Osnove statistike v geografiji 2

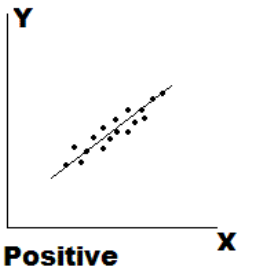
- Metodologija geografskega raziskovanja -

dr. Gregor Kovačič, doc.

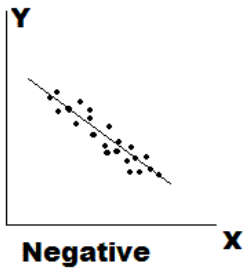
Bivariantna analiza

- Lastnosti so med sabo **odvisne** (vzročno-posledično povezane), kadar ena lastnost (spremenljivka X) vpliva na drugo (spremenljivka Y) x → y
- Lastnosti so korelacijsko(stohastično) **povezane**, kadar ni (neposredne) odvisnosti Y od X oz. je ne poznamo. x ←→ y; dani vrednosti x ne ustreza ena sama vrednost y temveč interval možnosti
- Obliko in jakost povezanosti med dvema spremenljivkama lahko ponazorimo:
  - Grafično (razsevni grafikon)
  - S števili, koeficienti (različno glede na tip spremenljivk)

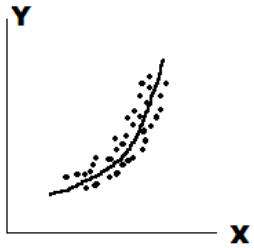
Razsevni grafikon ali korelacijski diagram



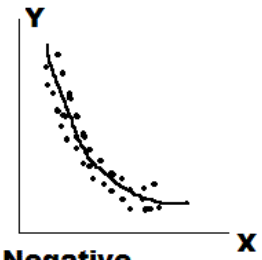
Positive linear correlation



Negative linear correlation

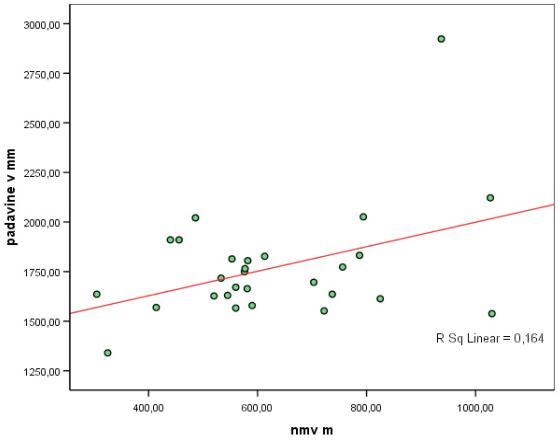


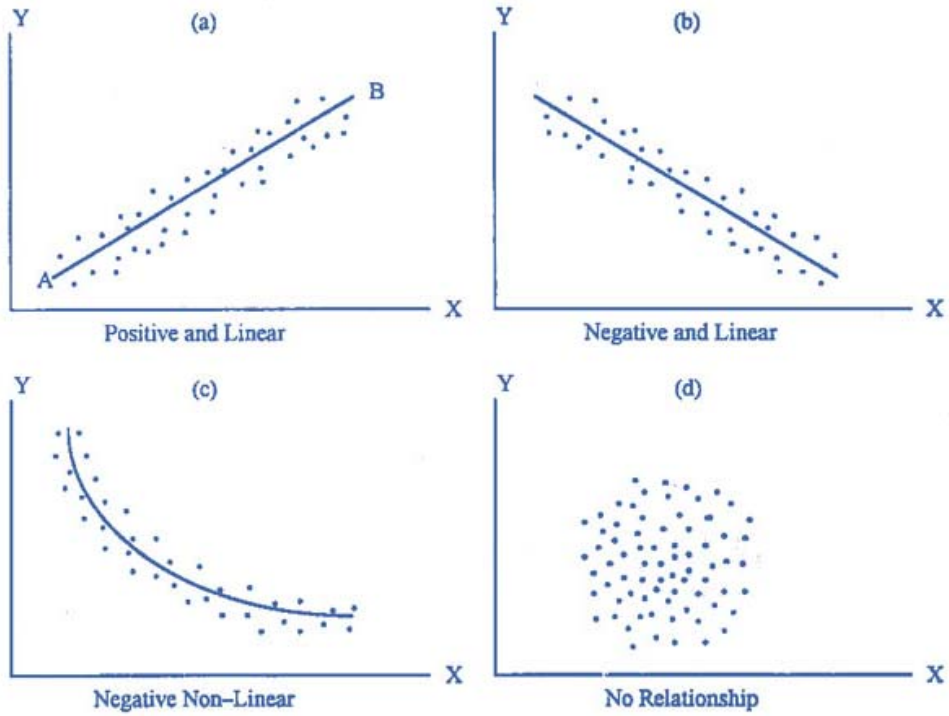
Positive non - linear correlation



Negative non - linear correlation

Razsevni diagram je grafični prikaz povezanosti med dvema spremenljivkama.
Regresijska črta kaže obliko povezanosti:
- Linearna
- Nelinearna (krivuljna)





Vrste korelacij

Glede na smer povezanosti:

- Pozitivna: z naraščanjem X narašča tudi Y
- Negativna: z naraščanjem X se vrednosti Y zmanjšujejo

Glede na obliko povezanosti:

- Linearna - premočrtna
- Nelinearna – krivuljna (več vrst, krivulje 2, 3 in nadaljnjega reda; MS Excel!)

Glede na jakost povezanosti: ni povezanosti, neznatna, šibka, …, močna (funkcijska) → glej naprej

Merjenje povezanosti

- Za merjenje povezanosti obstaja veliko koeficientov, ki so odvisni od nivoja merjenja (nominalni, ordinalni, številski) in tudi oblike povezanosti
  - Nominalni tip spremenljivk (ena od spremenljivk je nominalna) → $\chi 2$, kontingenčni koeficienti, koeficienti asociacije
  - Ordinalni tip para spremenljivk (ena spremenljivka je ordinalna, druga ordinalna ali boljša) → koeficient korelacije rangov (Spearmanov koeficient korelacije)
  - Številski (intervalne in/ali razmernostne) tip para spremenljivk (obe sta številski) → Pearsonov koeficient korelacije (linearno povezani številski spremenljivki)
  - Eta koeficient za nelinearno povezani številski spremenljivki (ali eno nominalno in eno številsko spremenljivko).

Družbena geografija - $\chi 2$, Spearmanov in Pearsonov koeficient

Fizična geografija – Pearsonov in Spearmanov koeficient, saj je večina spremenljivk kvantitativnih številskih in razmernostnih

HI kvadrat test (χ2)

- Samo za računanje s frekvencami
  - Kadar imamo opravka s podatki, ki niso normalno razporejeni
  - Ko preverjamo, ali dejanske frekvence odstopajo od pričakovanih pri določeni hipotezi
  - Včasih z njim ugotavljamo ali obstaja povezanost med spremenljivkama
- Pri korelaciji (ordinalne, razmernostne spremenljivke) → stopnja povezanosti
- χ^2 test → verjetnost povezanosti (nominalne spremenljivke)
- Zanima nas, ali so razlike v porazdelitvi med spremenljivkami statistično značilne in le učinek vzorca
- χ^2 test → sloni na primerjavi empiričnih (dejanskih) frekvenc s teoretičnimi frekvencami, to so frekvence, ki bi bile v kontingenčni preglednici, če spremenljivki ne bi bili povezani med seboj

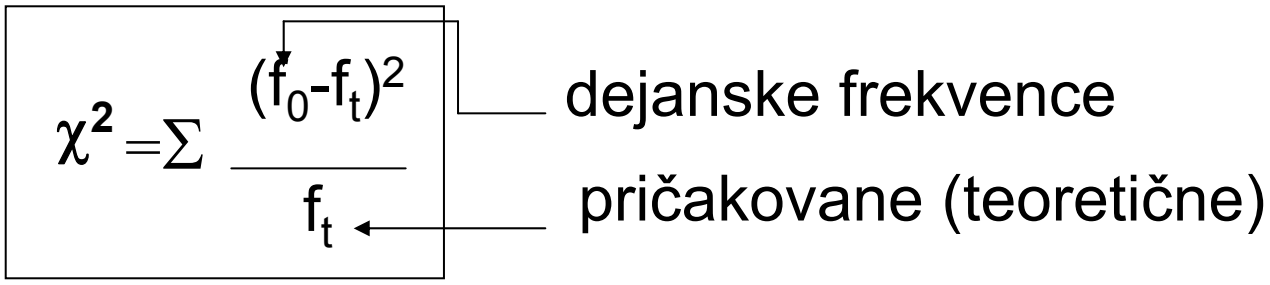
$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t}$$

f_0: dejanske frekvence

f_t: pričakovane (teoretične)

HI kvadrat test (χ2)

- Uporaba (ugotovitve)
  - Frekvence enega vzorca → ugotavljamo lahko, ali odstopajo od pričakovanih ob določeni hipotezi
  - Frekvence dveh neodvisnih vzorcev → ugotavljamo lahko, ali se razlikujeta v značilnostih
  - Dva odvisna vzorca → ugotavljamo lahko, ali je prišlo do spremembe, ali je razlika v različnih značilnostih



HI kvadrat test (χ2)

1. EN VZOREC – ena spremenljivka

- Ničelna hipoteza npr.: H_0: ni razlike v odgovorih

	DA	NE VEM	NE	SKUPAJ
f_0	26	12	10	48
f_t	16	16	16	48

stopnje svobode: število celic - 1=2

$$x^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$X^2 = \frac{(26-16)^2}{16} + \frac{(12-16)^2}{16} + \frac{(10-16)^2}{16} = 9{,}5$$

- $\chi 2 = 9{,}50$
- Če ne bi bilo razlik, bi bil x^2 enak 0
- Mejni $\chi 2$ pri 0,05 oz. 5 % stopnji tveganja pri stopnji svobode 2 je (glej preglednico mejnih vrednosti) enak 5,991 → s tem zavrnemo ničelno hipotezo
- Manjši $\chi 2$ pomeni večjo verjetnost, da je ničelna hipoteza resnična
- Če je $\chi 2$ manjši od stopnje svobode, je zanesljivo mogoče sprejeti ničelno hipotezo

Tabela mejnih vrednosti χ^2
Naslednja stran!

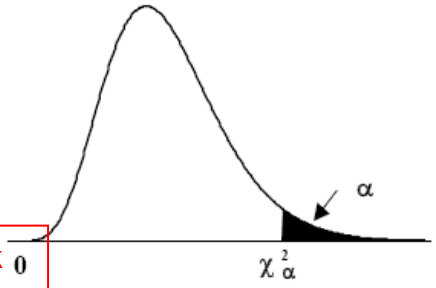
Preglednica mejnih vrednosti

- Preglednica ima na levi stopnje svobode (SP) v glavi pa verjetnost, da razlik med dejansko in teoretično razporeditvijo ni (to je delež pod krivuljo, ki je pobarvan črno – večji, ko je HI-kvadrat, manjša je površina pod krivuljo od njegove vrednosti na desno).
- Če smo pripravljeni zavrniti ničelno hipotezo s tveganjem, ki je 5 % ali manj, potem je pri 3 stopinjah svobode (SP = 3) to mogoče storiti, če je izračunani HI-kvadrat 7,81 ali več, če pa zmanjšamo tveganje na 1 % mora biti izračunani HI-kvadrat najmanj 11,34.
- V MS Excelu vam funkcija CHIINV izračuna mejno vrednost HI-kvadrata pri določenem tveganju in stopnji svobode in vam zato ni treba gledati v preglednico npr. CHIINV (0.95;4).
  - Probability – verjetnost = vstavimo želeno stopnjo tveganja (0-1)
  - Degrees freedom – vstavimo stopnjo svobode (prostostne stopnje), ki je odvisna od št. vrstic in stolpcev; SP = (št. vrstic – 1) x (št. stolpcev – 1)
- V MS Excelu vam funkcija CHITEST izračuna, kolikšno je tveganje in vam zato ni treba gledati v preglednico npr. CHITEST(B4:D6;B11:D13).
  - Izračuna vrednost tveganja pri izračunani vrednosti x^2 →
- CHIINV pa vam za to stopnjo tveganja (verjetnosti, da ničelna hipoteza drži) izračuna vrednost funkcije HI-kvadrat.
  - Izračunan x^2 bo enak šele pri tej stopnji tveganja
  - Če je izračunan x^2 pri večjih stopnjah tveganja večji od tistega v preglednici, zavrnemo ničelno hipotezo

TABELA 3: HI KVADRAT PORAZDELITEV

V tabeli je za verjetnost α in za stopinje prostosti **SP** navedena vrednost, za katero velja $P(X^2 \geq) = \alpha$.

Primer: $\alpha = 0{,}05$, SP = 1
$\chi^2_\alpha = 3{,}841$



Običajno 1 % ali 5 % tveganje razlik med dejansko in teoretično razporeditvijo

SP	α 0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
60	35,534	37,485	40,482	43,188	79,082	83,298	88,379	91,952
80	51,172	53,540	57,153	60,391	101,879	106,629	112,329	116,321
100	67,328	70,065	74,222	77,929	124,342	129,561	135,807	140,170

HI kvadrat test (χ2)

- DVA ALI VEČ NEODVISNIH VZORCEV

	-	+	skupaj
moški	$9_{(a)}$	$14_{(b)}$	$23_{(a+b)}$
ženske	$17_{(c)}$	$9_{(d)}$	$26_{(c+d)}$
skupaj	$26_{(a+c)}$	$23_{(b+d)}$	49

izračun teoretičnih frekvenc: (a+c)(a+b)/skupaj: 9 23/49=12,2

število stopenj svobode: (št. stolpcev - 1) x (št. vrstic - 1)

Preverjanje domnev pri HI-kvadrat testu

H_0 (ničelna hipoteza) : $\chi^2 = 0$ – spremenljivki nista povezani

- dejanske frekvence so enake teoretičnim

H_1 : $\chi^2 > 0$ – spremenljivki sta povezani

- iz preglednice porazdelitve $\chi 2$ razberemo kritično vrednost te statistike pri določeni stopnji značilnosti oz. tveganja (običajno 1 ali 5 %)

<u>Ničelno hipotezo lahko sprejmemo</u>

1. Če je izračunani $\chi 2$ manjši od stopenj svobode
2. Če je izračunani $\chi 2$ večji od kritične vrednosti, pomeni, da pade v kritično območje → ničelna domneva (hipoteza) je zavrnjena
    1. Pri (običajno 1 ali 5 %) stopnji značilnosti sprejmemo osnovno domnevo → spremenljivki sta statistično značilno povezani med seboj
3. Če smo pripravljeni sprejeti odločitev z višjo stopnjo tveganja, potem sprejmemo ničelno hipotezo, če je verjetnost, da razlik med dejansko in teoretično razporeditvijo frekvenc ni, med 1,00 in 0,05 ali izraženo v % med 100 in 5 %.
4. Če hočemo biti bolj prepričani o veljavnosti hipoteze, potem ničelno hipotezo sprejmemo, kadar je verjetnost, da razlik med teoretičnimi in dejanskimi frekvencami ni, večja od 0,01 oziroma 1 %.

Računanje z dvema ali več neodvisnima vzorcema

	Odgovor 1	Odgovor 2	Odgovor 3	Vsota
1. skupina	27	15	39	81
2. skupina	25	27	18	70
3. skupina	12	33	21	66
Vsota	64	75	78	217

	Odgovor 1	Odgovor 2	Odgovor 3	Vsota
1. skupina	23,8894	27,99539	29,11521	81
2. skupina	20,64516	24,19355	25,16129	70
3. skupina	19,46544	22,81106	23,7235	66
Vsota	64	75	78	217

Izračun po formuli: Σ((f0-ft)2)/ft			
	a) odgovor	b) odgovor	c) odgovor
1. skupina	0,405026	6,032429	3,355948
2. skupina	0,918599	0,325548	2,038213
3. skupina	2,863165	4,55106	0,312663

(81 x 64) / 217 = 23,88

$(27 - 23{,}88)^2 / 23{,}88$

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t}$$

$$X^2 = \frac{(27-23{,}88)^2}{23{,}88} + \frac{(15-27{,}99)^2}{27{,}99} + \frac{(39-29{,}11)^2}{29{,}11} + \ldots + \frac{(21-23{,}72)^2}{23{,}72} = 20{,}80$$

- H_0 - odgovori po skupinah niso med seboj povezani
- $\chi 2 = 20{,}80$
- Mejni $\chi 2$ pri 0,05 oz. 5 % stopnji tveganja pri stopnji svobode 4 je (glej preglednico mejnih vrednosti) enak 9,488
- Izračunani $\chi 2$ je večji od te vrednosti → s tem zavrnemo ničelno hipotezo; torej so povezani → spremenljivki sta statistično značilno povezani med seboj

število stopenj svobode: (št. stolpcev - 1) x (št. vrstic - 1)

HI kvadrat test (χ2)

- Lahko se računa le s frekvencami
- Vsota pričakovanih frekvenc mora biti enaka vsoti dejanskih
- Kadar imamo opravka z značilnostjo, ki se pojavlja ali pa ne, je vedno treba upoštevati tudi število primerov, ko se značilnost ne pojavi
- Vsaka frekvenca v posameznem polju mora pripadati drugi enoti
- Nobena pričakovana frekvenca ne sme biti premajhna
  - 20 % celic s $f_t < 5$ združevanje celic
  - Pri 2 x 2 preglednicah mora biti N > 40, sicer nobena celica ne sme imeti $f_t < 5$
  - Preglednice s številom stopenj svobode > 1: nobena celica ne sme imeti $f_t < 1$
- Pri številu stopenj svobode = 1 je treba izvesti Yatesovo korekcijo
- Statistika $\chi 2$ je lahko le pozitivna

Yatesov popravek

- Kadar delamo s preglednico 2 x 2 ali kadar imamo celice s frekvencami manjšimi od 5
- Pri številu stopenj svobode = 1 je treba izvesti Yatesov popravek

Popravek

- za 0,5 se zmanjša vsaka f_0, kjer velja $f_t < f_0$
- za 0,5 se poveča vsaka f_0, kjer velja $f_t > f_0$

Dejanska F	Dejanska F -Yp	Teoretična F
27	27,5	29
25	24,5	18
12	12,5	21
18	17,5	16

Literatura in viri

- Ferligoj, A., 1995: Osnove statistike na prosojnicah. Samozaložba z Batagelj.
- Statistika, 2007: Electronic statistic textbook, StatSoft. Dostopno na medmrežju.
- Rogerson, P. A., 2000: Statistical Methods for Geography. Sage Publications Inc.
- Kastelec. D. in K. Košmelj, 2010: Osnove statistike z Excelom 2007. Dostopno na medmrežju.