

Procedura za Hi kvadrat analizu

Aleksandar Stojanović , 147/2010

Marija Milićević , 22/2010

(Pearson-ov) Hi kvadrat test

- Neparametarski test
- Najpoznatiji su test frekvencija i test nezavisnosti

Hi kvadrat – test frekvencija

- The goodness of fit test

Analyze -> Nonparametric Tests -> Legacy Dialogs ->
Chi – square...

Hi kvadrat– test nezavisnosti

- Ispitivanje da li su dva obeležja X i Y nezavisna.
- Hipotezu o nezavisnosti $H_0(F_{X,Y} = F_X F_Y)$ testiramo protiv alternativne hipoteze $H_1(F_{X,Y} \neq F_X F_Y)$.
- Odgovor treba dati na osnovu dvodimenzionalnog prostog uzorka obima $n : (X_i, Y_i), i = 1, 2, \dots, n$.
- Podaci se daju u tabeli kontigencije (tabeli povezanosti).

Tabela kontigencije

$X \backslash Y$	y_1	y_2	...	y_s	zbir
x_1	n_{11}	n_{12}	...	n_{1s}	$n(x_1)$
x_2	n_{21}	n_{22}	...	n_{2s}	$n(x_2)$
...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n(x_r)$
zbir	$n(y_1)$	$n(y_2)$	...	$n(y_s)$	n

Podaci su raspodeljeni u r kategorija obeležja X i u s kategorija po vrednostima obeležja Y .

Broj n_{ij} označava broj pojavljivanja para (x_i, y_j) u uzorku.

- $n(x_1), \dots, n(x_r)$ - marginalni zbirovi po vrstama obeležja X
- $n(y_1), \dots, n(y_s)$ -marginalni zbirovi po kolonama obeležja Y
- Test statistika:

$$T = \sum_{i=1}^r \sum_{j=1}^s \frac{(n \cdot n_{ij} - n(x_i) \cdot n(y_j))^2}{n \cdot n(x_i) \cdot n(y_j)}$$

- Ako je tačna nulta hipoteza, test-statistika ima približno $\chi^2_{(r-1)(s-1)}$ raspodelu.
- Za dati prag značajnosti α , pomoću tablice za Hi kvadrat raspodelu, dobija se c tako da važi $P\{T \geq c\} = \alpha$.
- Hipoteza o nezavisnosti se odbacuje ako je realizovana vrednost test statistike veća od c , dok se u suprotnom prihvata.

- **Napomena 1:** U Pearson-ovom Hi kvadrat testu (i svim njegovim varijantama) obim uzorka treba da bude veliki, bar toliko da se obezbedi da očekivane frekvencije u svim klasama budu veće od 5.
- **Napomena 2:** Ovakva vrsta analize se vrši kada su promenljive kategoričkog tipa.

SPSS implementacija

Analyze -> Descriptive Statistics -> Crosstabs...

63-baza.SAV [DataSet1] - IBM SPSS Statistics Data Editor

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

10 : pol 1

Visible: 11 of 11 Variables

id pplata

id	pplata	kat_posla	rasa	pol_rasa	starost	r_staz	var	var	var	var	var
1	748	408	1	0	3	63,83	22,92				
2	832	408	1	0	3	55,25	3,58				
3	754	390	1	0	3	55,50	,00				
4	869	408	1	0	3	60,00	,00				
5	969	408	1	0	3	62,33	15,00				
6	825	408	6540	12	1	1	4	60,50	13,58		
7	1083	408	6600	12	1	0	3	62,42	24,00		
8	1107	390	6660	8	1	0	3	62,50	34,33		
9	886	408	6720	8	1	0	3	59,08	6,33		
10	1127	408	6780	8	1	0	3	56,92	26,58		
11	921	360	6780	12	1	1	4	60,67	10,33		
12	935	408	6780	12	1	1	4	51,50	22,58		
13	831	408	6840	12	1	0	3	55,08	6,00		
14	940	408	6840	8	1	1	4	51,50	,00		
15	1128	408	6900	12	1	0	3	53,92	29,83		
16	749	408	6960	8	1	0	3	57,83	32,50		
17	647	408	6960	12	1	1	4	46,50	9,67		
18	826	408	7080	8	1	0	3	61,50	15,33		
19	995	390	7260	12	1	0	3	62,00	6,00		
20	1121	408	7380	12	1	1	4	51,00	19,00		
21	1034	420	7500	15	1	0	3	58,00	4,50		
22	1096	3600	7680	15	1	1	4	60,50	1,92		
23	784	5220	7860	8	1	0	3	55,92	8,50		

Data View Variable View

Crosstabs...

IBM SPSS Statistics Processor is ready

15:03 30.5.2015.

Otvara se prozor za odabir dva obeležja čiju nezavisnost ispituјemo. Opcije u polјima *Statistics* i *Cells* su značajne u Hi kvadrat analizi.

63-baza.SAV [DataSet1] - IBM SPSS Statistics Data Editor

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

10 : pol 1 Visible: 11 of 11 Variables

	id	pplata	pol	vreme	splata	obrazovanje	kat_posla	rasa	pol_rasa	starost	r_staz	var	var	var	var	var
1	748	4080	1	70												
2	832	4080	1	74												
3	754	3900	1	92												
4	869	4080	1	82												
5	969	4080	1	68												
6	825	4080	1	66												
7	1083	4080	1	84												
8	1107	3900	1	88												
9	886	4080	1	76												
10	1127	4080	1	72												
11	921	3600	1	97												
12	935	4080	1	72												
13	831	4080	1	85												
14	940	4080	1	81												
15	1128	4080	1	82												
16	749	4080	1	81												
17	647	4080	1	72												
18	826	4080	1	72												
19	995	3900	1	86												
20	1121	4080	1	85												
21	1034	4200	1	90	7500	15	1	0	3	58,00	4,50					
22	1096	3600	1	96	7680	15	1	1	4	60,50	1,92					
23	784	5220	1	70	7860	8	1	0	3	55,92	8,50					

Crosstabs

Row(s):

Column(s):

Layer 1 of 1

Previous Next

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK Paste Reset Cancel Help

Exact... Statistics... Cells... Format... Bootstrap...

Data View Variable View

IBM SPSS Statistics Processor is ready

ENG 15:13 30.5.2015.

Statistics

63-baza.SAV [DataSet1] - IBM SPSS Statistics Data Editor

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

10: pol 1 Visible: 11 of 11 Variables

	id	pplata	pol	vreme	splata	obrazovanje	kat_posla	rasa	pol_rasa	starost	r_staz	var	var	var	var	var
1	748	4080	1	70												
2	832	4080	1	74												
3	754	3900	1	92												
4	869	4080	1	82												
5	969	4080	1	68												
6	825	4080	1	66												
7	1083	4080	1	84												
8	1107	3900	1	88												
9	886	4080	1	76												
10	1127	4080	1	72												
11	921	3600	1	97												
12	935	4080	1	72												
13	831	4080	1	85												
14	940	4080	1	81												
15	1128	4080	1	82												
16	749	4080	1	81												
17	647	4080	1	72												
18	826	4080	1	72												
19	995	3900	1	86												
20	1121	4080	1	85												
21	1034	4200	1	90	7500	15	1	0	3	58,00	4,50					
22	1096	3600	1	96	7680	15	1	1	4	60,50	1,92					
23	784	5220	1	70	7860	8	1	0	3	55,92	8,50					

Crosstabs

Crosstabs: Statistics

- ☒ Chi-square
- ☐ Correlations
- ☐ Contingency coefficient
- ☐ Phi and Cramer's V
- ☐ Lambda
- ☐ Uncertainty coefficient
- ☐ Gamma
- ☐ Somers' d
- ☐ Kendall's tau-b
- ☐ Kendall's tau-c
- ☐ Kappa
- ☐ RISK
- ☐ McNemar
- ☐ Cochran's and Mantel-Haenszel statistics
- Test common odds ratio equals: 1

Continue Cancel Help

OK Paste Reset Cancel Help

IBM SPSS Statistics Processor is ready

15:34 30.5.2015.

Statistics

- **Chi-square.** Za tabele 2×2 , ovo polje računa statistike Hi kvadrat testa, testa količnika verodostojnosti, Fisher-ov test tačnosti verovatnoće i Yates-ovu korekciju Hi kvadrat testa (korekcija neprekidnosti). Za tabele proizvoljnih dimenzija, ovo polje računa statistike Pearson-ovog Hi kvadrat testa i testa količnika verodostojnosti. Dodatno, kada su oba obeležja kvantitativnog tipa, računa se linearna zavisnost.
- **Correlations.** Ukoliko su oba obeležja ordinalnog tipa, računa Spearman-ov korelacioni koeficijent, ρ , koji je mera zavisnosti rangova. Kada su oba obeležja kvantitativnog tipa, računa se Pearson-ov korelacioni koeficijent, r , koji je mera linearne zavisnosti između obeležja.

- **Nominal.** Za obeležja nominalnog tipa.

- **Contingency coefficient.** Vrednost između 0 i 1, gde 0 znači da nema zavisnosti a vrednosti blizu 1 označavaju jaku zavisnost između obeležja. Maksimalna i minimalna moguća vrednost zavise od broja vrsta i kolona u tabeli.
- **Phi and Cramer's V.** *Phi* je broj koji se dobija kao koren količnika test statistike χ^2 kvadrat testa i obima uzorka (2×2 tablice). Cramer-ov broj *V* je mera zavisnosti zasnovana na χ^2 kvadrat testu, ista kao *Phi* ali se koristi za nekvadratne tablice.
- **Lambda.** Mera zavisnosti koja odražava proporcionalno smanjenje greške kada su vrednosti nezavisne promenljive upotrebljene za predviđanje vrednosti zavisne promenljive. Vrednost 1 znači da nezavisna promenljiva savršeno određuje zavisnu promenljivu, dok vrednost 0 znači da nezavisna promenljiva nema ulogu u određivanju zavisne promenljive.
- **Uncertainty coefficient.** Slična mere kao Lambda. Opisuje proporcionalno smanjenje greške kada su vrednosti jedne promenljive upotrebljene za predviđanje vrednosti druge promenljive. Npr. vrednost 0.83 znači da to što nam je poznata jedna promenljiva smanjuje grešku određivanja druge za 83%.

- **Ordinal.**

- **Gamma.** Simetrična mera zavisnosti koja uzima vrednosti između -1 i 1 . Ukoliko je apsolutna vrednost blizu 1 to ukazuje na jaku zavisnost, dok 0 ukazuje na slabu ili nikakvu.
- **Sommers' d.** Asimetrično proširenje mere Gamma. Uključuje i broj onih parova koji nemaju jednake vrednosti nezavisne promenljive.
- **Kendall's tau-b.** Neparametarska mera korelacije promenljivih koje su ordinalnog tipa. Uzima u obzir iste parove podataka. Znak određuje da li je pozitivna ili negativna, a apsolutna vrednost određuje jačinu korelacije. Vrednosti su između -1 i 1 , dok se tačne vrednosti -1 i 1 mogu dobiti samo za kvadratne tablice.
- **Kendall's tau-c.** Slično prethodnoj meri. Neparametarska mera zavisnosti promenljivih koje su ordinalnog tipa. U ovom slučaju se isti parovi podataka ne uzimaju u obzir.

- **Nominal by interval.** Kada je jedna promenljiva kategoričkog tipa a druga intervalnog, čekiramo *Eta*. Kategorička promenljiva mora biti kodirana numerički.
 - **Eta.** Mera zavisnosti. Uzima vrednosti između 0 i 1, gde 0 znači da nema zavisnosti između kolone i vrste, a vrednost bliska jedinici označava jaku zavisnost. Eta je pogodno koristiti kada imamo zavisnu promenljivu intervalnog tipa i nezavisnu promenljivu sa ograničenim brojem kategorija. Računaju se dve Eta vrednosti: prva tretira promenljive u redu kao intervalne, a druga kolone tretira kao intervalne promenljive.
- **Cochran's and Mantel-Haenszel statistics.** Ove dve statistike se koriste za ispitivanje nezavisnosti promenljive predstavljene dihotomnim faktorima i odgovarajuće dihotomne zavisne promenljive.

- **Kappa.** *Cohen-ov kappa* meri slaganje između dve procene istog objekta. Vrednost 1 ukazuje da je slaganje odlično, dok vrednost 0 ukazuje da slaganje nije ništa bolje od puke slučajnosti. Kappa je zasnovan na kvadratnim tablicama u kojoj su vrednosti u redovima i kolonama predstavljene istom skalom.
- **Risk.** Koristi se za tabele 2×2 i meri jačinu zavisnosti prisustva faktora i pojave nekog događaja. Ukoliko interval poverenja sadrži 1 ne možemo zaključiti da je faktor povezan sa događajem. Odnos kvota se može koristiti kao ocena relativnog rizika kad se faktor pojavljuje retko.
- **McNemar.** Neparametarski test za dve zavisne dihotomne promenljive. Testira promene u odzivu koristeći Hi kvadrat test. Koristan u određivanju promena u eksperimentima „pre i posle“ (takva vrsta eksperimenata se često koristi u medicini).

Cells

63-baza.SAV [DataSet1] - IBM SPSS Statistics Data Editor

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

10 : pol 1 Visible: 11 of 11 Variables

	id	pplata	pol	vreme	splata	obrazovanje	kat_posla	rasa	pol_rasa	starost	r_staz	var	var	var	var	var
1	748	4080	1	70												
2	832	4080	1	74												
3	754	3900	1	92												
4	869	4080	1	82												
5	969	4080	1	68												
6	825	4080	1	66												
7	1083	4080	1	84												
8	1107	3900	1	88												
9	886	4080	1	76												
10	1127	4080	1	72												
11	921	3600	1	97												
12	935	4080	1	72												
13	831	4080	1	85												
14	940	4080	1	81												
15	1128	4080	1	82												
16	749	4080	1	81												
17	647	4080	1	72												
18	826	4080	1	72												
19	995	3900	1	86												
20	1121	4080	1	85												
21	1034	4200	1	90	7500	15	1	0	3	58,00	4,50					
22	1096	3600	1	96	7680	15	1	1	4	60,50	1,92					
23	784	5220	1	70	7860	8	1	0	3	55,92	8,50					

Crosstabs: Cell Display

- Counts**
 - ☒ Observed
 - ☐ Expected
 - ☐ Hide small counts (Less than 5)
- z-test**
 - ☐ Compare column proportions
 - ☐ Adjust p-values (Bonferroni method)
- Percentages**
 - ☐ Row
 - ☐ Column
 - ☐ Total
- Residuals**
 - ☐ Unstandardized
 - ☐ Standardized
 - ☐ Adjusted standardized
- Noninteger Weights**
 - ☒ Round cell counts
 - ☐ Round case weights
 - ☐ Truncate cell counts
 - ☐ Truncate case weights
 - ☐ No adjustments

Continue Cancel Help

Data View Variable View

IBM SPSS Statistics Processor is ready

16:15 30.5.2015.

Cells

- **Counts.** Opcija *Observed* ispisuje broj slučajeva iz uzorka, dok opcija *Expected* ispisuje očekivani broj slučajeva uz pretpostavku da su redovi i kolone nezavisni. Možemo izabrati i opciju *Hide small counts* ako želimo da ne prikažemo frekvencije manje od zadatog celog broja. Ukoliko koristimo ovu opciju, zadati broj mora biti veći ili jednak 2.
- **Compare column proportions.** Ova opcija računa srazmeru kolona u parovima i ukazuje na to koji parovi kolona (u odnosu na zadati red) se značajno razlikuju. Značajne razlike su prikazane u tabeli kontigencije pomoću malih slova i računaju se sa pragom značajnosti od 0.05.
 - **Adjust p-values (Bonferroni method).** Srazmera kolana u parovima se računa Bonferroni-ovom metodom koja ispravlja prag značajnosti za posmatrane podatke.

- **Percentages.** Možemo dodati procenke po kolonama i redovima u tabeli kontingencije. Takođe, možemo dodati i procenke u odnosu na ceo uzorak.
- **Residuals.** Prikazuju odgovarajuće rezidualne.
 - **Unstandardized.** Razlika između posmatranih i očekivanih vrednosti. Očekivana vrednost je očekivani broj slučajeva u ćeliji pod pretpostavkom da su promenljive nezavisne.
 - **Standardized.** Nestandardizovani reziduali podeljeni sa procenom njihove standardne devijacije. Zovu se i Pearson-ovi reziduali i imaju očekivanje 0 i disperziju 1.
 - **Adjusted standardized.** Nestandardizovani reziduali podeljeni sa procenom svoje standardne greške. Rezultujući standardizovani rezidual se izražava u jedinicama standardne devijacije iznad ili ispod očekivane vrednosti.
- **Noninteger weights.** Kada su podaci pomnoženi težinskom racionalnom konstantom onda brojevi slučajeva u ćelijama ne moraju biti celobrojni. Možemo im skratiti razlomljeni deo ili zaokružiti na ceo broj pre ili posle računanja broja slučaja ili koristiti racionalne brojeve i za prikazivanje i za računanje statistika.

Primer 1.

Testirati nezavisnost promenljivih *obrazovanje* i *pol* u bazi1.

Primer 1.

Tabela kontigencije
sa stvarnim i
očekivanim
frekvencijama



Nivo obrazovanja * pol zaposlenog Crosstabulation					
		pol zaposlenog		Total	
		muskarci	zene		
Nivo obrazovanja	8	Count	23	30	53
		Expected Count	28,8	24,2	53,0
	12	Count	62	128	190
		Expected Count	103,4	86,6	190,0
	14	Count	6	0	6
		Expected Count	3,3	2,7	6,0
	15	Count	83	33	116
		Expected Count	63,1	52,9	116,0
	16	Count	35	24	59
		Expected Count	32,1	26,9	59,0
	17	Count	10	1	11
		Expected Count	6,0	5,0	11,0
	18	Count	9	0	9
		Expected Count	4,9	4,1	9,0
	19	Count	27	0	27
		Expected Count	14,7	12,3	27,0
	20	Count	2	0	2
		Expected Count	1,1	,9	2,0
	21	Count	1	0	1
		Expected Count	,5	,5	1,0
Total		Count	258	216	474
		Expected Count	258,0	216,0	474,0

Primer 1.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	96,856 ^a	9	,000
Likelihood Ratio	115,880	9	,000
Linear-by-Linear Association	59,941	1	,000
N of Valid Cases	474		

a. 8 cells (40,0%) have expected count less than 5. The minimum expected count is ,46.

- Vidimo da je došlo do prekoračenja, 8 (40.0%) ćelija ima manje od 5 slučajeva. Stoga, test statistika Hi kvadrat testa neće biti dobro izračunata pa koristimo test količnika verodostojnosti. Realizana vrednost test statistike iznosi 115.880 a p -vrednost testa je manja od 0.001 pa prihvatamo alternativnu hipotezu o zavisnosti promenljivih obrazovanje i pol.

Primer 1.

Nivo obrazovanja * pol zaposlenog Crosstabulation					
			pol zaposlenog		Total
			muskarci	zene	
Nivo obrazovanja	1	Count	23	30	53
		Expected Count	28,8	24,2	53,0
	2	Count	62	128	190
		Expected Count	103,4	86,6	190,0
	3	Count	89	33	122
		Expected Count	66,4	55,6	122,0
	4	Count	84	25	109
		Expected Count	59,3	49,7	109,0
	Total	Count	258	216	474
		Expected Count	258,0	216,0	474,0

Problem prekoračenja pretpostavke da u svakoj ćeliji postoji bar 5 slučajeva rešava se spajanjem ćelija što se u *SPSS*-u implementira prekodiranjem promenljivih. U ovom slučaju smo nivoe obrazovanja prekodirali u 1, 2, 3 i 4 (1-osnovno, 2-srednje, 3-viša škola, 4-minimum visoka škola).

Primer 1.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	78,385 ^a	3	,000
Likelihood Ratio	81,019	3	,000
Linear-by-Linear Association	55,674	1	,000
N of Valid Cases	474		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 24,15.

- Vidimo da sada ne dolazi do prekoračenja pretpostavke.
- Test statistika χ^2 kvadrat testa je 78,385, a p -vrednost je manja od 0.001 pa odbacujemo nultu hipotezu, odnosno prihvatamo hipotezu o zavisnosti.

Symmetric Measures

	Value	Approx. Sig.
Phi	,407	,000
Nominal by Nominal Cramer's V	,407	,000
Contingency Coefficient	,377	,000
N of Valid Cases	474	

- Testovi jačine zavisnosti nam kazuju da postoji slaba zavisnost između nivoa obrazovanja i pola osobe.

Primer 2.

Testirati jačinu veze između promenljivih *kategorija* *posla* i *pol* iz baze1.

Primer 2.

Tabela
kontigencije sa
stvarnim i
očekivanim
frekvencijama.



Kategorija zaposlenih * pol zaposlenog Crosstabulation					
			pol zaposlenog		Total
			muskarci	zene	
Kategorija zaposlenih	sluzbenik	Count	110	117	227
		Expected Count	123,6	103,4	227,0
	obuceni sluzbenik	Count	47	89	136
		Expected Count	74,0	62,0	136,0
	sluzbenik obezbedjenja	Count	27	0	27
		Expected Count	14,7	12,3	27,0
	sluzbenik sa fakult. diplomom	Count	34	7	41
		Expected Count	22,3	18,7	41,0
	posebna grupa sluzbenika	Count	30	2	32
		Expected Count	17,4	14,6	32,0
	MBA diploma	Count	4	1	5
		Expected Count	2,7	2,3	5,0
	tehnicka sluzba	Count	6	0	6
		Expected Count	3,3	2,7	6,0
	Total	Count	258	216	474
		Expected Count	258,0	216,0	474,0

Primer 2.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	87,230 ^a	6	,000
Likelihood Ratio	106,109	6	,000
Linear-by-Linear Association	48,161	1	,000
N of Valid Cases	474		

a. 4 cells (28,6%) have expected count less than 5. The minimum expected count is 2,28.

Symmetric Measures

		Value	Approx. Sig.
	Phi	,429	,000
Nominal by Nominal	Cramer's V	,429	,000
	Contingency Coefficient	,394	,000
N of Valid Cases		474	

Slično kao u primeru 1, zaključujemo da postoji jak dokaz o postojanju zavisnosti između kategorije posla i pola.

Kako tablica kontigencije u ovom slučaju nije kvadratna, korišćemo Cramer-ov V test. Dobijamo vrednost od 0.429 što bi bila neka slaba do umerena zavisnost.

Testirali smo i koeficijent kontigencije gde je dobijena vrednost 0.394 što potvrđuje naš zaključak o slaboj do umerenoj zavisnosti.



Hvala na pažnji!