KORELACIJA

PREDAVANJE BR.12

Tipovi veza među varijablama

- Funkcionalne veze ($Y=f(X)$)
- Uzročno-posljedične veze (X uzrok Y)
- Stohastičke veze (X povezana sa Y)

- Linearne
- Nelinearne veze

Značenje korelacije

Među varijablama postoji povezanost:

- Varijable zajedno variraju: promjene na jednoj varijabli su praćene "srodnim" promjenama na drugoj varijabli;
- Poznavanje rezultata za neku jedinicu posmatranja na jednoj varijabli pomaže da bolje predvidimo njen rezultat na drugoj varijabli nego što bismo mogli bez tog poznavanja
- VAŽNO: Korelacija među varijablama ne znači nužno uzočno-posljedičnu vezu!

Koeficijent linearne korelacije

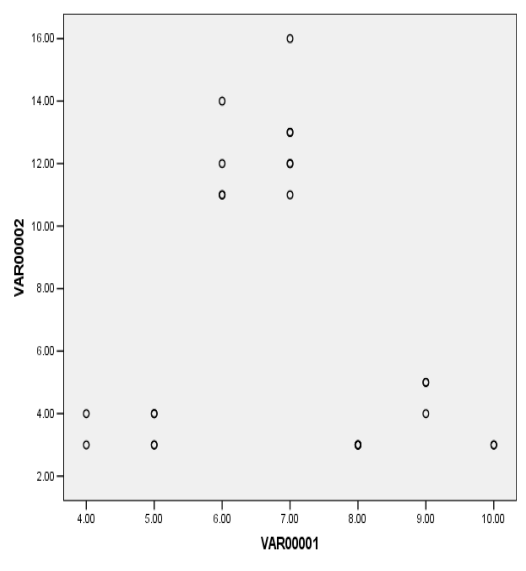
- Brojčana vrijednost koja nam pokazuje stepen linearne povezanosti između dvije varijable (kreće se od -1 do 1)
- Varijable mogu biti u jakoj nelinearnoj vezi a da koeficijentom linearne korelacije to ne otkrijemo!
- Zato je važno da uvijek prvo pogledamo dijagram raspršenja prije računanja koeficijenta linearne korelacije.
- Na dijagramu se mogu učiti i ekstremne vrijednosti (nestandardne opservacije ili autlajer) - engl.outliers

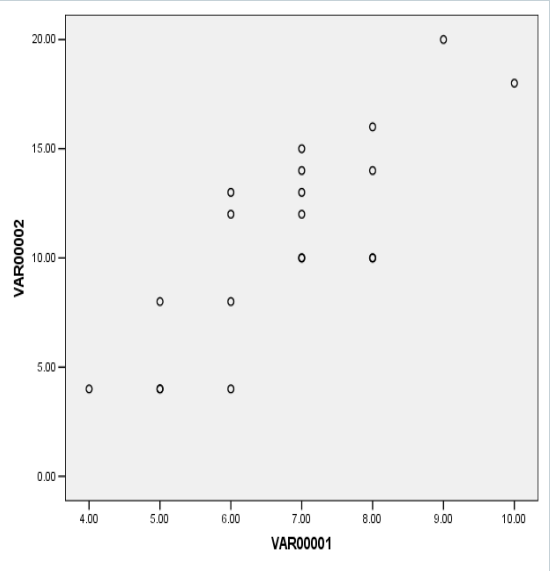
- Pearsonov koeficijent linearne korelacije:

$$r = \frac{\sum(z_x z_y)}{N-1} \qquad r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]}\sqrt{\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

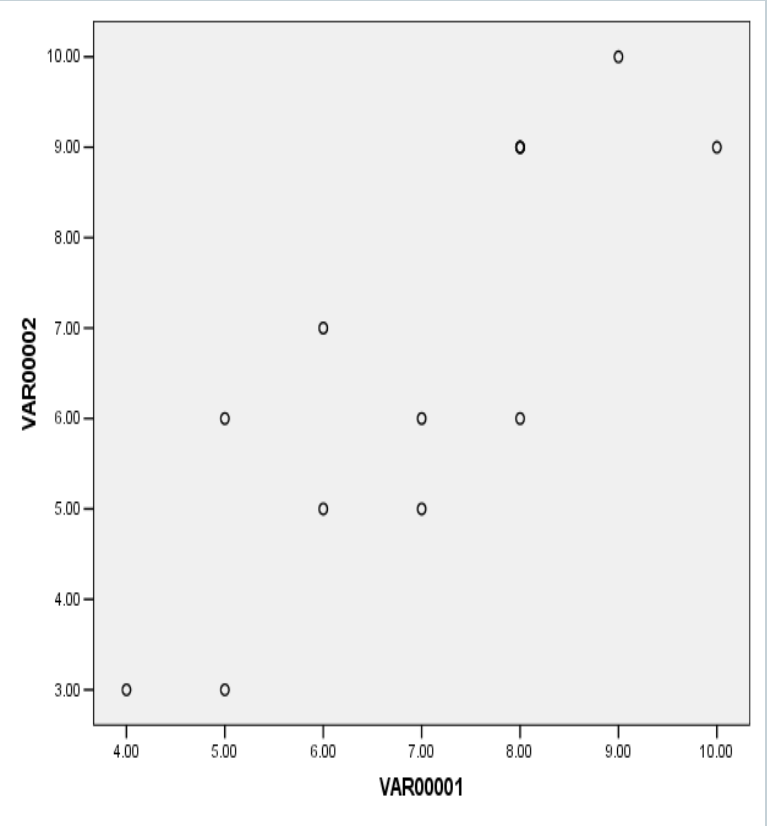
Grafički prikaz povezanosti između varijabli

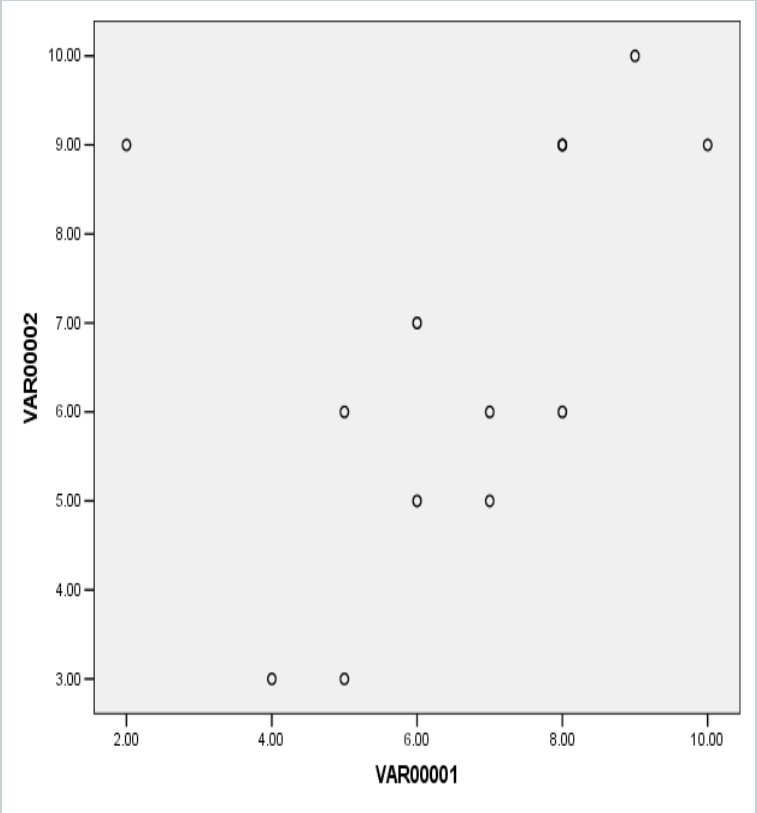
- Postoji li povezanost?
- Linearna ili nelinearna?
- Pozitivna ili negativna?
- Jaka ili slaba?





Grafički prikaz – uticaj outliers-a

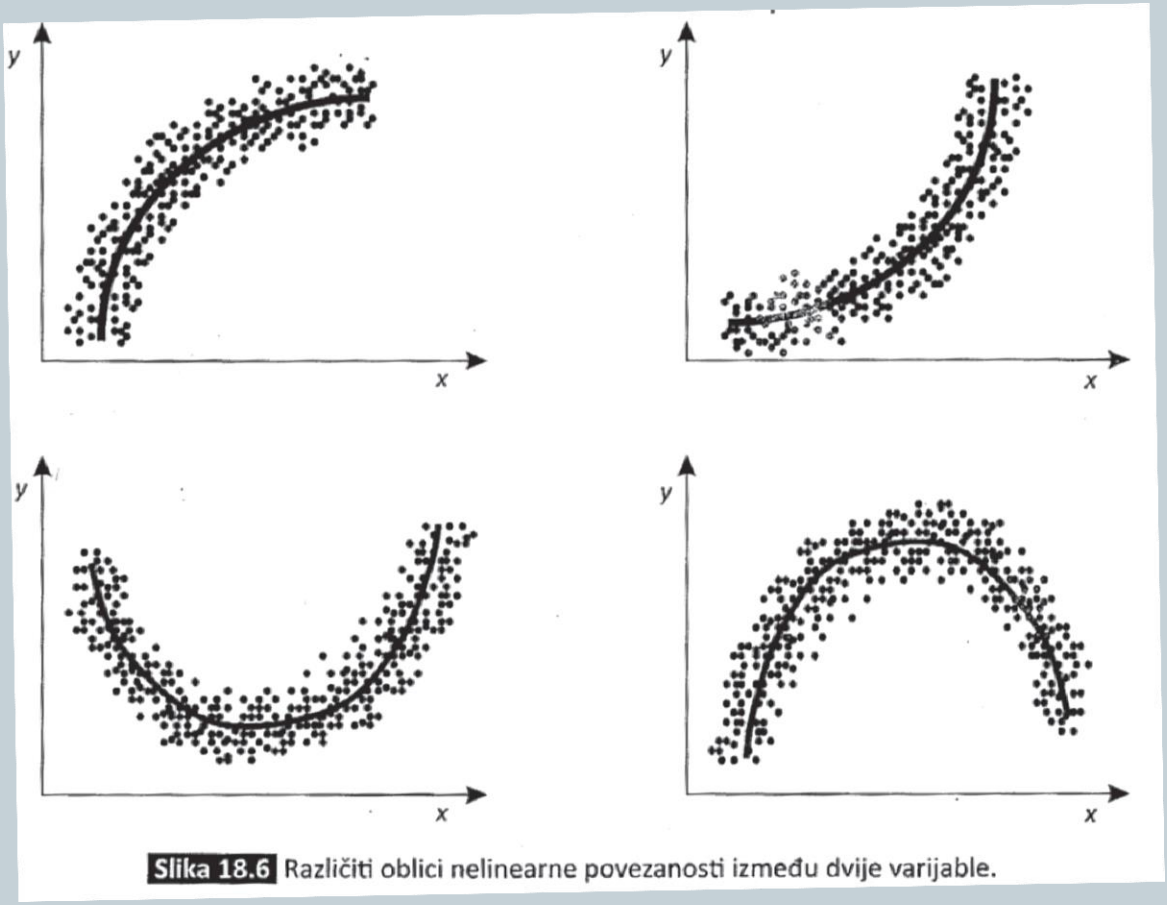




Slika 1 – bez outlier-a (r=0,84)

Slika 2 – sa outlier-om (r=0,44)

Oblici nelinearne povezanosti



Slika 18.6 Različiti oblici nelinearne povezanosti između dvije varijable.

Opravdanost računanja Pearsonovog koeficijenta linearne korelacije

- Rezultati moraju biti prave mjerne vrijednosti, izražene barem na intervalnoj skali
- Mora postojati dovoljan broj rezultata (obično 30)
- Distribucije za obije varijable moraju biti simetrične
- Smije se računati samo ako je povezanost linearna
- Zadovoljena pretpostavka homoskedastičnosti (pojednakog varijabiliteta u Y za sve nivoe X)

Primjer 1 – računanje koef. linearne korelacije

Hipotetički primjer: Osmorica ispitanika je testirana uz pomoć testova X i Y, a postignuti rezultati su prikazani u narednoj tabeli.

Ispitanici	Test X	Test Y	Zx	Zy	ZxZy
A	7	19	1,5	1,5	2,25
B	6	17	1,0	1,0	1
C	5	15	0,5	0,5	0,25
D	4	13	0,0	0,0	0
E	4	13	0,0	0,0	0
F	3	11	-0,5	-0,5	0,25
G	2	9	-1,0	-1,0	1
H	1	7	-1,5	-1,5	2,25
∑	**32**	**104**			**7**
M	**4**	**13**			
SD	**2**	**4**			

$$r = \frac{\sum(z_x z_y)}{N-1} = \frac{7}{7} = 1$$

Primjer 2 – računanje koeficijenta linearne korelacije

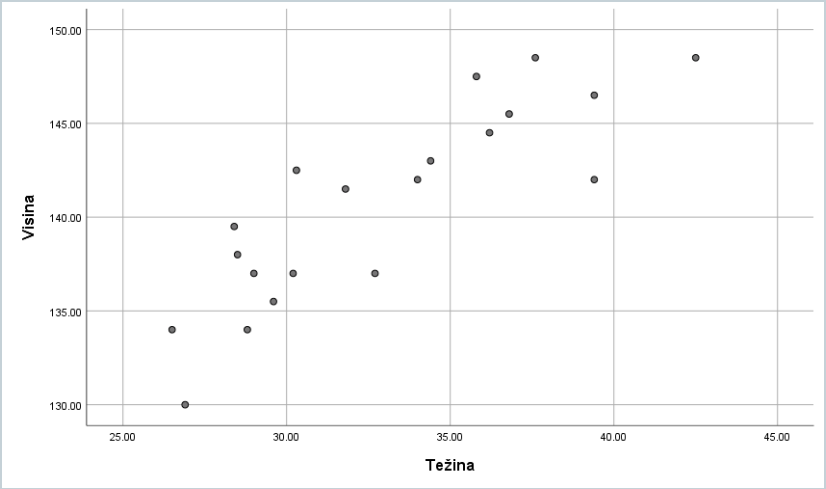
Mjerene su visina i težina 12-godišnjih dječaka u jednoj školi:

Ispitanici	Visina (X)	Težina (Y)	X^2	Y^2	XY
1	139,5	28,4	19.460,25	806,56	3.961,80
2	148,5	37,6	22.052,25	1.413,76	5.583,60
3	138	28,5	19.044,00	812,25	3.933,00
4	142	39,4	20.164,00	1.552,36	5.594,80
5	141,5	31,8	20.022,25	1.011,24	4.499,70
6	137	32,7	18.769,00	1.069,29	4.479,90
7	134	26,5	17.956,00	702,25	3.551,00
8	137	30,2	18.769,00	912,04	4.137,40
9	143	34,4	20.449,00	1.183,36	4.919,20
10	135,5	29,6	18.360,25	876,16	4.010,80
11	142	34	20.164,00	1.156,00	4.828,00
12	137	29	18.769,00	841,00	3.973,00
13	147,5	35,8	21.756,25	1.281,64	5.280,50
14	134	28,8	17.956,00	829,44	3.859,20
15	144,5	36,2	20.880,25	1.310,44	5.230,90
16	146,5	39,4	21.462,25	1.552,36	5.772,10
17	130	26,9	16.900,00	723,61	3.497,00
18	142,5	30,3	20.306,25	918,09	4.317,75
19	148,5	42,5	22.052,25	1.806,25	6.311,25
20	145,5	36,8	21.170,25	1.354,24	5.354,40
∑	**2.814,00**	**658,80**	**396.462,50**	**22.112,34**	**93.095,30**

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]}\sqrt{\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

$$= \frac{20 * 93095,3 - 2814 * 658,8}{\sqrt{(20 * 396462,5281 - 2814^2)(20 * 22112,34 - 658,8^2)}}$$

$$r = 0,86$$



Testiranje značajnosti koeficijenta korelacije: t test

- Ho: ρ = 0 (ρ je koef.linearne korelacije u populaciji)
- T statistik je:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

- Ako je Ho tačna, onda t-stat ima Studentovu distribuciju sa n-2 broja stepeni slobode
- Vrijednost t-stat zavisi od veličine koef.linearne korelacije i od veličine uzorka
- O interpretacija koeficijenta se može raspravljati samo nakon što je utvrđena njegova značajnost

Nastavak primjera 2

Kritična vrijednost iz tablica B u Dodatku (2,101)

$$t = r\sqrt{\frac{N-2}{1-r^2}} = 0,86\frac{\sqrt{18}}{\sqrt{1-0,74}} = 7,15$$

Odbacujemo nultu hipotezu. Korelacija je statistički značajna.

Korelacija rangova

Spearmanov koeficijent korelacije rangova

$$\rho_0 = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

x	y	d
1	3	1-3
2	2	2-2
3	1	3-1

- D – razlike parova rangova
- Testiranje statističke značajnosti – t statistik ima studentovu raspodjelu sa df=n-2
- N – broj parova rangova
- Rang korelacija daje samo približnu indikaciju asocijacije između varijabli
- Za razliku od Pearsonovog koeficijenta, kod rang korelacije nije neophodno da varijable budu u linearnom odnosu

Primjer 3

- Jedan Profesor je testirao svojih 15 studenata jednim testom i poređao ih po uspjehu koji su postigli na prijemnom ispitu. Da li postoji i kolika je korelacija između uspjeha na testu i uspjeha na prijemnom ispitu?

Student	Rang na prijemnom ispitu	Bodovi na testu	Rang na testu	D	D^2
A	1	22	1	0	0
B	2	19	3	-1	1
C	3	6	15	-12	144
D	4	18	4	0	0
E	5	20	2	3	9
F	6	16	5	1	1
G	7	11	10	-3	9
H	8	9	12	-4	16
I	9	15	6	3	9
J	10	12	8,5	1,5	2,25
K	11	10	11	0	0
L	12	7	14	-2	4
M	13	13	7	6	36
N	14	12	8,5	5,5	30,25
O	15	8	13	2	4
					265,5

$$\rho_0 = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$$\rho_0 = 1 - \frac{6*265,5}{15(225-1)} = 1 - 0,474 = 0,526$$

- Ako postoji veliki broj zajedničkih rangova, koristi se korigovana formula.

Primjer 3

Correlations

			Rang_prijemni	Rank of Test
Spearman's rho	Rang_prijemni	Correlation Coefficient	1.000	.525*
		Sig. (2-tailed)	.	.044
		N	15	15
	Rank of Test	Correlation Coefficient	.525*	1.000
		Sig. (2-tailed)	.044	.
		N	15	15

*. Correlation is significant at the 0.05 level (2-tailed).

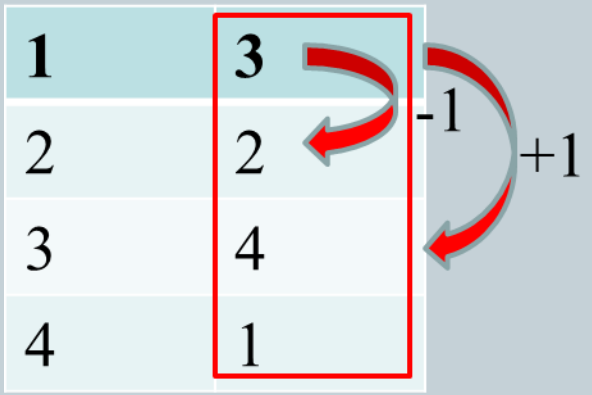
- SPSS: Analyze/Correlate/Bivariate…
- Ista procedura kao kod Pearsonovog koeficijenta

Korelacija rangova

Kendallov "tau" koeficijent rang korelacije

- Poređani rangovi jednog niza u prirodnom rastućem redosljedu, a drugi niz rangova se dodaje
  - +1 ako je prvo niži pa viši (prirodni redosljed)
  - -1 ako je prvo viši pa niži (inverzni redosljed)
  - S = suma svih +1 i -1

$$\tau = \frac{S}{(N(N-1)/2)}$$



Primjer 4

Primjer računanja Kendallovog "tau" koeficijent rang korelacije na hipotetičkom primjeru:

Rang X	1	2	3	4	5							
Rang Y	**1**	**4**	**3**	**5**	**2**							**S**
		+1	**+1**	**+1**	**+1**	**-1**	**+1**	**-1**	**+1**	**-1**	**-1**	**2**

$$\tau = \frac{S}{(N(N-1)/2)} = \frac{2}{20/2} = 0,2$$

Correlations

			VAR00001	VAR00002
Kendall's tau_b	VAR00001	Correlation Coefficient	1.000	.200
		Sig. (2-tailed)	.	.624
		N	5	5
	VAR00002	Correlation Coefficient	.200	1.000
		Sig. (2-tailed)	.624	.
		N	5	5

KORELACIJA RANGOVA:
Spirmanov i Kendalov postupak

- Kendalov tau koeficijent korelacije rangova niži je od Spirmanovog za iste podatke, ali je zaključak o značajnosti uobičajeno isti za oba koeficijenta;

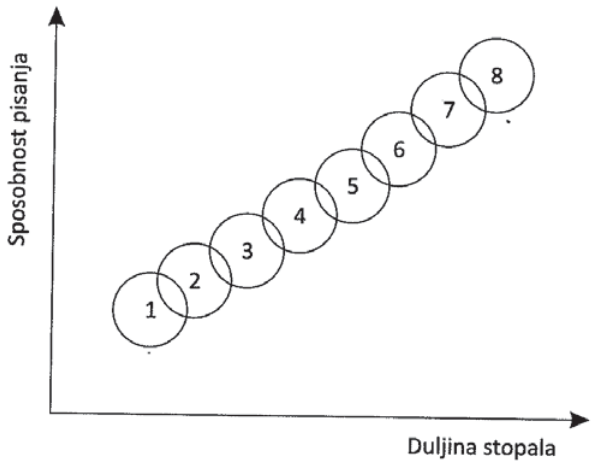
- Kendalov koeficijent bolje ocjenjuje korelaciju između rangova u populaciji;

- Kendalov postupak razrađen je i za računanje parcijalne korelacije dva niza rangova kada se isključi uticaj treće varijable.

Parcijalna korelacija

- Korelacija između dvije varijable kod koje isključujemo uticaj jednog (ili više) faktora koji nam smetaju, odnosno koji izazivaju pogrešne zaključke
- Primjer: Na uzorku školske djece uzrasta od 7 do 15 godina, u svakom školskom razredu mjerena je dužina stopala i sposobnost pisanja.
- Primijetimo: kod oba svojstva prosječni rezultati rastu s godinama djeteta
- Da bismo dobili stvarnu korelaciju, treba isključiti uticaj odrastanja
- Varijabla 1 – dužina stopala; varijabla 2 – sposobnost pisanja; varijabla 3 – starosna dob

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$



Parcijalna korelacija - primjer

Primjer: Na uzorku školske djece uzrasta od 7 do 15 godina, u svakom školskom razredu mjerena je dužina stopala i sposobnost pisanja.

- Korelacija između dužine stopala i sposobnosti pisanja (u rasponu od 7 do 15 god) je 0,69.
- Korelacija između dužine stopala i starosne dobi je 0,90
- Korelacija između sposobnosti pisanja i starosne dobi je 0,75

Parcijalna (stvarna) korelacija između dužine stopala I pisanja koja isključuje uticaj starosne dobi je:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0,69 - 0,90 * 0,75}{\sqrt{1-0,81}\sqrt{1-0,56}} = 0,05$$

Šta možemo zaključiti?

Još neki koeficijenti korelacije

"Point-biserijalni" koeficijent korelacije

- Kad nas zanima korelacija između jedne kontinuirane varijable i jedne dihotome varijable (muško/žensko; položio/pao)

Koeficijent konkordancije

- Slaganje između više nizova rangova

Fi koeficijent

- Sjetimo se veze sa hi-kvadrat testom

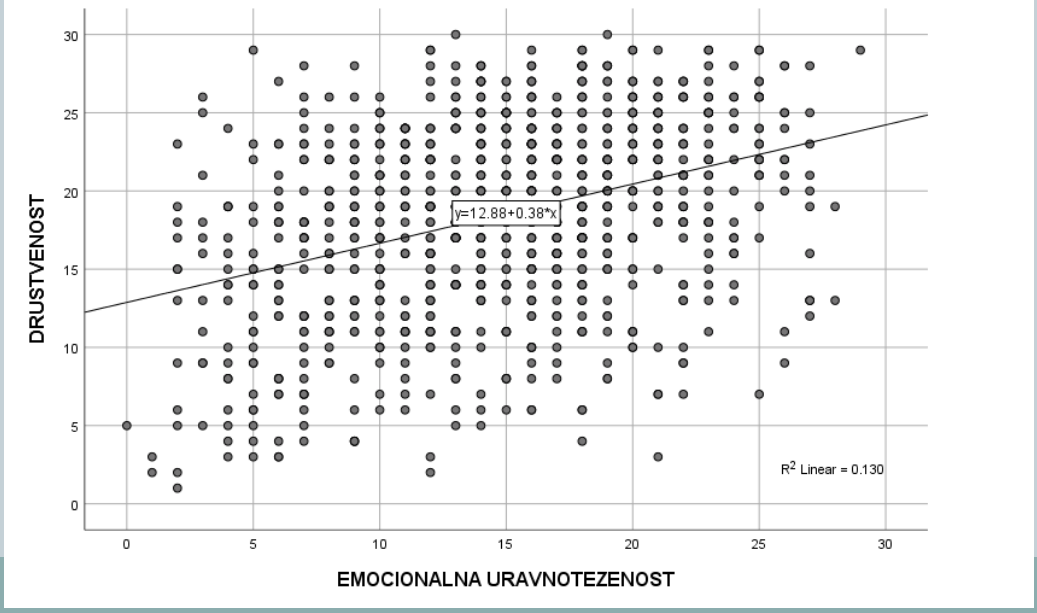
Koeficijent kontigencije

- Takođe kod hi-kvadrat testa

Zadatak / SPSS

Na bazi podataka o društvenosti i emocionalnoj uravnoteženosti grupe od 960 studenata:

- Grafički prikazati vezu između društvenosti i emocionalne uravnoteženosti
- Izračunati koef.linearne korelacije između ove 2 varijable
- Testirati statističku značajnost koeficijenta lin.korelacije



Correlations

		Drustvenost	Emocionalna uravnotezenost
Drustvenost	Pearson Correlation	1	.361**
	Sig. (2-tailed)		.000
	N	960	960
Emocionalna uravnotezenost	Pearson Correlation	.361**	1
	Sig. (2-tailed)	.000	
	N	960	960

**. Correlation is significant at the 0.01 level (2-tailed).