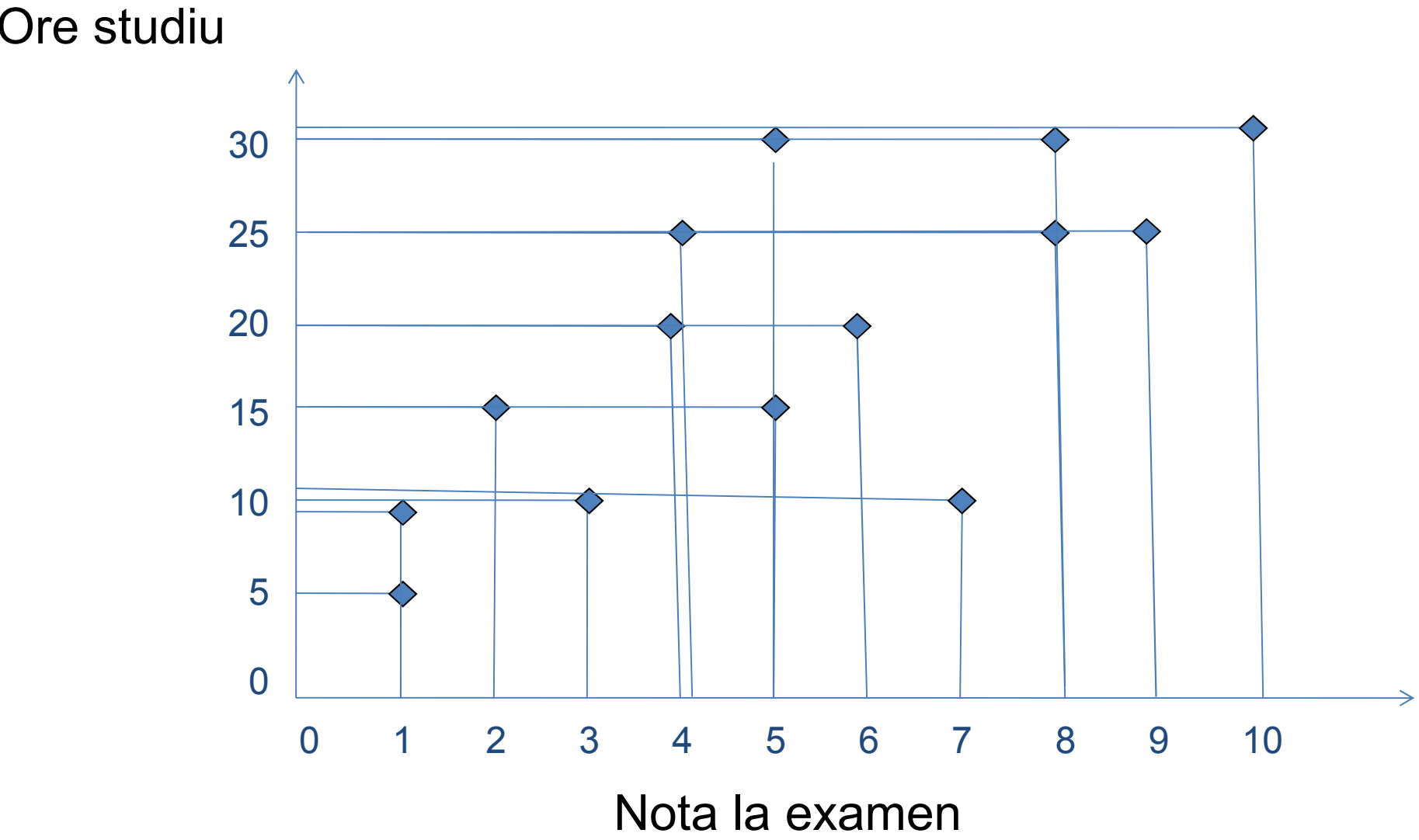
Coeficientul de corelaţie Pearson (r)

M. Popa

Asocierea valorilor perechi



Conceptul de corelaţie (Galton şi Pearson)

- cauzalitatea este doar limita extremă categoriei de relaţie între două fenomene
  - prea complexă pentru a fi întotdeauna demonstrată
- ”asocierea” poate fi un principiu explicativ
  - aduce în domeniul ştiinţelor sociale şi umane rigoarea specifică ştiinţelor fizice şi naturale.

probleme de cercetare tipice…

- *„există o legătură între numărul atitudini pozitive pe care le manifestă oamenii şi numărul atitudinilor pozitive pe care le primesc din partea celor din jur?”*.
- *„există o legătură între timpul de reacţie şi nivelul extraversiunii, ca trăsătură de personalitate?”*.
- *„există o legătură între greutate şi înălţime?*
- *„există o relaţie între frecvenţa pulsului şoferilor şi viteza maşinii pe care o conduc?”*
- *„există o relaţie între numărul orelor de studiu la statistică şi punctajul obţinut la evaluări?”*

Coeficientul de covarianţă

$$\mathrm{cov}_{xy} = \frac{\sum x * y}{N}$$

- precursorul coeficientului de corelaţie
- ridică probleme de utilizare în cazul variabilelor exprimate în unităţi de măsură diferite

Coeficientul de corelaţie Pearson

- z_x şi z_y sunt transformările z ale variabilelor corelate
- formula poate fi utilizată indiferent de unitatea de măsură
- "r" poate lua valori între
  - -1, corelaţie perfectă negativă
  - +1, corelaţie perfectă pozitivă
  - 0, absenţa corelaţiei

$$r = \frac{\sum z_x * z_y}{N}$$

$$r = \frac{\sum z_x^{\,2}}{N}$$

Formula de calcul

$$r = \frac{\sum (X - m_x) * (Y - m_y)}{N * s_x * s_y}$$

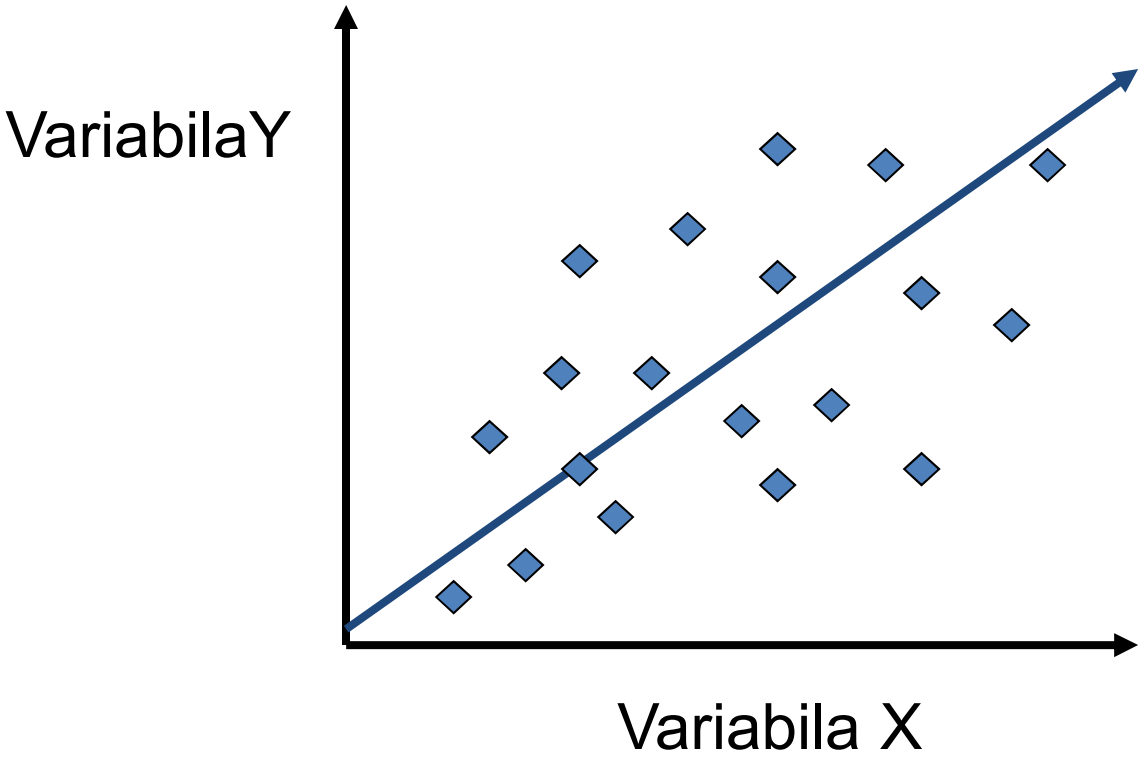
Plaja de valori Pearson *r*

- un număr între -1 şi +1 care indică intensitatea relaţiei dintre variabile
  - Semnul (- sau +) indică direcţia relaţiei
  - Valoarea indică intensitatea relaţiei

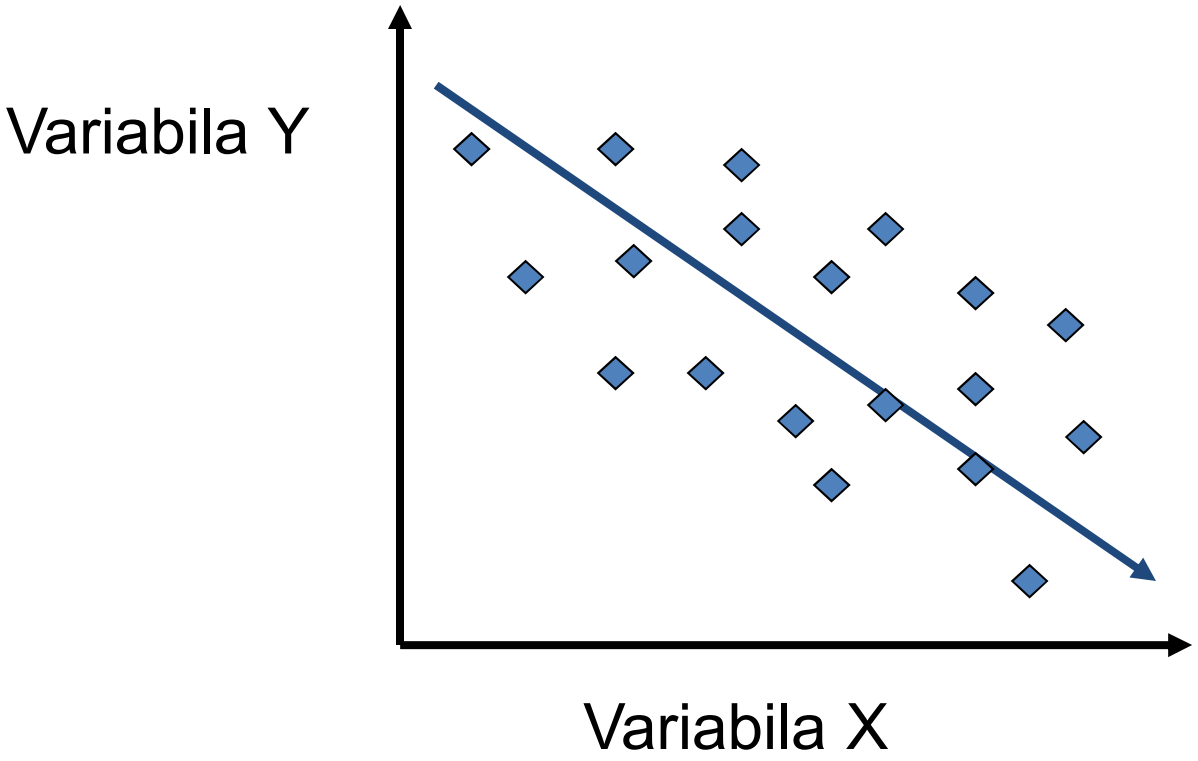
-1 ---------- 0 ---------- +1

corelaţie perfectă negativă	corelaţie nulă	corelaţie perfectă pozitivă

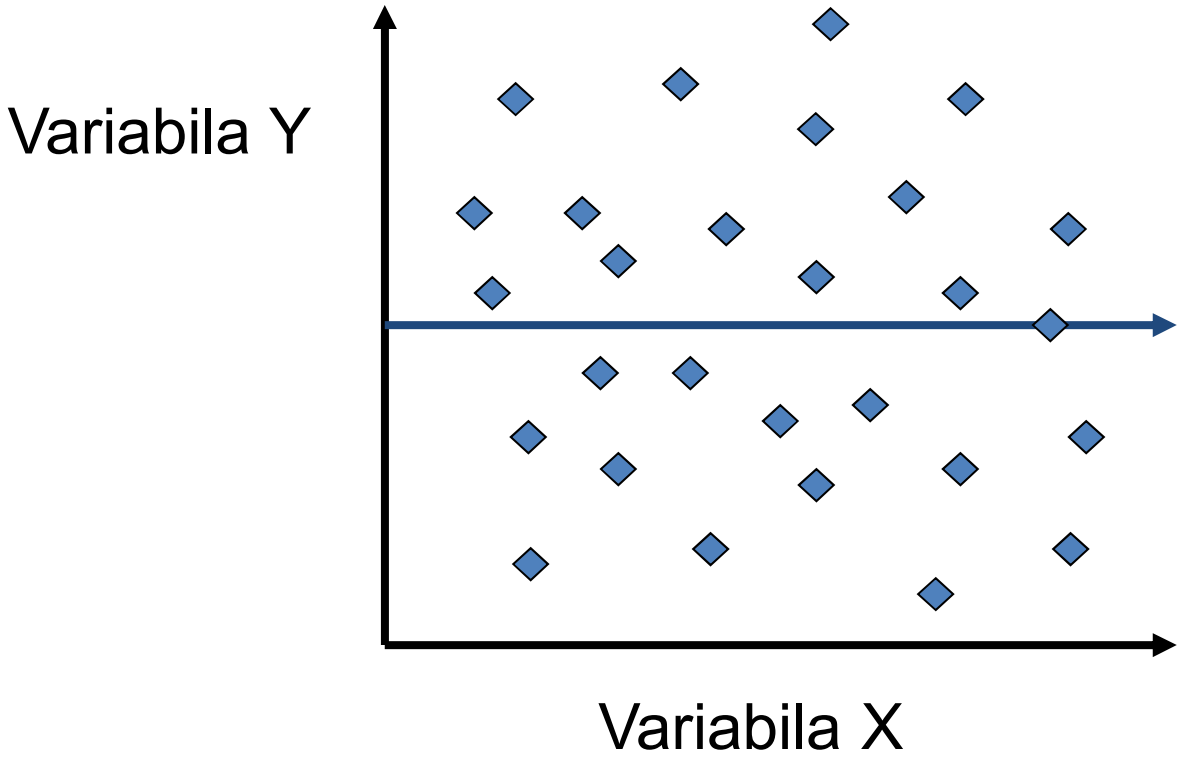
corelaţie pozitivă



corelaţie negativă

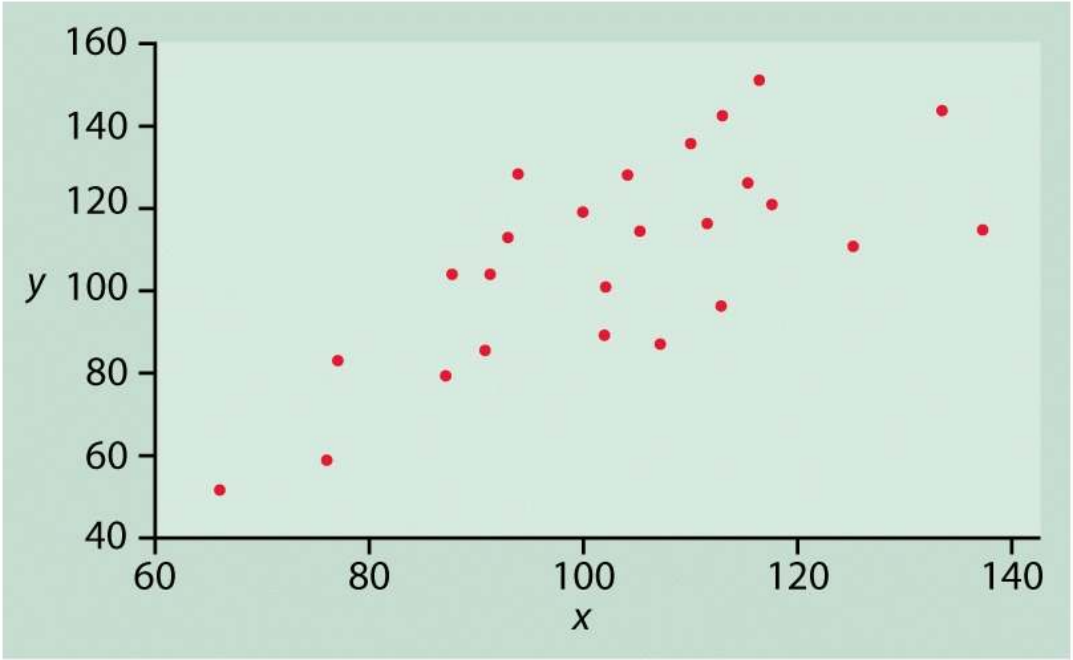


corelaţie inexistentă (0)

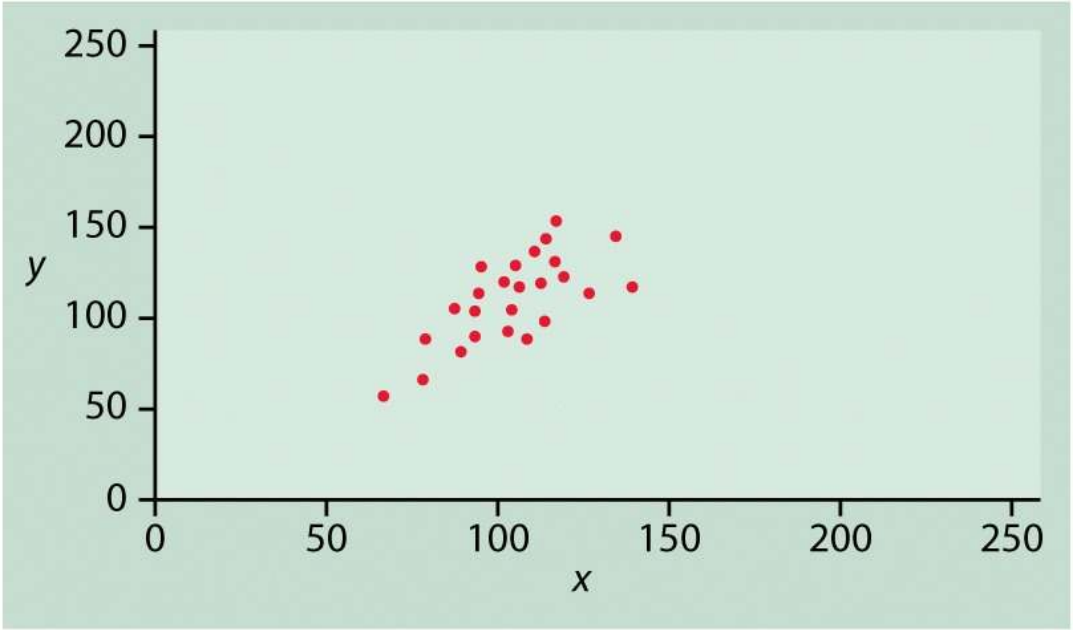


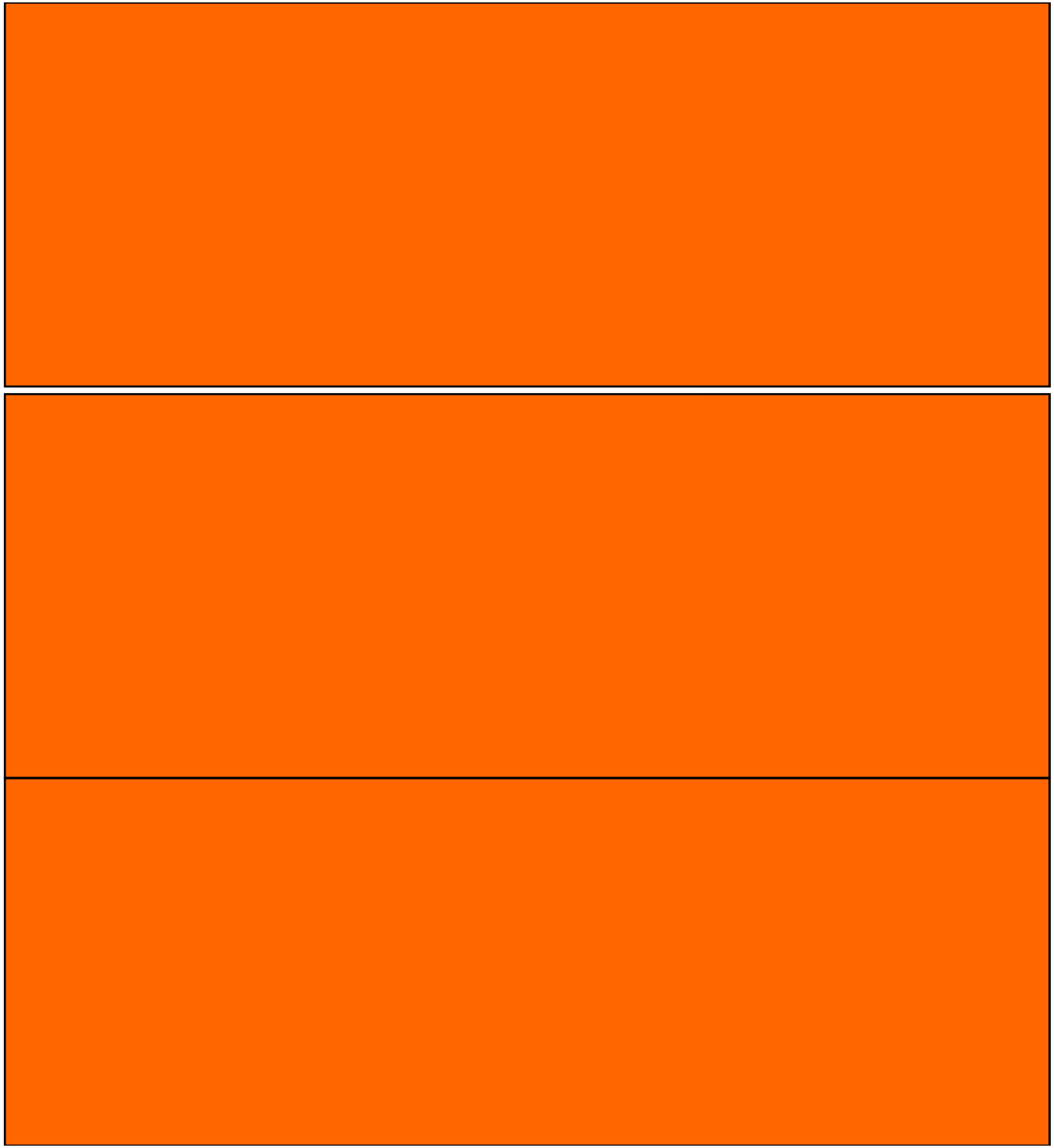
Scatterplot 1

care indică o corelaţie mai puternică?



Scatterplot 2





Un exemplu

- Cercetătorii au observat o relaţie între timpul de reacţie şi numărul erorilor la diverse tipuri de sarcini.
- Această relaţie este denumită “compensarea viteză-corectitudine”.
- Datele reprezintă timpul de reacţie (milisecunde) şi numărul total de erori înregistrate pentru un număr de 8 subiecţi.

tr	erori
184	10
213	6
234	2
197	7
189	13
221	10
237	4
192	9

Criteriile deciziei statistice

- coeficientul ***r*** se raportează la o distribuţie teoretică derivată din distribuţia ***t***
- df=N-2
- tabel special cu praguri de semnificaţie ale coeficientului de corelaţie ***r***

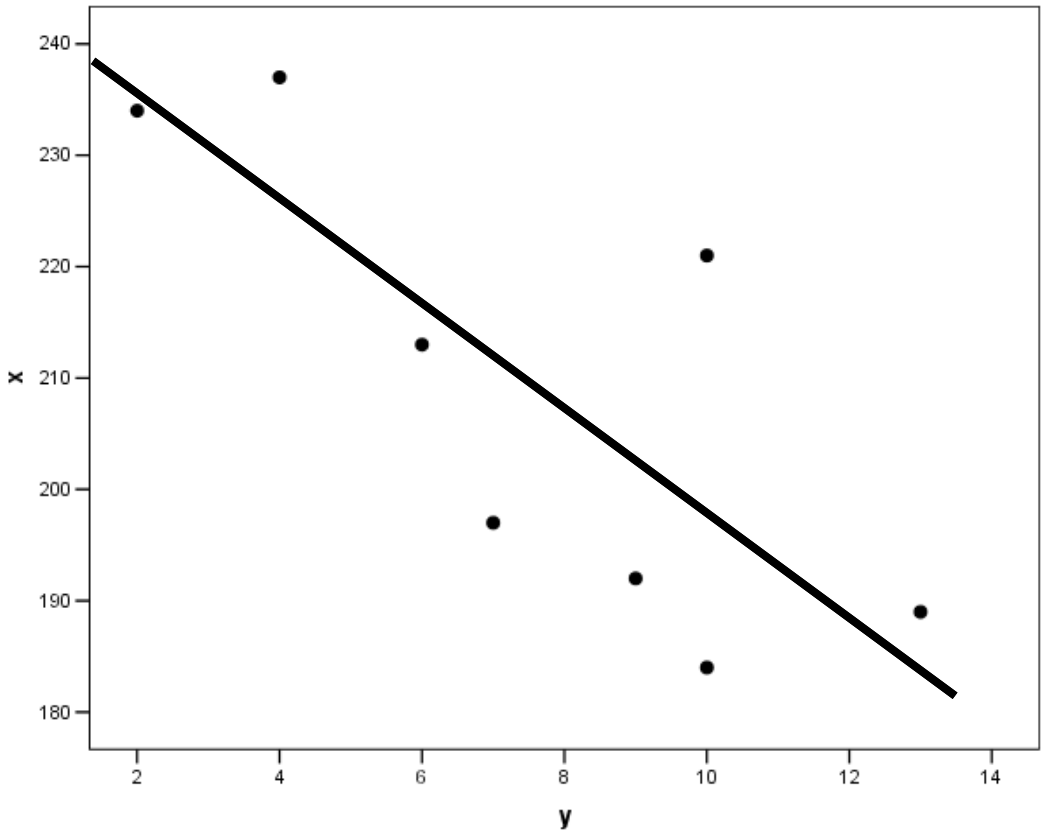
pentru test bilateral, α=0.05 şi df=6 (8-2)

	Levels of Significance for a One-Tailed Test			
	.05	.025	.01	.005
	Levels of Significance for a Two-Tailed Test			
df	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.755	.833	.875
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735

r critic=0.707

	tr (X)	X-m	$(X-m)^2$	erori (Y)	Y-m	$(Y-m)^2$	(X-m)* (Y-m)
1	184	-24,38	594,38	10	2,37	5,62	-57,78
2	213	4,62	21,34	6	-1,63	2,66	-7,53
3	234	25,62	656,38	2	-5,63	31,70	-144,24
4	197	-11,38	129,50	7	-,63	,40	7,17
5	189	-19,38	375,58	13	5,37	28,84	-104,07
6	221	12,62	159,26	10	2,37	5,62	29,91
7	237	28,62	819,10	4	-3,63	13,18	-103,89
8	192	-16,38	268,30	9	1,37	1,88	-22,44
Σ	1667		3023,88	61		89,88	-402,87
m_X	208,38			7,63			
s_X	20,784			3,583			

$$r = \frac{\sum (X - m_x) * (Y - m_y)}{N * s_x * s_y} = \frac{-402{,}87}{8 * 20{,}78 * 3{,}583} = \frac{-402{,}87}{595{,}14} = -0.68$$



r calculat = -0.68 < r critic=0.70

Decizia statistică?

Decizia cercetării?

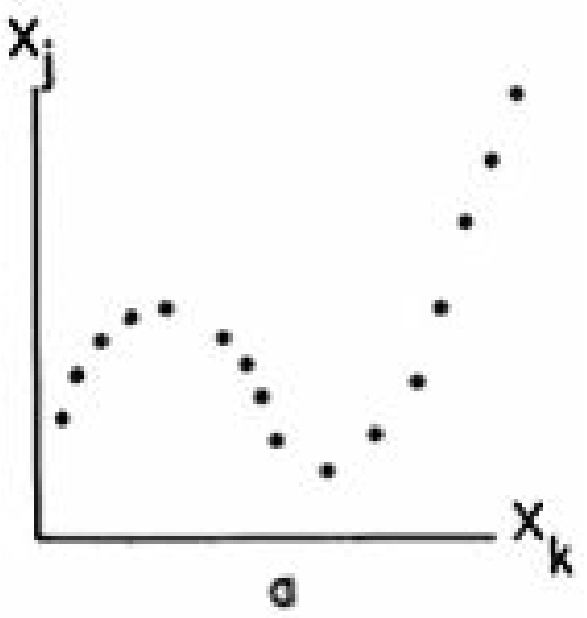
Interpretarea coeficientului de corelaţie

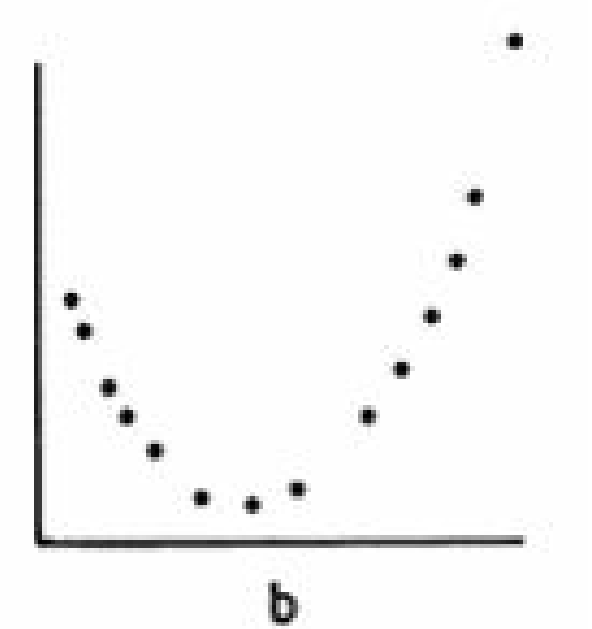
1) ***Corelaţie şi cauzalitate***
2) ***Natura liniară a corelaţiei Pearson***
3) ***Interpretarea valorii testului r***
4) ***Coeficientul de determinare***

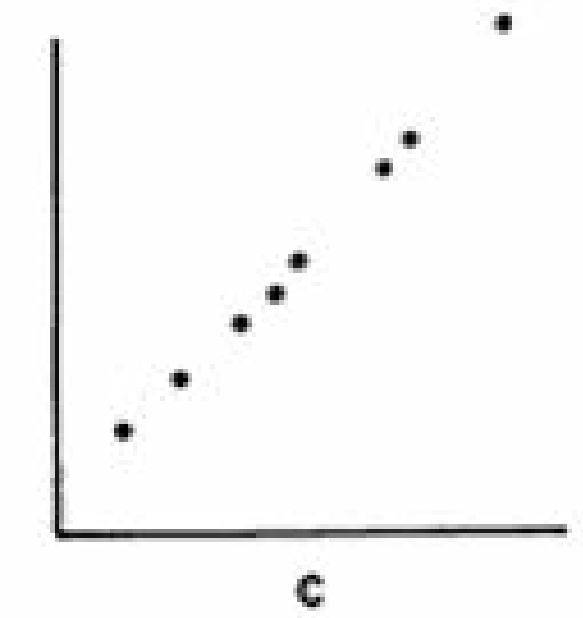
(1) Corelaţie şi cauzalitate

- Pearson (r) NU are semnificaţie cauzală
- relevă “legătura”, “asocierea”, variaţia concomitentă” a valorilor
- poate fi interpretat cauzal numai dacă variabilele sunt măsurate în condiţii de experiment

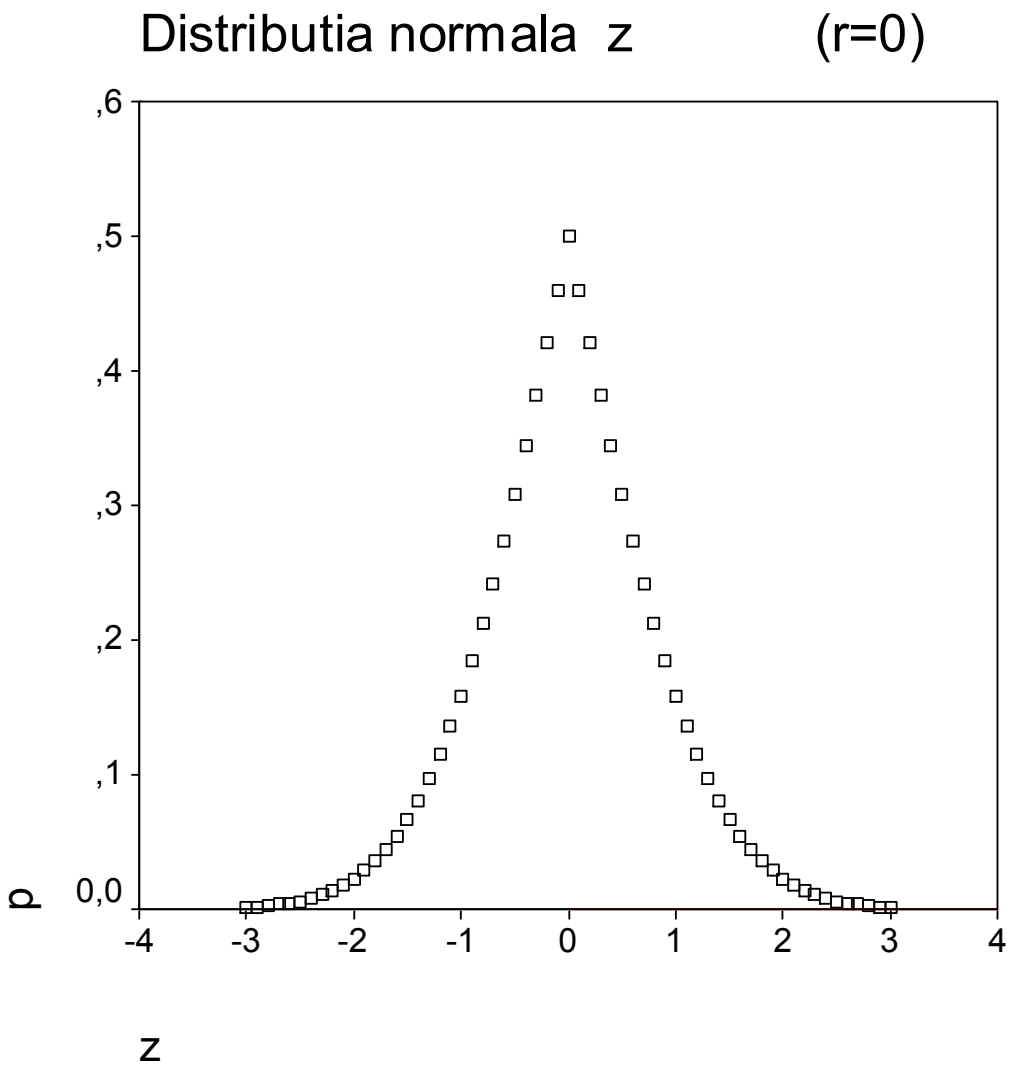
(2) Natura liniară a corelaţiei Pearson





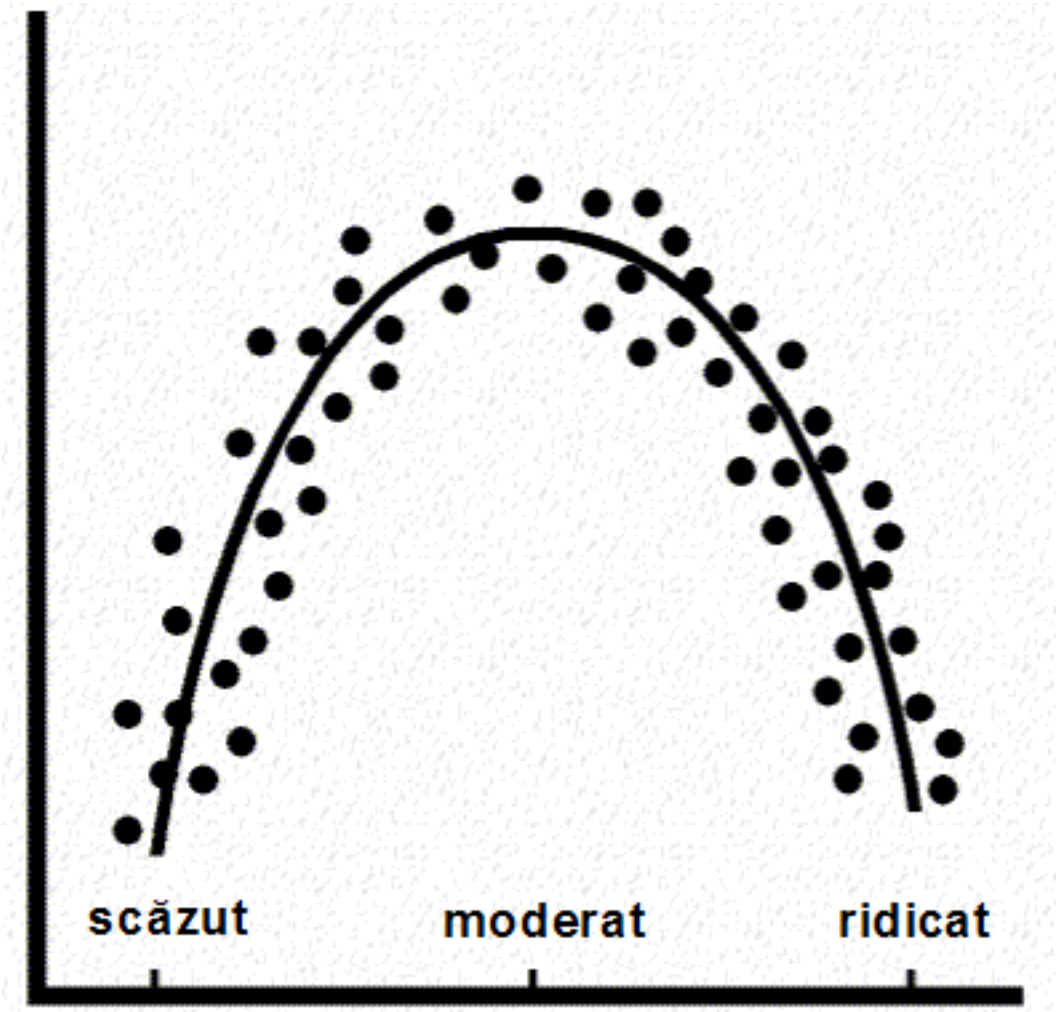


Corelaţia dintre valorile lui z şi probabilitatea aferentă de sub curba normală



Corelaţie curbilinie

performanţă şi nivelul stresului

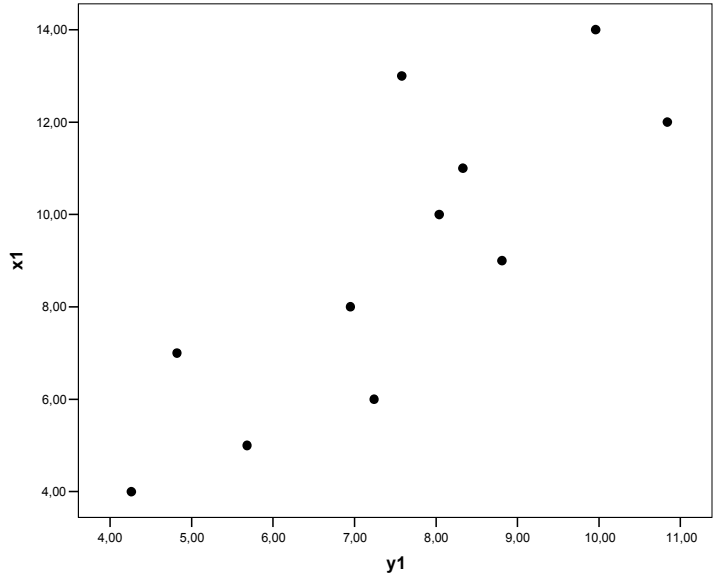


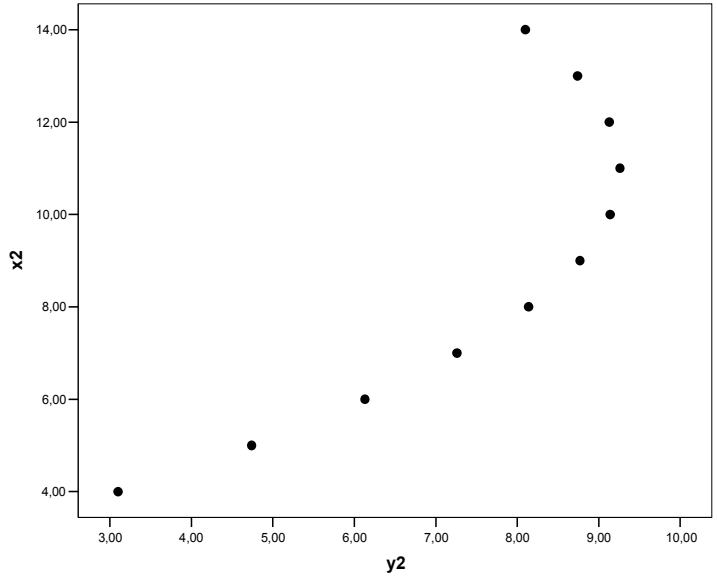
F. J. Anscombe, "Graphs in Statistical Analysis,"
American Statistician, 1973, **27**, 17-21

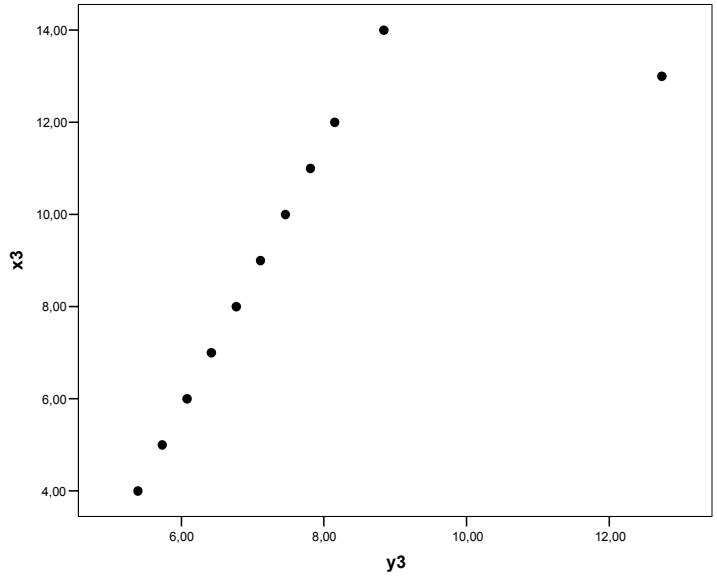
set #1		**set #2**		**set #3**		**set #4**	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10,00	8,04	10,00	9,14	10,00	7,46	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	5,76
13,00	7,58	13,00	8,74	13,00	12,74	8,00	7,71
9,00	8,81	9,00	8,77	9,00	7,11	8,00	8,84
11,00	8,33	11,00	9,26	11,00	7,81	8,00	8,47
14,00	9,96	14,00	8,10	14,00	8,84	8,00	7,04
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,25
4,00	4,26	4,00	3,10	4,00	5,39	19,00	12,50
12,00	10,84	12,00	9,13	12,00	8,15	8,00	5,56
7,00	4,82	7,00	7,26	7,00	6,42	8,00	7,91
5,00	5,68	5,00	4,74	5,00	5,73	8,00	6,89

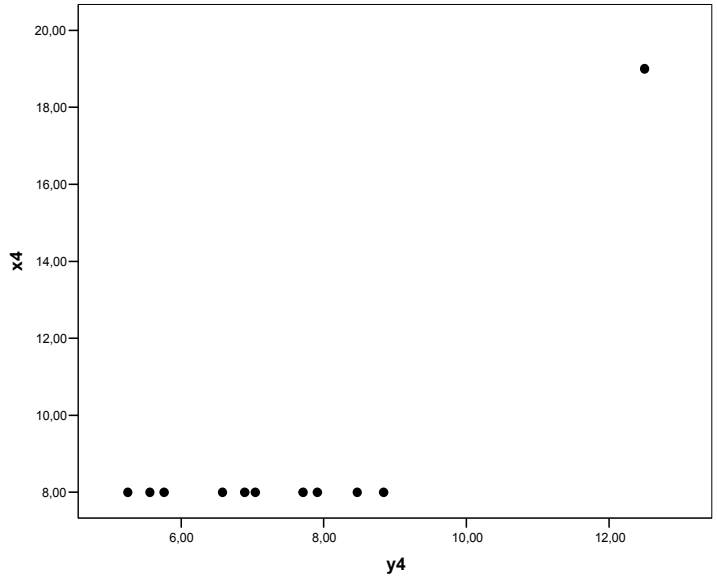
corelaţiile dintre toate cele patru seturi de date, două câte două, au aceeaşi valoare: r=0.816... şi totuşi...

Reprezentări scatterplot
pentru cele patru seturi de date Anscombe (r=0.81)









Mărimea efectului

- Valoarea însăşi a lui r
- Coeficientul de determinare (r^2)

*Interpretarea valorii testului r (*Hopkins*)*

Coeficientul de corelaţie	**Descriptor**
←0.1	Foarte mic, neglijabil, nesubstanţial
0.1↔0.3	Mic, minor
0.3↔0.5	Moderat, mediu
0.5↔0.7	Mare, ridicat, major
0.7↔0.9	Foarte mare, foarte ridicat
0.9→	Aproape perfect, descrie relaţia dintre două variabile practic indistincte

Interpretarea valorii testului r (Davis)

0.70 →	asociere foarte puternică
0.50 – 0.69	asociere substanţială
0.30 – 0.49	asociere moderată
0.10 – 0.29	asociere scăzută
0.01 – 0.09	asociere neglijabilă

Coeficientul de determinare (r^2)

r	r^2
1.00	**1.00**
.90	**.81**
.80	**.64**
.70	**.49**
.60	**.36**
.50	**.25**
.40	**.16**
.30	**.09**
.20	**.04**
.10	**.01**
.0	**.0**

coeficientul de determinare

r=0,68 r^2=0,46

46% din variaţia valorilor uneia dintre variabile este determintă de variaţia valorilor celeilalte variabile

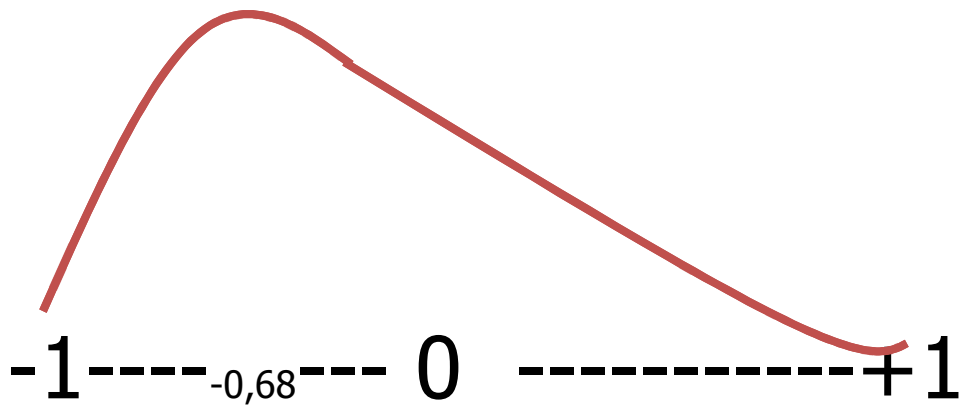
r^2 (Cohen)	0.0196	efect mic
	0.1300	efect mediu
	0.2600	efect mare

Limite de încredere pentru coeficientul de corelaţie

- Semnificaţia limitelor de încredere
  - ***r*** (calculat pentru eşantion) → estimare pentru ***ρ*** (*ro*)
  - putem evalua probabilitatea ca intensitatea asocierii în populaţie să se afle între anumite limite
  - aceste limite vor fi cu atât mai largi, cu atât acurateţea estimării r este mai scăzută
  - „distanţa" dintre limitele de încredere (superioară şi inferioară) este dată de „eroarea standard" a valorii calculate a lui ***r*** (simbolizată cu r_e)
    - variabilitatea estimată pentru o distribuţie de coeficienţi ***r***, pe care o vom numi ***rs*** (de la *sample distribution*, distribuţia de eşantionare)
  - principiul de calcul este acelaşi ca pentru media populaţiei

Calcularea limitelor de încredere pentru r

- Particularităţi:
  - Distribuția valorilor *r* la nivelul populației nu este simetrică decât pentru valoarea *r*=0



  - Fisher a elaborat un algoritm pe baza căruia valorile ***rs*** sunt transformate în valori Z, a căror arie de distribuţie sub curba normală este cunoscută:

$$Z = 0.5*\ln[(1 + r)/(1 - r)]$$

calculul limitelor de încredere pentru r

- r=-0.68
- Z ($r_{-0.68}$) = *-0.8291*
- Z critic=±1.96

$$\rho = r \pm z_{critic} * r_e$$

$$r_e = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{8-3}} = 0{,}447$$

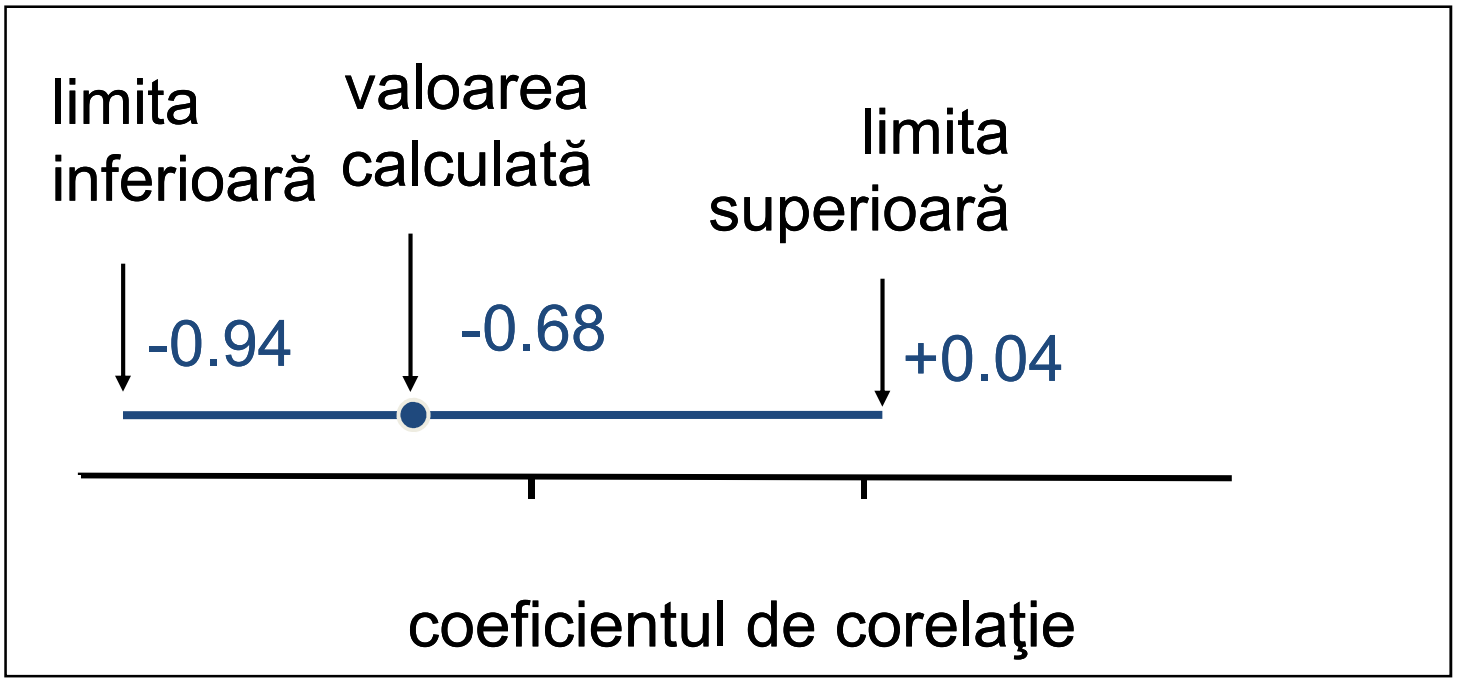
Limita superioară a intervalului (Z)..... (r)

$$\rho = -0.8291 + 1.96 * 0.447 = +0.04$$ **r=+0.04**

Limita inferioară a intervalului (Z)..... (r)

$$\rho = -0.8291 - 1.96 * 0.447 = -1.70$$ **r=-0.94**

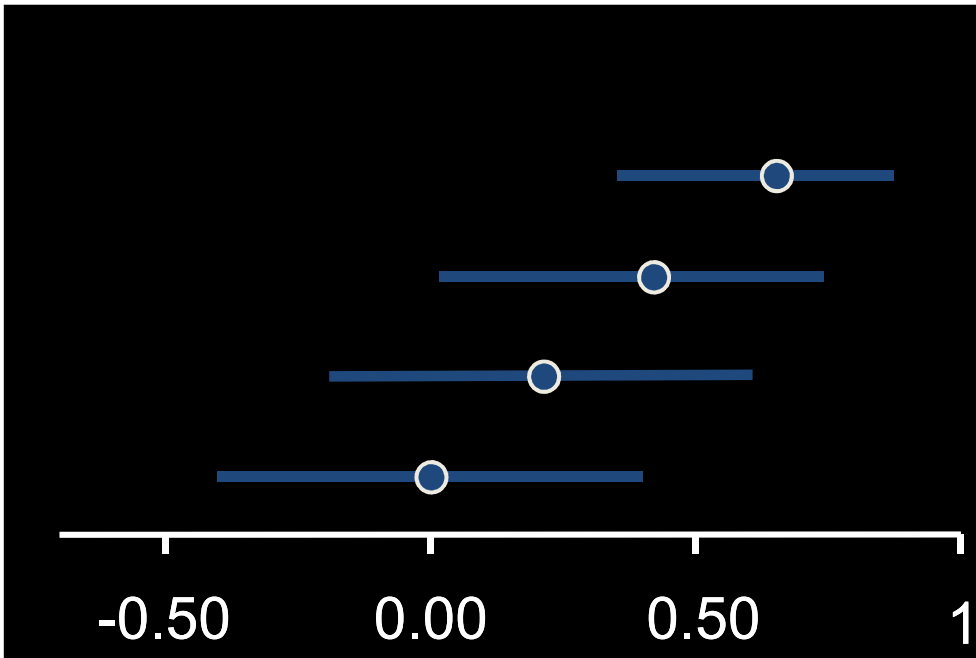
Z (r)



– am obţinut o corelaţie mare, dar valoarea adevărată, la nivelul populaţiei, se poate află oriunde, pe intervalul de la o valoare negativă, la una aproape perfectă.

Utilizarea limitelor de încredere

- r “statistic semnificativ"
  - P < 0.05
  - Zero cade în afara intervalului de încredere
    - Exemple: patru corelaţii pentru eşantioane de 20 subiecţi



r	interval	p
0.70	0.37 -- 0.87	0.007
0.44	0.00 -- 0.74	0.05
0.25	-0.22 -- 0.62	0.29
0.00	-0.44 -- 0.44	1.00

Un exemplu

- Limitele de încredere pentru acesta sunt între -0.07 şi +0.60
  - nesemnificativ, (între cele două limite este şi valoarea zero)
- Cu cât N va fi mai mare, cu atât valoarea lui r_e va fi mai mică iar limitele intervalului de încredere pentru ***r***, mai aproape de ***r***.
- Dacă am creşte volumul eşantionului la 50 de subiecţi, limita inferioară ar trece dincolo de valoarea zero.
- Celelalte linii din tabel prezintă efectul de mărime al eşantionului în cazul creşterii lui N până la 100 de subiecţi.

Eşantion N=30; r=0.30

N	Pearson *r*	Niv. de încredere (%)	Limite de încredere	
			inferioară	Superioară
30	0,30	95	-0,07	0,60
40	0,30	95	-0,01	0,56
50	0,30	95	0,02	0,53
60	0,30	95	0,05	0,51
70	0,30	95	0,07	0,50
80	0,30	95	0,09	0,49
90	0,30	95	0,10	0,48
100	0,30	95	0,11	0,47

pentru exemplul nostru

- dacă N=10
- r_e=1/sqrt(7)=0.38
- lim. sup.=-0.8291+1.96*0.38=-0.08 (r= -0.08)
- lim. inf. =-0.8291-1.96*0.38=-1.57 (r= -0.93)
- cu numai 2 subiecţi în plus, rezultatul devenea semnificativ

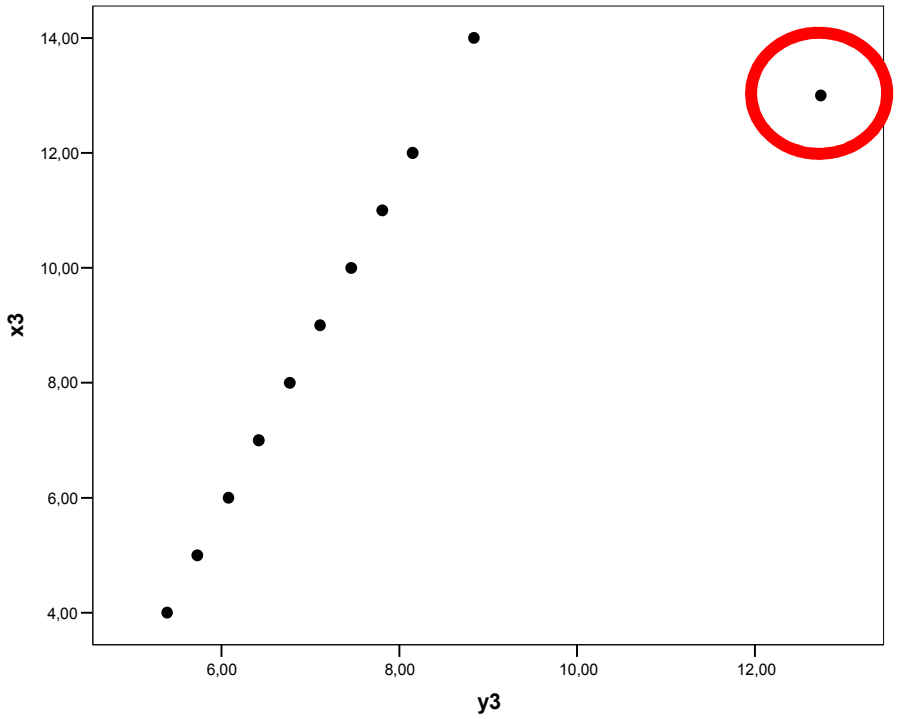
Semnificaţia diferenţei dintre doi coeficienţi de corelaţie

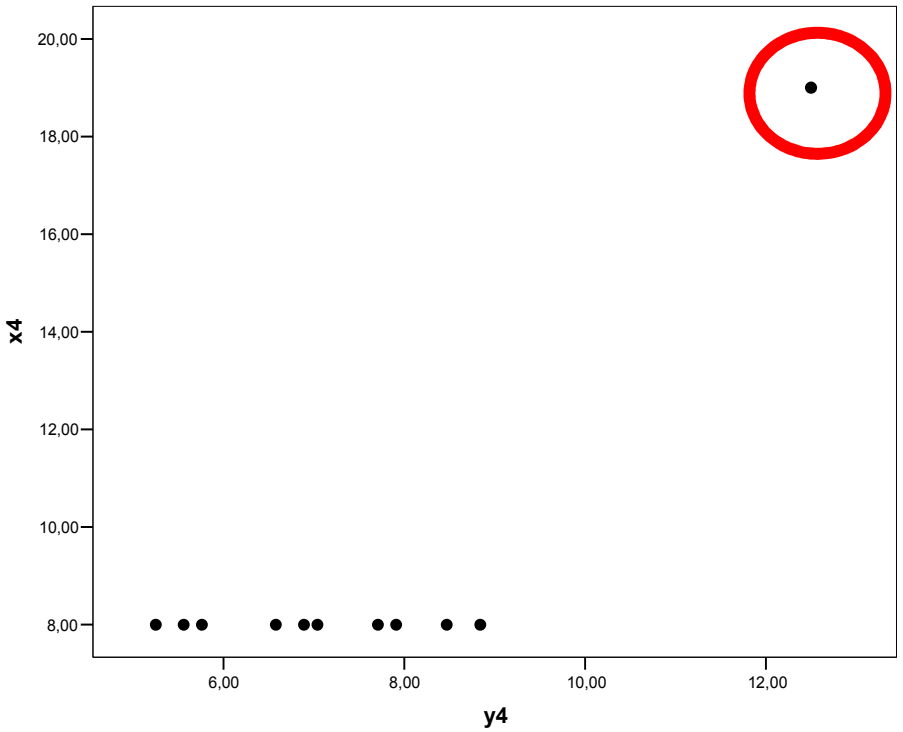
- corelaţia dintre extraversie şi agresivitate
  - separat, pentru bărbaţi şi pentru femei
    - *r*=0.50 pentru bărbaţi
    - *r*=0.30 pentru femei
    - Ambii semnficativi
- Semnificaţia diferenţei ia în considerare:
  - diferenţa dintre valorile *r*
  - mărimea eşantioanelor
  - Mărimea a celor doi coeficienţi
- De exemplu, o diferenţă de 0.1 între doi indici de corelaţie
  - poate fi nesemnficativă dacă cei doi *r* sunt 0.15 şi 0.25
  - poate fi semnificativă dacă valorile *r* comparate sunt 0.80 şi 0.90.

Condiţii pentru calcularea coeficientului de corelaţie Pearson

- eşantionul aleatoriu
- variabile cu distribuţie care să nu se abată grav de la distribuţia normală
- condiţie este cu atât mai importantă cu cât eşantionul este mai mic
- atenţie aparte trebuie acordată valorilor excesive, prezenţa acestora putând avea efecte neaşteptate asupra valorii coeficientului de corelaţie
  - vezi seturile Anscombe

Efectul valorilor extreme (bivariate) asupra lui r
Anscombe (r=0.81)





Utilizarea coeficientul de corelaţie

- Analiza de corelaţie este una dintre cele mai uzuale proceduri statistice în cercetarea psihologică
  - *consistenţa testelor (internă, test-retest)*
  - *validităţii* testelor psihologice
- testul t (dep) sau r?

Publicarea rezultatului corelaţiei

- *„A fost evaluată relaţia dintre numărul conduitelor agresive emise şi cel al aprecierilor primite, pe un grup de 8 elevi. Media conduitelor agresive a fost de m=20.68 (s=20.78) iar a aprecierilor primite m=7.63 (s=3.58). Am rezultat o corelaţie negativă, nesemnificativă, între cele două tipuri de conduite, r(6)=-0.68, p>0.05, bilateral.”*

Tabela Fisher de transformare a valorilor *r* în scoruri Z

(Sursa: http://davidmlane.com/hyperstat/rtoz_table.html)

r	Z	r	Z	r	Z	R	Z
0.0000	*0.0000*	0.2600	*0.2661*	0.5200	*0.5763*	0.7800	*1.0454*
0.0100	*0.0100*	0.2700	*0.2769*	0.5300	*0.5901*	0.7900	*1.0714*
0.0200	*0.0200*	0.2800	*0.2877*	0.5400	*0.6042*	0.8000	*1.0986*
0.0300	*0.0300*	0.2900	*0.2986*	0.5500	*0.6184*	0.8100	*1.1270*
0.0400	*0.0400*	0.3000	*0.3095*	0.5600	*0.6328*	0.8200	*1.1568*
0.0500	*0.0500*	0.3100	*0.3205*	0.5700	*0.6475*	0.8300	*1.1881*
0.0600	*0.0601*	0.3200	*0.3316*	0.5800	*0.6625*	0.8400	*1.2212*
0.0700	*0.0701*	0.3300	*0.3428*	0.5900	*0.6777*	0.8500	*1.2562*
0.0800	*0.0802*	0.3400	*0.3541*	0.6000	*0.6931*	0.8600	*1.2933*
0.0900	*0.0902*	0.3500	*0.3654*	0.6100	*0.7089*	0.8700	*1.3331*
0.1000	*0.1003*	0.3600	*0.3769*	0.6200	*0.7250*	0.8800	*1.3758*
0.1100	*0.1104*	0.3700	*0.3884*	0.6300	*0.7414*	0.8900	*1.4219*
0.1200	*0.1206*	0.3800	*0.4001*	0.6400	*0.7582*	0.9000	*1.4722*
0.1300	*0.1307*	0.3900	*0.4118*	0.6500	*0.7753*	0.9100	*1.5275*
0.1400	*0.1409*	0.4000	*0.4236*	0.6600	*0.7928*	0.9200	*1.5890*
0.1500	*0.1511*	0.4100	*0.4356*	0.6700	*0.8107*	0.9300	*1.6584*
0.1600	*0.1614*	0.4200	*0.4477*	0.6800	*0.8291*	0.9400	*1.7380*
0.1700	*0.1717*	0.4300	*0.4599*	0.6900	*0.8480*	0.9500	*1.8318*
0.1800	*0.1820*	0.4400	*0.4722*	0.7000	*0.8673*	0.9600	*1.9459*
0.1900	*0.1923*	0.4500	*0.4847*	0.7100	*0.8872*	0.9700	*2.0923*
0.2000	*0.2027*	0.4600	*0.4973*	0.7200	*0.9076*	0.9800	*2.2976*
0.2100	*0.2132*	0.4700	*0.5101*	0.7300	*0.9287*	0.9900	*2.6467*
0.2200	*0.2237*	0.4800	*0.5230*	0.7400	*0.9505*		
0.2300	*0.2342*	0.4900	*0.5361*	0.7500	*0.9730*		
0.2400	*0.2448*	0.5000	*0.5493*	0.7600	*0.9962*		
0.2500	*0.2554*	0.5100	*0.5627*	0.7700	*1.0203*