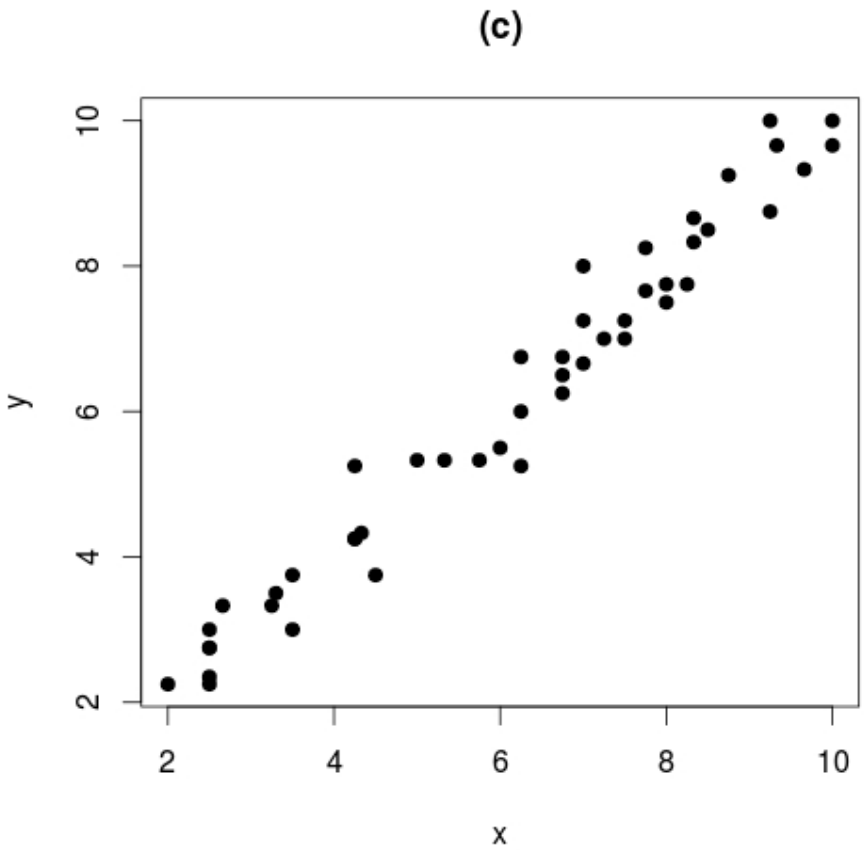
Indicaţi coeficientul care corespunde fiecărei diagrame.

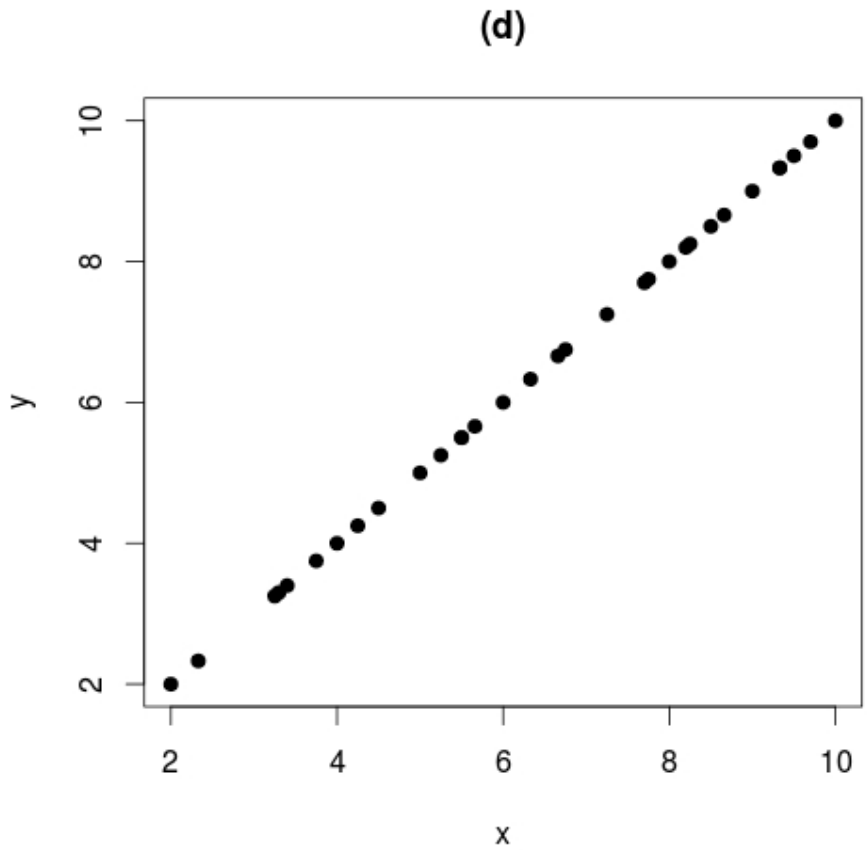
Corelaţie - Exerciţii





Corelaţie - Exerciţii





Regresia

- Dacă metoda corelaţiei 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 faţă de x** care estimează valoarea medie a lui y corespunzătoare fiecărei valori a lui x şi **inia de regresie a lui x faţă de y**. Vom discuta doar despre primul tip regresie.
- Metoda regresiei poate fi descrisă astfel: *unei creşteri cu o deviaţie standard a lui x îi corespunde o creştere de r deviaţii standard deviations 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.

Linia de regresie

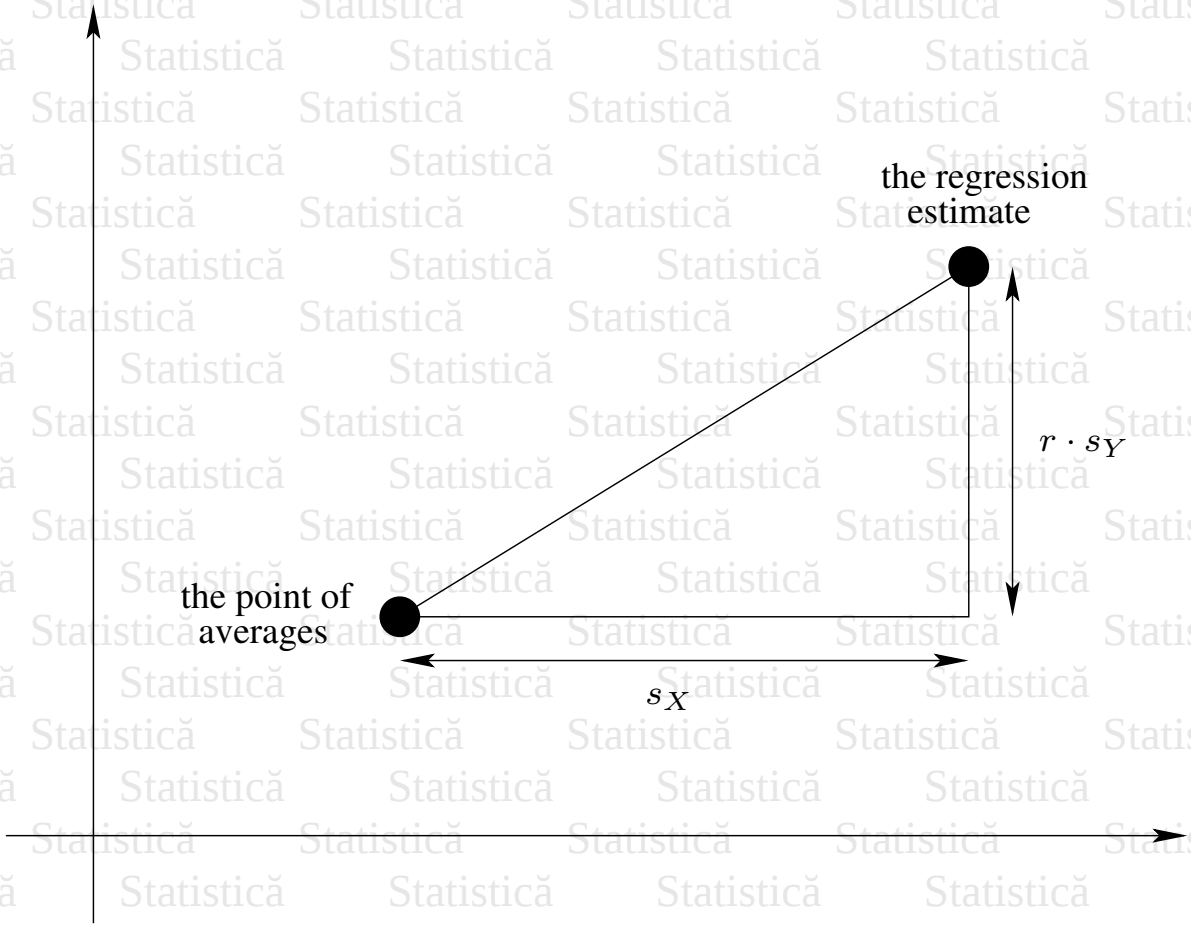


Figure : Estimarea regresiei

Linia de regresie

- Formal, dacă această dreaptă este $y = mx+n$, atunci m şi n trebuie să minimizeze următoarea sumă a pătratelor distanţelor la toate punctele diagramei (cea mai bunăpotrivire corespunde metodei celor mai mici pătrate):

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

- Soluţia acestei probleme de minimizare este

$$m = rs_y, n = \overline{y}_n - rs_y\overline{x}_n.$$

- Cunoscând ecuaţia acestei drepte putem prezice valoarea uneia dintre variabile prin cealaltă.
- Linia de regresie nu trebuie sa fie folosită dacă nu există o asociere liniară între variabile: dacă există o asociere ne-liniară linia de regresie va rata predicţia corectă.

Linia de regresie - exemple

- Revenim la punctajele de la Probabilităţi/Statistică
- Pentru anul 2014/2015:

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

$$\text{Linia de regresie este } Y = 0.9797X - 0.5430.$$

- Dacă alegem un student la întâmplare din acest an, de exemplu cu punctajul 5.25 la examenul de Probabilităţi, atunci putem prezice că nota lui la Statistică a fost 4.6004.
- Pentru anul 2015/2016:

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

$$\text{Linia de regresie este } Y = 1.1517X - 1.9874.$$

- Dacă alegem un student la întâmplare din 2015/2016, cu punctajul 4.00 la examenul de Statistică, atunci putem prezice că punctajul lui la Probabilităţi a fost 5.1987.

Regresie - Exerciţii

1. O universitate face o analiză statistică a relaţiei dintre scorul Math SAT (cu valori în tre 200 şi 800) şi scorul GPA (Grade Point Average, cu valori între 0 şi 40, pentru studenţii 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 şi are scorul Math SAT 650, care a fost valoarea GPA?

2. Un profesor şi-a standardizt examenele de la mijlocul semestrului şi cel final astfel ca mediile lor să fie 50 cu deviaţia standard 10 (la ambele teste). Coeficientul de corelaţie dintre teste este 0.60. Ştiind că analiza corelaţiei a evidenţiat o asociere liniară, estimaţi scorul la cel de-al doilea test al unui student care a obţinut sub 30 de puncte la primul test.

Regresie - Exerciţii

3. Într-un studiu relativ la stabilitatea IQ-ului, un grup mare de indivizi aleşi aleator sunt testaţi mai întâi la 18 ani şi apoi, din nou, la 35 de ani. S-au obţinut următoarele rezultate

$$\text{18 ani: IQ-ul mediu} = 100,\ s = 15$$

$$\text{35 ani: IQ-ul mediu} = 100,\ s = 15,3 = 0.80$$

Estimaţi 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:

$$\text{înălţimea medie a soţului} = 68in, s = 2.7in$$

$$\text{înălţimea medie a soţiei} = 63in, s = 2.5in, r = 0.25$$

Preziceţi înălţimea unei soţii al cărei soţ are (a) 72 in; (b) 64 in; (c) 68 in. (Diagrama are o formă eliptică.)

Sfârşit

Bibliography

- 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.