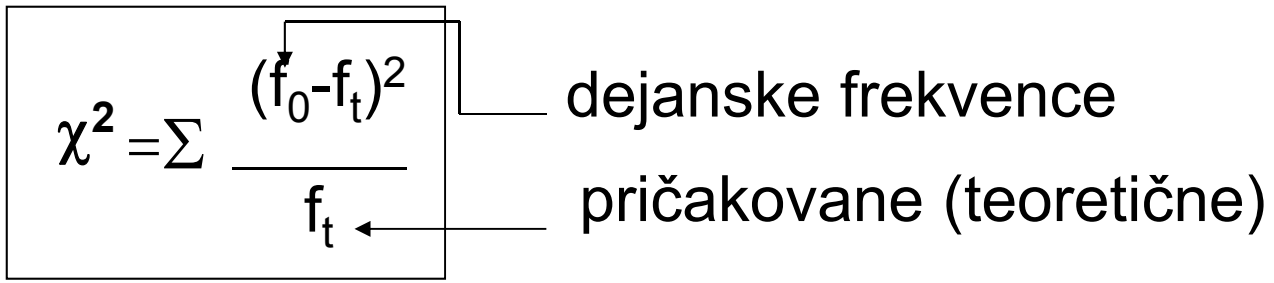
Hi-kvadrat preizkusi

- Kvantitativne metode v geografiji in uvod v GIS -

dr. Gregor Kovačič, doc.

Hi-kvadrat preizkus (χ2)

1. Preverjanje domneve o porazdelitvi. Frekvence enega vzorca → ugotavljamo lahko, ali odstopajo od pričakovanih ob določeni domnevi
    1. Lahko tudi primerjamo, ali dejanske frekvence odstopajo od pričakovanih frekvenc, ki so jih pokazale predhodne raziskave (npr. enaka raziskava, opravljena v različnih letih).
2. Preverjanje povezanosti dveh opisnih spremenljivk. Frekvence dveh neodvisnih vzorcev → ugotavljamo lahko, ali so razlike v porazdelitvi med spremenljivkami statistično značilne ali le učinek vzorca → najpogostejša raba
3. Preverjanje homogenosti struktur



Hi-kvadrat preizkus (χ2)

- Samo za računanje s frekvencami
- Primeren za imenske spremenljivke
- Kadar imamo opravka s podatki, ki niso normalno razporejeni
- χ^2 test → verjetnost povezanosti, ne dobimo pa stopnje povezanosti
- χ^2 test → sloni na primerjavi dejanskih (empiričnih) frekvenc s pričakovanimi (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 preizkus ($\chi 2$) – pregled osnovnih pojmov

- Dejanska frekvenca – število dejanskih enot (odgovorov) v skupini
- Pričakovana (teoretična) frekvenca – pričakovano število enot (odgovorov) v skupini.
  - Izračunamo s pomočjo vsote vseh dejanskih frekvenc (velikost vzorca) in vsot dejanskih frekvenc po vrsticah in stolpcih kontingenčne preglednice ali
  - S pomočjo vsote vseh dejanskih frekvenc in pričakovanih verjetnosti
- H_0 (ničelna domneva): $\chi^2 = 0$ – dejanske frekvence so enake pričakovanim (npr. spremenljivki nista povezani)
- H_1 (alternativna domneva): $\chi^2 > 0$ – dejanske frekvence so različne od pričakovanih (npr. spremenljivki sta povezani)
  - Iz preglednice porazdelitve $\chi 2$ razberemo kritično vrednost te statistike pri določeni stopnji značilnosti oz. tveganja (α; običajno 1 ali 5 %)
- α – stopnja značilnosti (stopnja tveganja): Je verjetnost napake, ki jo storimo v primeru, ko s statističnim sklepom ničelno domnevo zavrnemo v korist alternativne domneve. Njeno vrednost izberemo sami glede na to, kolikšno je za nas največje še sprejemljivo tveganje. Določena mora biti vnaprej, s tem je območje sprejema in zavrnitve ničelne domneve določeno pred izvedbo računskega postopka.
- SP – stopinje prostosti. Določimo s pomočjo preglednice $\chi 2$ porazdelitve ali s pomočjo funkcije CHIINV v MS Excelu. SP se vežejo na število celic in ne na velikost vzorca.
- *p*–vrednost: Je vezana na vzorec in izraža, v kolikšni meri so vzorčni podatki skladni z ničelno domnevo. Glej naslednji slide.

p - vrednost

- *p*–vrednost je vezana na vzorec in izraža, v kolikšni meri so vzorčni podatki skladni z ničelno domnevo
- Večja *p*–vrednost pomeni večjo podporo ničelni domnevi
- Majhna *p*–vrednost govori v prid alternativne domneve
- Če velja, da je *p*–vrednost manjša od predpisane vrednosti za α, ničelno domnevo zavrnemo
- Če velja, da je *p*–vrednost večja od predpisane vrednosti za α, ničelno domnevo obdržimo
- Računalniški programi (tudi MS Excel) izračunajo točne *p*–vrednosti; sicer se jih lahko odčita iz statističnih preglednic
    - Za izračun *p*–vrednosti pri Hi-kvadrat preizkusu se uporabljata funkciji CHISQ.TEST ali pa CHISQ.DIST.RT
- *p*–vrednost uporabimo pri vsebinskem sklepu

Preverjanje domnev pri Hi-kvadrat preizkusu

Ničelno domnevo lahko sprejmemo

1. Če je izračunani $\chi 2$ manjši od stopinj prostosti. Manjši $\chi 2$ pomeni večjo verjetnost, da je ničelna domneva resnična.
2. Če je izračunani $\chi 2$ večji od kritične vrednosti pri določenem α, pomeni, da pade v kritično območje → ničelna domneva je zavrnjena.
3. Če je $\chi 2$ manjši od stopinje prostosti, je zanesljivo mogoče sprejeti ničelno domnevo.
4. Če smo pripravljeni sprejeti odločitev z višjo stopnjo tveganja (α), potem sprejmemo ničelno domnevo, če je verjetnost, da razlik med dejansko in pričakovano razporeditvijo frekvenc ni, med 1 (100 %) in 0,05 (5 %).

1. Preverjanje domneve o porazdelitvi

- Preverjamo domnevo o verjetnostni porazdelitvi
- Preverjamo dejanske verjetnosti s pričakovanimi (teoretičnimi) verjetnostmi

PRIMER 1

- Ali je obisk Škocjanskega zatoka enakomerno razporejen po letnih časih?
- Če da, potem bi veljalo, da je v vsakem letnem času 1/4 letnega obiska. S Hi-kvadrat preizkusom lahko preverimo to domnevo.
- H_0 : porazdelitev obiska je po letnih časih enakomerna, torej verjetnost obiska v posameznem letnem času = 1/4.
- H_1 : H_0 ne velja - porazdelitev obiska po letnih časih ni enakomerna

PRIMER 2 (tega ne bomo računali!)

- Ali je v populaciji študentov UP FHŠ povprečna ocena porazdeljena po normalni porazdelitvi s povprečno vrednostjo 7,8 in standardnim odklonom 0,2, torej PO ~ (7,8;0,2)?
- Če da, potem bi, skladno z gostotami verjetnosti za normalno porazdelitev (Gaussova krivulja), veljalo, da ima:
  - 68,27 % študentov povprečno oceno v razponu 7,6–8,
  - 95,5 % študentov povprečno oceno v razponu 7,4–8,2 in
  - 99,73% študentov povprečno oceno v razponu 7,2–8,4.
- H_0 : porazdelitev je normalna
- H_1 : H_0 ne velja

1. Preverjanje domneve o porazdelitvi – primer 1

- Ničelna domneva - H_0: porazdelitev obiska po letnih časih je enakomerna pri α = 0,05
- Alternativna domneva - H_1: ničelna domneva ne drži, porazdelitev obiskov po letnih časih ni enakomerna

stopinje prostosti: število celic – 1 = 2

	Zima	Pomlad	Poletje	Jesen	SKUPAJ
f_0 (dejansko število obiskovalcev) – dejanska frekvenca	26	12	14	8	60
f_t (pričakovano število obiskovalcev) – teoretična frekvenca	15	15	15	15	60

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

$$X^2 = \frac{(26-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(8-15)^2}{15} = 9{,}33$$

Tabela mejnih vrednosti χ^2 Naslednja stran!

REŠITEV

- $\chi^2 = 9{,}33$
- Če ne bi bilo razlik, bi bil x^2 enak 0
- Mejni χ^2 pri 0,05 oz. 5 % stopnji tveganja pri stopinji prostosti 2 je (glej preglednico mejnih vrednosti) enak 5,991
- Izračunani χ^2 večji od teoretičnega → zato zavrnemo ničelno domnevo.

Statistični sklep:

- Ničelno domnevo zavrnemo pri α = 0,05. Porazdelitev obiska v Škocjanskem zatoku se po letnih časih statistično značilno razlikuje.

Vsebinski sklep:

- S tveganjem, manjšim od α = 0,05, trdimo, da se obisk v Škocjanskem zatoku po letnih časih razlikuje (p = 0,0073).

Interpretacija:

- Podatki nakazujejo, da je pozimi veliko več obiska, kot bi pričakovali pri enakomerni porazdelitvi. Prispevek k χ^2 statistiki je namreč največji pri zimi.

Preglednica mejnih vrednosti

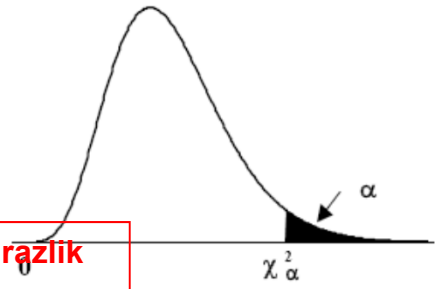
- Preglednica ima v prvem stolpcu na levi stopinje prostosti (SP), v glavi pa verjetnost, da razlik med dejansko in teoretično razporeditvijo ni (to je delež pod krivuljo, ki je pobarvan črno.
- Večji, ko je Hi-kvadrat, manjša je površina pod krivuljo od njegove vrednosti na desno).
- Če smo pripravljeni zavrniti ničelno domnevo s tveganjem, ki je 5 % ali manj, potem je pri 3 stopinje prostosti 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.

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

	α							
SP	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

Uporaba funkcij v MS Excelu pri Hi-kvadrat preizkusu

Funkcija CHIINV (probability; deg_freedom)

- Vrne vrednost Hi-kvadrat statistike pri določenem tveganju in stopinji prostosti
- Tako ni potrebno za vrednosti gledati v preglednico Hi-kvadrat porazdelitve
- Ima dva argumenta:
  - Probability – verjetnost = vstavimo želeno stopnjo tveganja (0–1)
  - Deg_freedom – vstavimo stopinjo prostosti, ki je odvisna od št. vrstic in stolpcev; SP = (št. vrstic – 1) x (št. stolpcev – 1)
- Npr. CHIINV (0,05;4).

Funkcija CHISQ.DIST.RT (x; deg_freedom)

- Vrne *p*–vrednost za izračunano Hi-kvadrat statistiko
- Ima dva argumenta:
  - x – vrednost Hi-kvadrat statistike, za katero funkcija vrne *p*–vrednost (verjetnost)
  - Deg_freedom – vstavimo stopinjo prostosti, ki je odvisna od št. vrstic in stolpcev; SP = (št. vrstic – 1) x (št. stolpcev – 1)

Funkcija CHISQ.TEST (actual_range; expected range)

- Vrne *p*–vrednost Hi-kvadrat preizkusa; torej izračuna tveganje, da alternativna domneva ne velja
- Prvi argument je obseg celic z dejanskimi frekvencami in drugi argument je obseg celic s pričakovanimi frekvencami
- CHISQ.TEST izračuna, kolikšno je tveganje in vam zato ni treba gledati v preglednico
- Npr. CHISQ.TEST(B4:D6;B11:D13)
- CHIINV pa vam za to stopnjo tveganja (verjetnosti, da ničelna domneva drži) izračuna vrednost funkcije Hi-kvadrat.
  - Izračunan x^2 bo enak šele pri tej stopnji tveganja

2. Preverjanje domneve o povezanosti dveh opisnih spremenljivk

- Preverjamo, ali sta spremenljivki povezani ali ne.
- Preverjamo dejanske frekvence s pričakovanimi frekvencami

PRIMER 1

- Preveriti želimo, ali sta spremenljivki "osebna predstava o TR" in "domicilnost anketiranih študentov po fakultetah UP" povezani pri $\alpha = 0{,}01$.
- H_0 : spremenljivki nista povezani
- H_1 : spremenljivki sta povezani

PRIMER 2

- Preveriti želimo, ali sta spremenljivki „dosežena stopnja izobrazbe“ in „velikost naselja stalnega prebivališča“ povezani pri $\alpha = 0{,}05$.
- H_0 : spremenljivki nista povezani
- H_1 : spremenljivki sta povezani

2. Preverjanje domneve o povezanosti dveh opisnih spremenljivk

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

	Bela	Siva	Črna	Vsota
Trda	27	15	39	81
Lomljiva	25	27	18	70
Mehka	12	33	21	66
Vsota	64	75	78	217

	Bela	Siva	Črna	Vsota
Trda	23,8894	27,99539	29,11521	81
Lomljiva	20,64516	24,19355	25,16129	70
Mehka	19,46544	22,81106	23,7235	66
Vsota	64	75	78	217

Izračun po formuli: Σ((f0-ft)2)/ft			
	Bela	Siva	Črna
Trda	0,405026	6,032429	3,355948
Lomljiva	0,918599	0,325548	2,038213
Mehka	2,863165	4,55106	0,312663

(81 x 64) / 217 = 23,88

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

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

- H_0 - spremenljivki „barva kamnine“ in „trdnost kamnine“ nista povezani pri α = 0,05
- H_1 – spremenljivki sta povezani
- $\chi 2 = 20{,}80$
- Mejni $\chi 2$ pri 0,05 oz. 5 % stopnji tveganja pri stopinji prostosti 4 je (glej preglednico mejnih vrednosti) enak 9,488
- Izračunani $\chi 2$ je večji od te vrednosti → s tem zavrnemo ničelno domnevo; spremenljivki sta statistično značilno povezani med seboj

Statistični sklep:

- Ničelno domnevo zavrnemo pri α = 0,05. Barva kamnin in njihova trdnost sta povezani značilnosti.

Vsebinski sklep:

- S tveganjem, manjšim od α = 0,05, trdimo, da sta barva kamnine in njena trdnost povezani spremenljivki (*p* = 0,00034).

Interpretacija:

- Podatki nakazujejo, da je več kamnin sive barve, ki so obenem trdne, kot bi pričakovali, če povezave med barvo in trdnostjo kamnine ne bi bilo. Prispevek te celice k $\chi 2$ statistiki je namreč največji.

stopinje svobode: (št. stolpcev - 1) x (št. vrstic - 1)

3. Preverjanje homogenosti struktur

- Pri določeni stopnji značilnosti (α) preverjamo enakost struktur
- Preverjamo dejanske frekvence s pričakovanimi frekvencami

PRIMER 1

- Preveriti želimo, ali je izobrazbena struktura po spolu enaka pri α = 0,01.
- H_0 : izobrazbena struktura po spolu je enaka
- H_1 : izobrazbena struktura po spolu ni enaka

PRIMER 2

- Preveriti želimo, ali je struktura zaposlenosti med statističnimi regijami enaka pri α = 0,05.
- H_0 : struktura zaposlenosti je po statističnih regijah enaka
- H_1 : struktura zaposlenosti po statističnih regijah ni enaka

3. Preverjanje homogenosti struktur

• Ničelna domneva - H_0: struktura po stolpcih je enaka; višina plače po spolu je enaka pri α = 0,01
• Alternativna domneva - H_1: struktura po stolpcih ni enaka; višina plače po spolu ni enaka

	Ženske	Moški	Vsota
Do 800 evr	22	18	40
801-1000	70	80	150
Nad 1000	52	63	115
Vsota	144	161	305

	Ženske	Moški	Vsota
Do 800 evr	18,88	21,11	40
801-1000	70,81	79,18	150
Nad 1000	54,29	60,7	115
Vsota	144	161	305

	Ženske	Moški
Do 800 evr	0,51	0,45
801-1000	0,009	0,008
Nad 1000	0,096	0,087

REŠITEV

- $\chi 2 = 1{,}173$
- Mejni $\chi 2$ pri 0,01 oz. 1% stopnji tveganja pri stopinji prostosti 2 je (glej preglednico mejnih vrednosti) enak 9,210
- Izračunani $\chi 2$ manjši od teoretičnega → zato ničelno domnevo obdržimo.

Statistični sklep:

- Ničelno domnevo obdržimo pri α = 0,01. Višina plače se po spolu ne razlikuje.

Vsebinski sklep:

- Tveganje, da se višina plače po spolu razlikuje, je zelo veliko (*p* = 0,5561).

Hi-kvadrat preizkus ($\chi 2$)

- 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 stopinj svobode > 1: nobena celica ne sme imeti $f_t < 1$
- Pri številu stopinj svobode = 1 je treba izvesti Yatesov popravek
- 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 stopinj 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, Anuška. 1995. *Osnove statistike na prosojnicah.* Ljubljana: Samozaložba Z. Batagelj.
- Statistika. 2013. »*Electronic statistic textbook, StatSoft* «. Http://www.statsoft.com.
- Rogerson, Peter A. 2006. *Statistical Methods for Geography: a student guide*. London: Sage Publications.
- Kastelec, Damijana in Katarina Košmelj, 2010. *Osnove statistike z Excelom 2007.* Ljubljana: Biotehniška fakulteta. - Dostopno tudi na medmrežju.