

Neparametarski testovi: Hi-kvadrat test



PREDAVANJE BR.7

Uvod



- P6: Testiranje pretpostavke o raspodjeli vjerovatnoća kategoričke varijable
- P7: Analiza dvodimenzionalnih kontigencijskih tabela

Testiranje pretpostavke o nezavisnosti dvije kategoričke varijable



- Tabela kontigencije ili tabela unakrsnog razvrstavanja
- Ako jedna varijabla ima 3 kategorije, a druga varijabla 2 kategorije – 3×2 (3 reda i 2 kolone)
- Ukrštanjem reda i kolone dobija se ćelija u tabeli kontigencije
- U ćelijama tabele kontigencije nalaze se zajedničke frekvencije, a zbrovi frekvencija u tabelama i kolonama nazivaju se marginalne frekvencije

Testiranje nezavisnosti



- $H_0: Q_1 \perp Q_2$ – varijable Q_1 i Q_2 su statistički nezavisne
 - indeks za kategorije $Q_1 - j$, (od 1 do r), a za kategorije varijable $Q_2 - k$, (od 1 do c).
- $H_0: \pi_{jk} = \pi_j * \pi_k$ (zajedničke vjerovatnoće jednake su proizvodu marginalnih).
 - Množenjem ocjena marginalnih vjerovatnoća, p_j i p_k dobijaju se ocjene zajedničkih vjerovatnoća p_{jk} .
 - Množenjem vjerovatnoća veličinom uzorka dobijaju se očekivane frekvence za ćeliju jk

Testiranje nezavisnosti



- Očekivana frekvencija za ćeliju u redu j i koloni k označava se sa f_{tjk}
- f_{tjk} se računa dijeljenjem proizvoda zbira frekvencija u redu j i zbira frekvencije u koloni k sa N , tj. sa veličinom uzorka.

Testiranje nezavisnosti



- Da li lijek pomaže pri liječenju neke bolesti?

		Zdrav		Total
		DA	NE	
Ljekovi	DA	10	50	60
	NE	50	10	60
Total		60	60	120

		Zdrav		Total
		DA	NE	
Ljekovi	DA	30	30	60
	NE	30	30	60
Total		60	60	120

- Ho: nema razlika između empirijskih i teorijskih frekvencija
- Ako teorijske frekvencije = nema povezanosti
 - Ho: između empirijskih frekvenci nema povezanosti
- Ako $p < 0.05$ – odbacujemo Ho
 - Postoji povezanost dvije varijable – zdravlja i uzimanja lijeka!

Određivanje Ft



Ljekovi / Zdravlje	DA	NE
DA	a	b
NE	c	d

Ljekovi / Zdravlje	DA	NE
DA	a1	b1
NE	c1	d1

- Zajednička vjerovatnoća statistički nezavisnih događaja jednaka je proizvodu njihovih vjerovatnoća.
- $Ft(a1) = (a+b) * (a+c) / (a+b+c+d)$
- $Ft(b1) = (a+b) * (b+d) / (a+b+c+d)$
- $Ft(c1) = (c+d) * (a+c) / (a+b+c+d)$
- $Ft(d1) = (b+d) * (c+d) / (a+b+c+d)$

$$f_{tjk} = p_j * p_k$$

$$f_{tjk} = \frac{f_{oj} * f_{ok}}{n}$$

χ^2 test nezavisnosti



- Pirsonova statistika

$$\chi^2 = \sum_{k=1}^c \sum_{j=1}^r \frac{(f_{o\ jk} - f_{t\ jk})^2}{f_{jk}}$$

- Stepeni slobode: $(c-1)*(r-1)$
- c – broj redova; r – broj kolona

Intenzitet povezanosti



- Različiti koeficijenti: viša vrijednost – viša povezanost
- Samo neki: od 0 (nema povezanosti) do 1 (visoka povezanost)
- Kramerov F_i koeficijent
 - n : veličina uzorka, q : manja vrijednost broja redova i kolona
- F_i koeficijent: Kramerov V za tabele 2×2

$$\text{Cramerov } F_i = \sqrt{\frac{\chi^2}{N(s-1)}}, s = \min(r, c)$$

Koeficijent kontigencije $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

- Drugi: Količnik šansi, količnik rizika, razlika rizika itd.

Nastavak primjera 1



- Da li lijek pomaže pri liječenju neke bolesti?

Ljekovi * Zdrav Crosstabulation

		Zdrav		Total	
		DA	NE		
Ljekovi	DA	Count	10	50	60
		Expected Count	30.0	30.0	60.0
	NE	Count	50	10	60
		Expected Count	30.0	30.0	60.0
Total	Count	60	60	120	
	Expected Count	60.0	60.0	120.0	

fo	ft	fo-ft	(fo-ft) ²	(fo-ft) ² /ft
10	30	-20	400	13.33
50	30	20	400	13.33
50	30	20	400	13.33
10	30	-20	400	13.33
			hi-kvadrat	53.33

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	53.333 ^a	1	.000		
Continuity Correction ^b	50.700	1	.000		
Likelihood Ratio	58.221	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	120				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 30.00.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.667	.000
	Cramer's V	.667	.000
	Contingency Coefficient	.555	.000
N of Valid Cases		120	

- SPSS: Analyse/Descriptives Statistics/Crostats...

Primjer 2



2. Da li žene voze lošije od muškaraca?

		Nezgode		Total
		Sa nezgodama	Bez nezgoda	
Pol	Muški	23	49	72
	Ženski	27	21	48
Total		50	70	120

		Nezgode		Total	
		Sa nezgodama	Bez nezgoda		
Pol	Muški	Count	23	49	72
		Expected Count	30.0	42.0	72.0
	Ženski	Count	27	21	48
		Expected Count	20.0	28.0	48.0
Total		Count	50	70	120
		Expected Count	50.0	70.0	120.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.000 ^a	1	.008		
Continuity Correction ^b	6.036	1	.014		
Likelihood Ratio	7.007	1	.008		
Fisher's Exact Test				.014	.007
Linear-by-Linear Association	6.942	1	.008		
N of Valid Cases	120				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 20.00.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.242	.008
	Cramer's V	.242	.008
	Contingency Coefficient	.235	.008
N of Valid Cases		120	

Primjer 3



- Jedan medicinski centar izvršio je ispitivanje oboljenja od epidemije gripa u preduzećima u kojima su radnici bili nevakcinisani, neki vakcinisani 11 mjeseci prije epidemije, a neki neposredno prije epidemije. Da li postoji zavisnost dobijanja gripa od načina vakcinisanja?

		Oboljeli		Total
		Oboljeli	Nisu oboljeli	
Vakcina	Nevakcinisani	402	2497	2899
	Vakcina 11 mjeseci ranije	378	3789	4167
	Vakcina neposredno prije epidemije	131	2009	2140
Total		911	8295	9206

Primjer 3



			Oboljeli		Total
			Oboljeli	Nisu oboljeli	
Vakcina	Nevakcinisani	Count	402	2497	2899
		Expected Count	286.9	2612.1	2899.0
	Vakcina 11 mjeseci ranije	Count	378	3789	4167
		Expected Count	412.4	3754.6	4167.0
	Vakcina neposredno prije epidemije	Count	131	2009	2140
		Expected Count	211.8	1928.2	2140.0
Total		Count	911	8295	9206
		Expected Count	911.0	8295.0	9206.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	88.637 ^a	2	.000
Likelihood Ratio	88.467	2	.000
Linear-by-Linear Association	86.472	1	.000
N of Valid Cases	9206		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 211.77.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.098	.000
	Cramer's V	.098	.000
	Contingency Coefficient	.098	.000
N of Valid Cases		9206	

Primjer 3 – pretvaranje frekvencija u procenete radi interpretacije



		Oboljeli		Total
		Oboljeli	Nisu oboljeli	
Vakcina	Nevakcinisani	13.9%	86.1%	100.0%
	Vakcina 11 mjeseci ranije	9.1%	90.9%	100.0%
	Vakcina neposredno prije epidemije	6.1%	93.9%	100.0%
Total		9.9%	90.1%	100.0%

		Oboljeli		Total
		Oboljeli	Nisu oboljeli	
Vakcina	Nevakcinisani	44.1%	30.1%	31.5%
	Vakcina 11 mjeseci ranije	41.5%	45.7%	45.3%
	Vakcina neposredno prije epidemije	14.4%	24.2%	23.2%
Total		100.0%	100.0%	100.0%

Test nezavisnosti kod 2 zavisna uzorka (McNemarov test)



4. 100 ispitanika podvrnuto je ispitivanju psihomotorike pomoću dva testa koji su takve prirode da daju samo alternativne rezultate: “zadovoljava” i “ne zadovoljava”. U prvom testu 60 ispitanika zadovoljava, a u drugom 70. Je li ta razlika statistički značajna, tj. je li drugi test stvarno lakši od prvog?

		test2		Total
		Zadovoljava	Nezadovoljav a	
test1	Zadovoljava	55	5	60
	Nezadovoljava	15	25	40
Total		70	30	100

	Value	Exact Sig. (2- sided)
McNemar Test		.041 ^a
N of Valid Cases	100	

a. Binomial distribution used.

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t} = \frac{(A - D)^2}{A + D}$$

$$A=5, B=55, C=25, D=15$$