

ZX510 Pokročilé statistické metody geografického výzkumu

Téma: Měření síly asociace mezi proměnnými (korelační analýza)



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



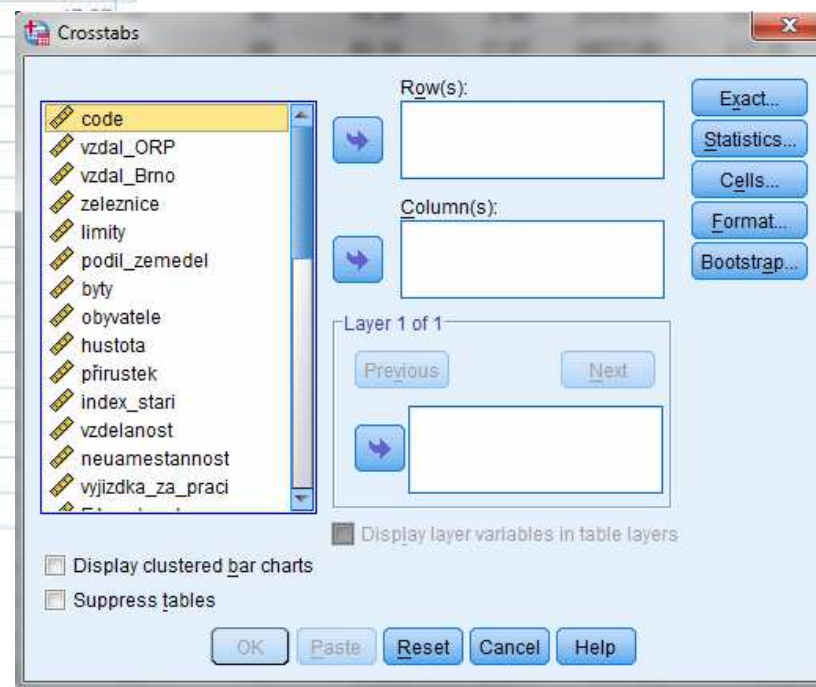
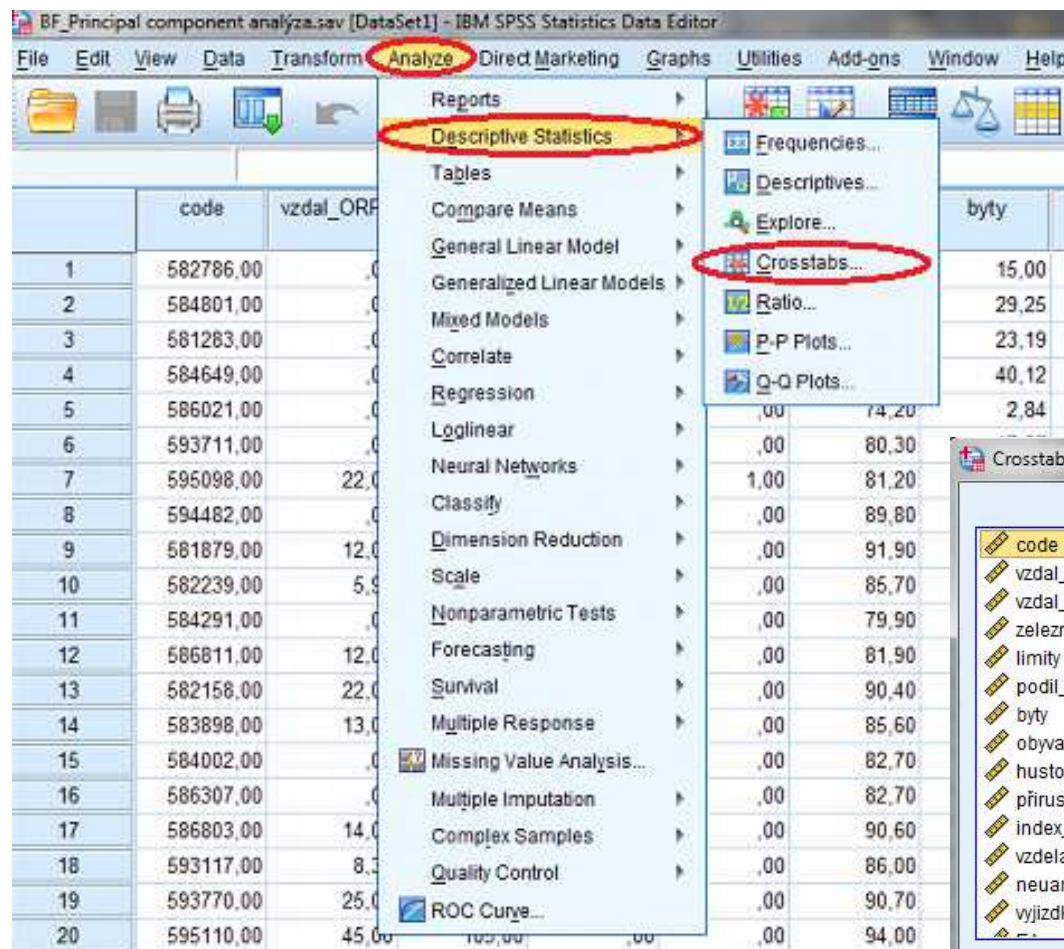
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Měření síly asociace (korelace) mezi proměnnými

- Vztah mezi dvěma proměnnými existuje, pokud hodnoty jedné proměnné jsou vztaženy k hodnotám druhé proměnné (tzv. kovariují).
- zajímáme se nejen o to, zdali je mezi proměnnými souvislost, ale také jakou má tato souvislost sílu (těsnost vztahu) a jakou má povahu (jaký má směr)
- pro zjištění síly či těsnosti vztahu počítáme tzv. koeficienty asociace (u nominálních znaků) nebo korelace (ordinální, kardinální)

Korelační analýza v SPSS (funkce Crosstabs)



Korelační analýza v SPSS (funkce Compare means)

The screenshot displays the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, showing the 'Compare Means' option. The 'Means' dialog box is open, showing 'code' as the dependent variable and 'vzda_ORP' as the independent variable. The 'Means: Options' sub-dialog box is also open, showing 'Median' selected under 'Statistics' and 'Anova table and eta' selected under 'Statistics for First Layer'.

Means Dialog Box:

- Dependent List: code
- Independent List: vzda_ORP

Means: Options Dialog Box:

- Statistics: Median, Grouped Median, Std. Error of Mean, Sum, Minimum, Maximum, Range, First, Last, Variance, Kurtosis, Std. Error of Kurtosis, Skewness, Std. Error of Skewness, Harmonic Mean
- Cell Statistics: Mean, Number of Cases, Standard Deviation
- Statistics for First Layer: ☒ Anova table and eta, ☐ Test for linearity

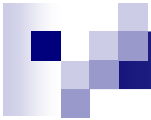
Data Table:

code	vzda_ORP
1	582786,00
2	584801,00
3	581283,00
4	584649,00
5	586021,00
6	593711,00
7	595098,00
8	594482,00
9	581879,00
10	582239,00
11	584291,00
12	586811,00
13	582158,00
14	583898,00
15	584002,00
16	586307,00
17	586803,00
18	593117,00
19	593770,00
20	595110,00



Typy korelačních koeficientů

- Hodnoty koeficientů se pohybují v intervalech $<0;1>$ (u nominálních znaků nemá smysl určovat směr vztahu) nebo $<-1;+1>$
- Čím vyšší je hodnota koeficientů, tím silnější je vztah mezi proměnnými
- Záporné koeficienty znamenají negativní asociaci a kladné koeficienty pozitivní asociaci.
- Nula má obvykle význam neexistence vztahu (vztah může existovat, ale je nelineární!), hodnota 1 má význam existence perfektního vztahu.
- Pro každou úroveň měření (pro typ proměnných - nominální, ordinální a kardinální) jsou určeny speciální koeficienty.



Typy korelačních koeficientů

- Jsou-li proměnné různého charakteru (např. nominální a intervalové), musíme volit vždy koeficient pro proměnnou nižší úrovně
- Některé typy koeficientů lze použít jen při lineárním vztahu, jiné i pro vztahy nelineární
- Některé koeficienty rozlišují, která z proměnných je závislá a která nezávislá proměnná (asymetrické vztahy), jiné to nerozlišují (symetrické vztahy)
- Některé koeficienty rozlišují i velikost kontingenční tabulky
- Prokázání asociace není důkazem kauzality vztahu

Metody analýzy a korelační koeficienty

<i>Úroveň měření</i>	<i>Počet kategorií</i>	<i>Vhodná metoda</i>	<i>Vhodný koeficient</i>
1. Nominální / Nominální	2 x 2	Crosstabs	Phi, Lambda
2. Nominální / Nominální	3+ x 2+	Crosstabs	Cramerovo V, Lambda
3. Nominální / Ordinální	3+ x 3+	Crosstabs	Cramerovo V, Lambda
4. Nominální / Intervalová	nominální nezávislá	a) Crosstabs (pokud má intervalová proměnná málo kategorií b) Means, ANOVA	Eta Eta
5. Ordinální / Ordinální	obě proměnné s malým počtem kategorií	Crosstabs	Gamma, Kendalovo tau b (pro čtvercovou tabulku, Sommersovo D, Kendalovo tau c (pro obdélníkovou tabulku)
6. Ordinální / Ordinální	jedna proměnná s mnoha kategoriemi	pořadová korelace	Kendalovo tau c
7. Ordinální / Ordinální	obě proměnné s mnoha kategoriemi	pořadová korelace	Kendalovo tau c Spearmanovo ρ
8. Ordinální / Intervalová	obě proměnné s několika kategoriemi	a) Crosstabs b) Srovnání průměrů pokud je závisle proměnná intervalová	Eta, stejné koeficienty jak v 5. Eta
9. Ordinální / Intervalová	ordinální s několika kategoriemi, intervalová s mnoha	a) Means b) Pořadová korelace	Eta Kendalovo tau
10. Ordinální / Intervalová	obě s mnoha kategoriemi	pořadová korelace	Kendalovo tau Spearmanovo ρ
11. Intervalová / Intervalová		bodový graf	Pearsonovo R, Regrese

Podle Vaus, D. A. (1990): Surveys in Social Research. Unwin Hyman, London , p.182. In. Mareš, P. Rabušic, L. (2002): Měření (síly) asociace mezi dvěma spojitými proměnnými: korelační koeficienty a grafy.

Korelační analýza v SPSS – výběr korelačních koeficientů

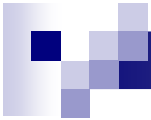
The image shows the SPSS Crosstabs dialog box and its Statistics sub-dialog box. The Crosstabs dialog has 'obyvatele' in the Row(s) field and 'prumyslove_podniky' in the Column(s) field. The Statistics sub-dialog is open, showing various statistical options.

Crosstabs Dialog:

- Row(s): obyvatele
- Column(s): prumyslove_podniky
- Buttons: Exact..., **Statistics...**, Cells..., Format..., Bootstrap...
- Layer 1 of 1: Previous, Next
- Display layer variables in table layers: ☒
- Display clustered bar charts: ☐
- Suppress tables: ☐
- Buttons: OK, Paste, Reset, Cancel, Help

Crosstabs: Statistics Sub-dialog:

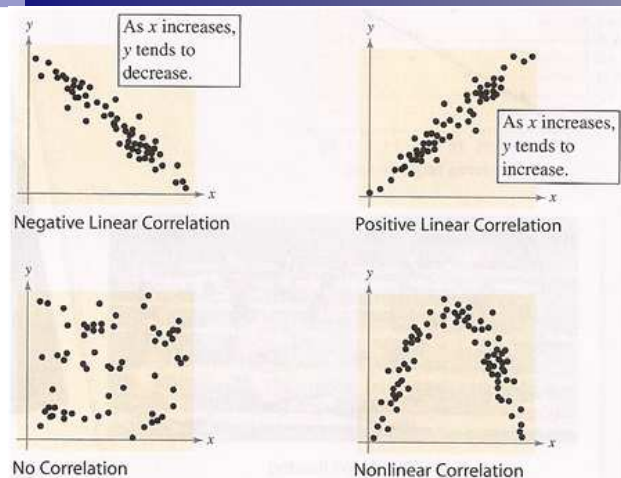
- ☐ Chi-square
- ☐ Correlations
- Nominal:**
 - ☐ Contingency coefficient
 - ☐ Phi and Cramer's V
 - ☐ Lambda
 - ☐ Uncertainty coefficient
- Ordinal:**
 - ☐ Gamma
 - ☐ Somers' d
 - ☐ Kendall's tau-b
 - ☐ Kendall's tau-c
- Nominal by Interval:**
 - ☐ Eta
- ☐ Kappa
- ☐ Risk
- ☐ McNemar
- ☐ Cochran's and Mantel-Haenszel statistics
 - Test common odds ratio equals: 1
- Buttons: Continue, Cancel, Help



Charakteristicky korelačních koeficientů

Koeficient	Velikost tabulky	Rozsah hodnot	Směr	Symetrický	Linearita
Phi	2 x 2	<0; 1>	ne	ano	ne
Cramerovo V	větší než 2 x 2	<0; 1>	ne	ano	ne
Lambda	jakákoliv velikost	<0; 1>	ne	obě verze	ne
Gamma	jakákoliv velikost	<-1; 1>	ano	ano	ano
Somersovo d	jakákoliv velikost	<-1; 1>	ano	obě verze	ano
Kendallovo tau b	čtvercové tabulky	<-1; 1>	ano	ano	ano
Kendallovo tau c	jakákoliv velikost	<-1; 1>	ano	ano	ano
Eta	jakákoliv velikost	<0; 1>	ne	ne	ne
Spearmanovo ρ	jakákoliv velikost	<-1; 1>	ano	ano	ano
Pearsonovo r	netabelovat	<-1; 1>	ano	ano	ano

Zdroj: Mareš, P. Rabušic, L. (2002): Měření (síly) asociace mezi dvěma spojitými proměnnými: korelační koeficienty a grafy.



Interpretace hodnot korelačního koeficientu v sociálních vědách (podle de Vaus, 2002)

Hodnota korelace	interpretace souvislosti
0,01 – 0,09	triviální, žádná
0,10 – 0,29	nízká až střední
0,30 – 0,49	střední až podstatná
0,50 – 0,69	podstatná až velmi silná
0,70 – 0,89	velmi silná
0,90 – 0,99	téměř perfektní



Statistická významnost (significance)

- Korelační koeficient vypovídá pouze o vztahu proměnných ve výběrovém souboru - cílem výzkumu je ale zobecnit (generalizovat) výsledky na základní soubor - to umožňuje test signifikance
- Výsledek testu statistické signifikance je závislý na velikosti souboru:
 - silné korelace lze získat spíše v malých souborech než ve velkých (malé soubory jsou více ovlivněny výběrovou chybou)
 - silné korelace budou často statisticky nevýznamné v malých souborech,
 - silná korelace v malém souboru může být statisticky nevýznamná, ale nízká korelace ve velkém souboru může být statisticky významná



Statisticky významný neznamená důležitý

- Statistická signifikance nemá nic společného s významností věcnou - i statisticky vysoce signifikantní výsledek může mít z věcného hlediska nulový význam
- Signifikance = jde o podmíněnou pravděpodobnost získání našich dat při platnosti nulové hypotézy v základním souboru a nikoliv o pravděpodobnost platnosti nulové hypotézy při existenci našich dat.
- ... neplatí tvrzení, že čím je hodnota statistické významnosti nižší (výsledek statisticky významnější), tím je výsledek důležitější nebo kvalitnější pro vědu či praxi.

Více viz:

Soukup, P. (2007): Statisticky významný neznamená důležitý. In. Socioweb 6/2007. Dostupné z:

<http://www.socioweb.cz/index.php?disp=teorie&shw=298&lst=108>

Soukup, P., Rabušic, L. (2007): Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti. *Sociologický časopis*, 43, 2: 379-395,



Příklad: jak souvisí ukazatele ekonomického potenciálu obce s velikostí obce a její periferií v rámci regionu

- Datový soubor: [Faktorovka_priklad.sav](#)
- Obsahuje data za obce JMK
(např. vzdálenost od regionální centra, počet obyvatel, míra nezaměstnanosti, bytová výstavba, index stáří, míra vzdělanosti, ekonomická aktivita, infrastruktura, daňový výtěžnost, atd.)
- Cílem korelační analýzy je zjistit, které proměnné mají statisticky významný vliv na vybrané ukazatele ekonomického potenciálu obce

Příklad použití funkce „Compare Means“

*BF_Principal component analýza.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform **Analyze** Direct Marketing Graphs Utilities Add-ons Window Help

Means

Závislé proměnné

Dependent List

code
vzdal_ORP
vzdal_Brno
zeleznice
limity
podil_zemedel
byty
obyvatele
hustota
přirustek
index_stari
vz

Nezávislé proměnné

Independent List

velikost
vzdalenost_Brno

Means: Options

Statistics:

Median
Grouped Median
Std. Error of Mean
Sum
Minimum
Maximum
Range
First
Last
Variance
Kurtosis
Std. Error of Kurtosis
Skewness
Std. Error of Skewness
Harmonic Mean

Cell Statistics:

Mean
Number of Cases
Standard Deviation

Statistics for First Layer

☒ Anova table and eta
☐ Test for linearity

Continue Cancel Help

	code	vzdal_ORP	vzdal_Brno	zeleznice	limity	podil_zemedel	byty	obyvatele	hustota	přirustek	index_stari	vz
1	582786,00	,00	,00	1,00	,00	61,80	15,00	378965,00	1646,24	2,03	136,50	
2	584801,00							4640,00	107,78	9,83	109,50	
3	581283,00							20769,00	461,84	2,31	128,40	
4	584649,00							7374,00	162,71	,22	101,00	
5	586021,00							25259,00	398,97	-8,43	120,30	
6	593711,00							34073,00	517,28	-2,50	114,60	
7	595098,00							860,00	64,08	-7,15	126,70	
8	594482,00											
9	581879,00											
10	582239,00											
11	584291,00											
12	586811,00											
13	582158,00											
14	583898,00											
15	584002,00											
16	586307,00											
17	586803,00											
18	593117,00											
19	593770,00											
20	595110,00	43,00	103,00	,00	,00	34,00						
21	581372,00	,00	39,00	1,00	,00	80,80						
22	546941,00	3,40	63,00	,00	,00	78,20						
23	583588,00	4,60	29,00	1,00	,00	84,60						
24	584550,00	18,00	34,00	,00	,00	92,00						
25	592927,00	11,00	45,00	1,00	,00	90,20						
26	593583,00	,00	24,00	1,00	,00	78,10						
27	581950,00	19,00	36,00	,00	,00	93,50						
28	582557,00	16,00	22,00	,00	,00	92,20						
29	582956,00	9,40	23,00	,00	,00	77,60						



Porovnání průměrných hodnot
- míry nezaměstnanosti
- ekonomicky aktivních v
terciéru
- míry podnikatelské aktivity a
- daňové výtěžnosti
u různě velkých kategorií obcí
(dle počtu obyvatel)

Report

velikost		nezamestannost	EA_tercier	podnik_activita	danova_vytěžnost
1,00	Mean	13,3091	43,2288	207,3894	9,2168
	N	66	66	66	66
	Std. Deviation	4,94084	9,43215	39,44846	2,16293
2,00	Mean	12,0487	45,1333	186,1718	8,6359
	N	39	39	39	39
	Std. Deviation	4,12405	5,83800	23,11458	2,09462
3,00	Mean	11,4758	50,2121	208,7727	8,6488
	N	33	33	33	33
	Std. Deviation	2,95540	6,26682	25,41779	1,17690
4,00	Mean	9,6333	54,4444	228,9333	9,5544
	N	9	9	9	9
	Std. Deviation	1,77482	6,26081	18,32069	,67125
5,00	Mean	10,5000	53,0750	223,0750	10,0675
	N	4	4	4	4
	Std. Deviation	2,01163	5,05198	7,01778	,36591
6,00	Mean	10,9600	61,4400	241,0800	10,6080
	N	5	5	5	5
	Std. Deviation	3,15246	7,20368	15,83010	,50415
7,00	Mean	9,2000	70,5000	330,1000	20,0200
	N	1	1	1	1
	Std. Deviation	.	.	.	.
Total	Mean	12,2274	46,8172	205,8987	9,1073
	N	157	157	157	157
	Std. Deviation	4,21876	8,98123	34,81914	2,07609

Hodnoty koeficientu ETA a statistická významnost vztahu mezi velikostí obce a indikátory ekonomického potenciálu

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
nezamestannost * velikost	Between Groups (Combined)	186,808	6	31,135	1,803	,102
	Within Groups	2589,665	150	17,264		
	Total	2776,472	156			
EA_tercier * velikost	Between Groups (Combined)	3651,005	6	608,501	10,219	,000
	Within Groups	8932,339	150	59,549		
	Total	12583,344	156			
podnik_aktivita * velikost	Between Groups (Combined)	43166,157	6	7194,360	7,393	,000
	Within Groups	145963,903	150	973,093		
	Total	189130,060	156			
danova_vyteznost * velikost	Between Groups (Combined)	152,231	6	25,372	7,317	,000
	Within Groups	520,155	150	3,468		
	Total	672,386	156			

Measures of Association

	Eta	Eta Squared
nezamestannost * velikost	,259	,067
EA_tercier * velikost	,539	,290
podnik_aktivita * velikost	,478	,228
danova_vyteznost * velikost	,476	,226

Příklad použití funkce „Crosstabs“

*BF_Principal component analýza.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

code vzdal_ORP vzdal_Brno zeleznice limity podil_zemedel byty obyvatel hustota přirustek index_stari

1	582786,00	,00	,00	1,00	,00	61,80	15,00	378965,00	1646,24	2,03	136,50
2	584801,00	,00	27,00	1,00	,00	73,20	20,25	1640,00	107,78	9,83	109,50
3	581283,00	,00	,00	,00	,00	,00	,00	69,00	461,84	2,31	128,40
4	584649,00	,00	,00	,00	,00	,00	,00	74,00	162,71	,22	101,00
5	586021,00	,00	,00	,00	,00	,00	,00	59,00	398,97	-8,43	120,30
6	593711,00	,00	,00	,00	,00	,00	,00	73,00	517,28	-2,50	114,60
7	595098,00	,00	,00	,00	,00	,00	,00	60,00	64,08	-7,15	126,70
8	594482,00	,00	,00	,00	,00	,00	,00	03,00	119,11	-1,99	119,60
9	581879,00	,00	,00	,00	,00	,00	,00	39,00	113,18	12,66	77,40
10	582239,00	,00	,00	,00	,00	,00	,00	45,00	232,76	4,64	111,00
11	584291,00	,00	,00	,00	,00	,00	,00	15,00	324,41	-3,03	112,40
12	586811,00	,00	,00	,00	,00	,00	,00	13,00	84,10	1,96	151,70
13	582158,00	,00	,00	,00	,00	,00	,00				
14	583898,00	,00	,00	,00	,00	,00	,00				
15	584002,00	,00	,00	,00	,00	,00	,00				
16	586307,00	,00	,00	,00	,00	,00	,00				
17	586803,00	,00	,00	,00	,00	,00	,00				
18	593117,00	,00	,00	,00	,00	,00	,00				
19	593770,00	,00	,00	,00	,00	,00	,00				
20	595110,00	,00	,00	,00	,00	,00	,00				
21	581372,00	,00	,00	,00	,00	,00	,00				
22	546941,00	3,40	63,00	,00	,00	78,20	1				
23	583588,00	4,60	29,00	1,00	,00	84,60	1				
24	584550,00	18,00	34,00	,00	,00	92,00	2				
25	592927,00	11,00	45,00	1,00	,00	90,20	1				
26	593583,00	,00	24,00	1,00	,00	78,10	2				
27	581950,00	19,00	36,00	,00	,00	93,50	1				
28	582557,00	16,00	22,00	,00	,00	92,20	1				
29	582956,00	9,40	23,00	,00	,00	77,60		8,61	2434,00	273,58	-9,53

Crosstabs

Row(s): vzdalenost_Brno

Column(s): nezamestnanost_kateg...

Layer 1 of 1

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK Paste Reset Cancel Help

Crosstabs: Statistics

Chi-square

Nominal

Contingency coefficient

Phi and Cramer's V

Lambda

Uncertainty coefficient

Ordinal

Gamma

Somers' d

Kendall's tau-b

Kendall's tau-c

Nominal by Interval

Eta

Kappa

Risk

McNemar

Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Continue Cancel Help

Vztah mezi kategorizovanou mírou nezaměstnanosti a vzdáleností obce od Brna – míra korelace

vzdalenost_Brno * nezamestnanost_kategorized Crosstabulation

			nezamestnanost_kategorized					Total
			1,00	2,00	3,00	4,00	5,00	
vzdalenost_Brno	1,00	Count	1	2	3	0	0	6
		% within nezamestnanost_kategorized	33,3%	11,1%	9,7%	0,0%	0,0%	3,8%
		% of Total	0,6%	1,3%	1,9%	0,0%	0,0%	3,8%
	2,00	Count	1	13	10	10	2	36
		% within nezamestnanost_kategorized	33,3%	72,2%	32,3%	27,8%	2,9%	22,9%
		% of Total	0,6%	8,3%	6,4%	6,4%	1,3%	22,9%
	3,00	Count	1	3	18	19	18	59
		% within nezamestnanost_kategorized	33,3%	16,7%	58,1%	52,8%	26,1%	37,6%
		% of Total	0,6%	1,9%	11,5%	12,1%	11,5%	37,6%
	4,00	Count	0	0	0	7	49	56
		% within nezamestnanost_kategorized	0,0%	0,0%	0,0%	19,4%	71,0%	35,7%
		% of Total	0,0%	0,0%	0,0%	4,5%	31,2%	35,7%
Total		Count	3	18	31	36	69	157

Directional Measures

			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	,647	,038	15,970	,000
		vzdalenost_Brno	,636	,039	15,970	,000
		nezamestnanost_kategorized	,658	,038	15,970	,000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.