

Testiranje statističkih hipoteza 2 (Testiranje statističke značajnosti koeficijenta linearne korelacije, hi-kvadrat test distribucije verovatnoća i testiranje hipoteze o statističkoj nezavisnosti dveju kategoričkih varijabli): rešenja

(Rešenja zadataka uradili: Marija Balović, studentkinja psihologije, broj indeksa PS120015, Andrej Gađanski, student psihologije, broj indeksa PS160059 i Luka Milošević, student psihologije, broj indeksa PS160032. Rešenja dopunio Lazar Tenjović)

Zadatak 1

U fajlu **gzptbu.sav** nalaze se, između ostalog, podaci o emocionalnoj uravnoteženosti (varijabla **EU**) i društvenosti (varijabla **DRUS**) slučajnog uzorka studenata.

- Utvrditi da li u populaciji studenata postoji linearna povezanost ovih dveju varijabli (testirati nultu hipotezu: koeficijent linearne korelacije u populaciji jednak je nuli).

Rešenje:

Postavićemo sledeće početne hipoteze:

$H_0: \rho = 0$ (ρ je koeficijent linearne korelacije u populaciji)

$H_1: \rho \neq 0$

U meniju **Analyze** idemo na **Correlate** i biramo opciju **Bivariate**. Nakon toga će nam se otvoriti prozor u kojem biramo između kojih varijabli želimo da izračunamo korelaciju i kakvu korelaciju zapravo želimo da izračunamo. U ovom zadatku nema potrebe podešavati nikakve opcije, jer je već sve podešeno za računanje Brave-Pirsonovog koeficijenta, tj. koeficijenta linearne korelacije. Jedino je potrebno prebaciti varijable **EU** i **DRUS** u desni okvir kako bismo izračunali korelaciju. Klikom na **OK** u ispisu dobijamo sledeću tabelu:

Correlations

		EMOCIONALNA URAVNOTEŽENOST	DRUSTVENOST
EMOCIONALNA URAVNOTEŽENOST	Pearson Correlation	1	.361**
	Sig. (2-tailed)		.000
	N	960	960
DRUSTVENOST	Pearson Correlation	.361**	1
	Sig. (2-tailed)	.000	
	N	960	960

** . Correlation is significant at the 0.01 level (2-tailed).

Iz tabele vidimo da Brave-Pirsonov koeficijent iznosi 0.361. Verovatnoća u redu **Sig. (2-tailed)** je manja od 0.05, što znači da možemo da odbacimo nultu hipotezu. Dakle, dobijeni koeficijent korelacije je statistički značajan. Na osnovu toga i činjenice da je koeficijent koji smo dobili na uzorku jednak 0.361 možemo zaključiti da u populaciji studenata postoji umerena (ili osrednja) pozitivna linearna povezanost između društvenosti i emocionalne uravnoteženosti.

Napomena: U statističkom zaključivanju treba biti pažljiv, u slučaju da je dobijeno Sig. (2-tailed) > 0.05 ne bismo imali osnova da odbacimo nultu hipotezu, ali ne bi bilo ni opravdano reći da ne postoji linearna povezanost u populaciji. Naime, neodbacivanje nulte hipoteze u jednom istraživanju nije dovoljan argument za takav zaključak. Jedino što bismo na osnovu toga mogli da tvrdimo jeste da dobijeni koeficijent linearne korelacije nije statistički značajan.

Zadatak 2.

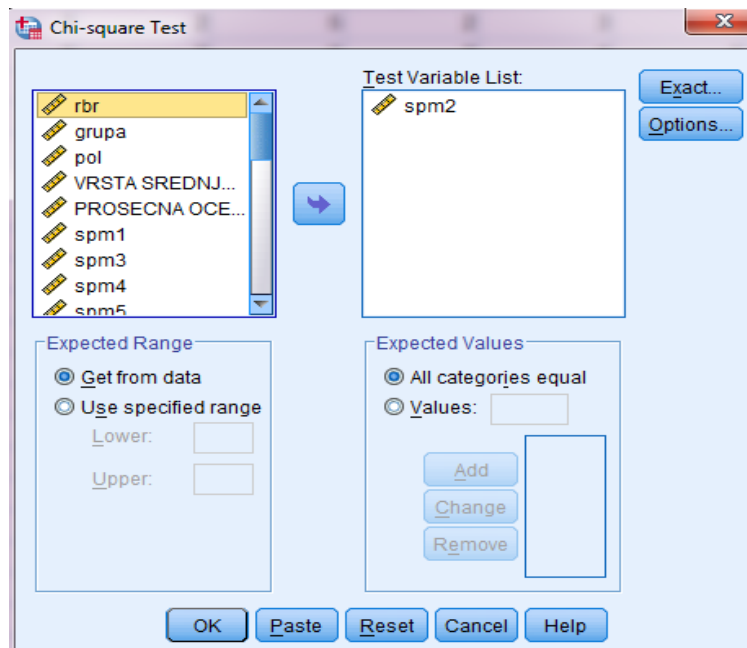
U fajlu **stavmat.sav** nalaze se, između ostalog, odgovori na jedno pitanje iz upitnika (varijabla **spm2**) koje su ispitanici dali koristeći petostepenu skalu za odgovaranje.

- Testirajte pretpostavku da su ispitanici nasumično odgovarali na ovo pitanje.

Rešenje:

Za testiranje pretpostavke da su ispitanici nasumično odgovarali na ovo pitanje, koristimo *hi-kvadrat test za testiranje pretpostavke o distribuciji verovatnoća jedne kategoričke varijable*. Dakle, odabrati meni **Analyze / Legacy Dialogs / Nonparametric tests / Chi-square**. Varijablu **spm2** ubacujemo u okvir **Test Variable list**, a u okviru **Expected Values** ostavljamo uključenu opciju **All categories equal**, jer proveravamo da li su ispitanici odgovarali na pitanje nasumično, što znači da je distribucija odgovora uniformna (verovatnoća svakog mogućeg odgovora na pitanje jednaka je 1/5, budući da ima

5 mogućih odgovora na pitanje). Dakle, nulta hipoteza glasi: verovatnoće svih odgovora su jednake, tj. jednake $1/5$ ili 0.2 .



U ispisu dobijamo tabele sa imenom varijable **spm2** i **Test Statistics**. U prvoj tabeli vidimo opažene (**Observed N**) i očekivane (**Expected N**) frekvencije, kao i njihove razlike, tj. rezidualne. Iz tabele **Test Statistics** vidimo da hi-kvadrat statistik iznosi 331.64 i da je statistički značajan, zbog čega možemo odbaciti nultu hipotezu i zaključiti da ispitanici po svoj prilici nisu nasumično odgovarali na ovo pitanje.

spm2			
	Observed N	Expected N	Residual
1	2	48.2	-46.2
2	1	48.2	-47.2
3	14	48.2	-34.2
4	78	48.2	29.8
5	146	48.2	97.8
Total	241		

Test Statistics

	spm2
Chi-Square	331,635 ^a
	4
Asymp. Sig.	,000

a. 0 cells (0,0%) have expected frequencies less than 5.
The minimum expected cell frequency is 48,2.

Zadatak 3.

U fajlu **estrogen_alchajm.sav** nalaze se podaci iz medicinskog istraživanja čiji rezultati su objavljeni u poznatom časopisu *Lancet* (Tang, Jacobs, & Stern, 1996): U varijabli **estrogen** za svaku od 1124 žene iz slučajnog uzorka dati su podaci o tome da li su dobijale estrogene tokom menopauze (0-NE; 1-DA) a u varijabli **alchajmb** da li su obolele od Alchajmerove bolesti tokom 5-godišnjeg perioda praćenja (0-NE; 1-DA).

- Tabelom kontingencije predstaviti bivarijatnu raspodelu varijabli **estrogen** i **alchajmb** i izračunati meru asocijacije ovih dveju varijabli (fi-koeficijent u ovom slučaju). (Varijablu **estrogen** stavite u redove, a varijablu **alchajmb** u kolone tabele kontingencije);
- Testirati statističku značajnost fi-koeficijenta.
- Izračunati šanse da se oboli od Alchajmerove bolesti za žene koje jesu primale estrogen tokom menopauze;
- Izračunati šanse da se oboli od Alchajmerove bolesti za žene koje nisu primale estrogen tokom menopauze;
- Izračunati količnik šansi za žene koje nisu primale estrogen u odnosu na žene koje jesu primale estrogen.

Rešenje:

Tabelu kojom se predstavlja bivarijatna raspodela dveju kategoričkih varijabli dobija se tako što se odabire meni **Analyze/Descriptive Statistics/Crosstabs**. U polje Row(s) stavljamo **estrogen**, dok u polje Column(s) stavljamo **alchajmb**. Da bi se izracunao fi-koeficijent neophodno je kliknuti na dugme **Statistics** i u okviru *Nominal* izabrati *Phi and Cramer's V*. Posle toga treba kliknuti na dugme **Continue** i na **OK**.

U ispisu se dobijaju dve tabele.

Prva tabela predstavlja bivarijatnu raspodelu frekvencija datih varijabli:

Da li je primala estrogene tokom menopauze * Da li je tokom 5 godina obolela od Alchajmerove bolesti Crosstabulation

Count		Da li je tokom 5 godina obolela od Alchajmerove bolesti		Total
		ne	da	
Da li je primala estrogene tokom menopauze	ne	810	158	968
	da	147	9	156
Total		957	167	1124

Klikom na dugme **Cells** pri definisanju procedure, i potom uključivanjem opcije **Raw** i **Column** u okviru Percentages možemo u tabeli dobiti i procenete u odnosu na zbir frekvencija u redu, odnosno u odnosu na zbir frekvencija u koloni kojoj pripada ćelija. Tada bi ova tabela izgledala ovako:

Da li je primala estrogene tokom menopauze ^ Da li je tokom 5 godina obolela od Alchajmerove bolesti
Crosstabulation

			Da li je tokom 5 godina obolela od Alchajmerove bolesti		Total
			0 ne	1 da	
Da li je primala estrogene tokom menopauze	0 ne	Count	810	158	968
		% within Da li je primala estrogene tokom menopauze	83.7%	16.3%	100.0%
		% within Da li je tokom 5 godina obolela od Alchajmerove bolesti	84.6%	94.6%	86.1%
	1 da	Count	147	9	156
		% within Da li je primala estrogene tokom menopauze	94.2%	5.8%	100.0%
		% within Da li je tokom 5 godina obolela od Alchajmerove bolesti	15.4%	5.4%	13.9%
Total	Count		957	167	1124
	% within Da li je primala estrogene tokom menopauze		85.1%	14.9%	100.0%
	% within Da li je tokom 5 godina obolela od Alchajmerove bolesti		100.0%	100.0%	100.0%

U drugoj tabeli se dobijaju vrednosti C-koeficijenta i Kramerovog V-koeficijenta:

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.103	.001
	Cramer's V	.103	.001
	Contingency Coefficient	.102	.001
N of Valid Cases		1124	

Statističku značajnost fi-koeficijenta testiramo Pirsonovim hi-kvadrat testom. Iz kolone **Approx. Sig.** vidimo da je fi-koeficijent statistički značajan.

Uočimo da je Kramerov V-koeficijent u slučaju tabele 2x2 jednak fi-koeficijentu jedino što SPSS za Fi-koeficijent dodeljuje predznak, ovog puta minus. Predznak fi-koeficijenta se određuje na osnovu raspodele frekvencija u dijagonalama tabele ali nije uvek koristan za tumačenje tabele.

*Napomena: Ukoliko želimo da vidimo u ispisu i tabelu sa hi-kvadrat testom potrebno je pri definisanju procedure po kliku na dugme **Statistics** uključiti **Chi-square**. Tada u ispisu dobijamo i sledeću tabelu:*

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson Chi-Square	11.828 ^a	1	.001		
Continuity Correction ^b	11.008	1	.001		
Likelihood Ratio	14.390	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	11.817	1	.001		
N of Valid Cases	1124				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 23.18.

b. Computed only for a 2x2 table

Iz tabele **Chi-Square Tests** vidimo da je Pearsonov hi-kvadrat statistik jednak 11.83 i da je statistički značajan. Uočimo da je verovatnoća u koloni *Asymptotic Significance (2-sided)* jednaka onoj u koloni *Approx. Sig.* za fi-koeficijent. To je uvek slučaj budući da se statistička značajnost fi-koeficijenta određuje preko značajnosti hi-kvadrat statistika. Ukoliko je hi-kvadrat statistik statistički značajan tada je nužno i fi-koeficijent statistički značajan.

Šansa da se oboli od Alchajmerove bolesti za žene koje su primale estrogen tokom menopauze dobija se tako što se frekvencija žena koje su primale estrogen tokom menopauze i obolele od Alchajmerove bolesti podeli sa frekvencijom žena koje su primale estrogen tokom menopauze ali nisu obolele od Alchajmerove bolesti.

Da bi se ovo uradilo potrebno je odabrati meni **Transform/Compute Variable** i u polju *Target Variable* upisati ime nove varijable čija će vrednost biti jednaka traženim šansama (npr. **soape**). U polje *Numeric Expression* upisujemo **9/147**. Klikom na dugme **OK**, u prozoru za podatke u opciji *Data View* dobijamo tražene šanse u vidu nove varijable **soape**:

	rbr	estrogen	alchajmb	soape
1	1	1	0	.06
2	2	1	0	.06
3	3	1	0	.06
4	4	1	0	.06
5	5	1	0	.06
6	6	1	0	.06
7	7	1	0	.06
8	8	1	0	.06
9	9	1	0	.06
10	10	1	0	.06
11	11	1	0	.06

Šanse da se oboli od Alchajmerove bolesti za žene koje nisu primale estrogen tokom menopauze dobija se tako što se frekvencija žena koje nisu primale estrogen tokom menopauze, a pritom su obolele od Alchajmerove bolesti podeli sa frekvencijom žena koje nisu primale estrogen tokom menopauze i nisu obolele od Alchajmerove bolesti.

Da bi se ovo uradilo potrebno je opet odabrati meni **Transform/Compute Variable** i u polju *Target Variable* upisati ime varijable čija će vrednost biti jednaka traženim šansama (npr. **soanpe**). U polje *Numeric Expression* upisujemo **158/810**. Klikom na dugme **OK**, u prozoru za podatke u opciji *Data View* dobijamo tražene šanse u vidu nove "varijable":

	rbr	estrogen	alchajmb	soape	soanpe
1	1	1	0	.06	.20
2	2	1	0	.06	.20
3	3	1	0	.06	.20
4	4	1	0	.06	.20
5	5	1	0	.06	.20
6	6	1	0	.06	.20
7	7	1	0	.06	.20

Količnik šansi oboljevanja od Alchajmerove bolesti za žene koje nisu primale estrogen u odnosu na žene koje jesu primale estrogen izračunava se tako što se šanse da se oboli od Alchajmerove bolesti za žene koje nisu primale estrogen podeli sa šansama da se oboli od Alchajmerove bolesti za žene koje su primale estrogen.

Da bi se ova vrednost dobila potrebno je, kao i u prethodnim slučajevima, odabrati meni *Transform/Compute Variable* i u polju *Target Variable* upisati ime varijable čija će vrednost biti jednaka traženom količniku šansi (npr. **ks**). U polje *Numeric Expression* upisujemo **soanpe/soape**. Klikom na dugme **OK**, u prozoru za podatke u opciji *Data View* dobijamo traženi količnik šansi u vidu nove "varijable" **ks**:

	rbr	estrogen	alchajmb	soape	soanpe	ks
1	1	1	0	.06	.20	3.19
2	2	1	0	.06	.20	3.19
3	3	1	0	.06	.20	3.19
4	4	1	0	.06	.20	3.19
5	5	1	0	.06	.20	3.19
6	6	1	0	.06	.20	3.19

Dakle, kada je o ženama iz ovog uzorka reč, žene koje nisu primale estrogen tokom menopauze imaju 3.19 puta veće šanse da obole od Alchajmerove bolesti nego žene koje jesu primale estrogen tokom menopauze.

*Napomena: Količnik šansi možemo automatski dobiti i u proceduri **Crosstabs** i pri tome testirati statističku značajnost količnika šansi: klikom na dugme **Statistics** pri definisanju procedure **Crosstabs** i odabirom opcije **Risk**. Međutim, u tabeli iz ove*

procedure koja sadrži količnik šansi teško je se snaći. Zavisno od toga kako su označene kategorije dobiće se traženi količnik šansi ili njegova recipročna vrednost. U konkretnom slučaju kada je kategorija „nije primala estrogen“ označena sa 0 a kategorija „jeste primala estrogen“ označena sa 1, ispis bi izgledao ovako:

Risk Estimate			
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Da li je primala estrogene tokom menopauze (0 ne / 1 da)	.314	.157	.628
For cohort Da li je tokom 5 godina obolela od Alchajmerove bolesti = 0 ne	.888	.847	.931
For cohort Da li je tokom 5 godina obolela od Alchajmerove bolesti = 1 da	2.829	1.477	5.420
N of Valid Cases	1124		

Količnik šansi (Odds ratio) u tabeli je 0.314 a traženi količnik šansi je zapravo recipročna vrednost količnika šansi iz tabele. Ukoliko varijablu **estrogen** rekodiramo tako da oznakom 1 označimo kategoriju „nije primala estrogen“ a oznakom 0 kategoriju „jeste primala estrogen“ ova tabela će u ispisu izgledati ovako:

Risk Estimate			
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for estrogen_nije1_jeste0 (.00 / 1.00)	3.186	1.591	6.379
For cohort Da li je tokom 5 godina obolela od Alchajmerove bolesti = 0 ne	1.126	1.074	1.181
For cohort Da li je tokom 5 godina obolela od Alchajmerove bolesti = 1 da	.353	.185	.677
N of Valid Cases	1124		

Sada je u tabeli prikazan baš traženi količnik šansi (3.186 $\approx$ 3.19). Budući da 95% interval poverenja za količnik šansi ne obuhvata jedinicu, količnik šansi je statistički značajan.

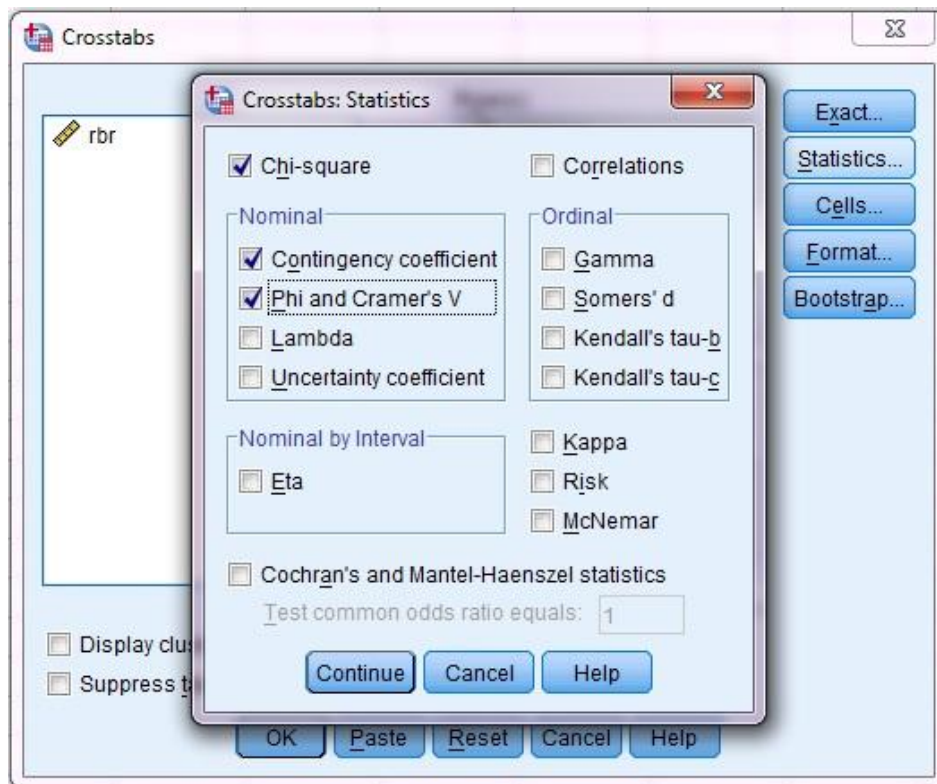
Zadatak 4.

U fajlu **abortus_velporpor.sav** su dati podaci sa slučajnog uzorka o veličini porodice porekla ispitanika – varijabla **velporpor** (iskazanoj prema broju dece: 1 – sa jednim detetom, 2 – sa dva deteta 3 – sa tri deteta 4 – sa četvoro ili više dece) i o odnosu ispitanika prema abortusu – varijabla **abortus** (kategorije: 1 – svaka žena ima pravo da odlučuje o tome da li će abortirati; 2 – odobravam abortus samo u određenim slučajevima kad je to medicinski opravdano; 3 – abortus uopšte ne treba dozvoliti).

- Utvrditi da li u populaciji iz koje je uzorak postoji povezanost između veličine porodice porekla i odnosa prema abortusu (izračunati koeficijente asocijacije – C-koeficijent i Kramеров V koeficijent i testirati njihovu statističku značajnost).

Rešenje:

Postupak započinjemo sa **Analyze/Descriptive statistics/Crosstabs**. U **Row(s)** ubacujemo varijablu **velporpor**, a u **Column(s)** varijablu **abortus**. Zatim kliknemo na **Statistics** i u okviru prozora **Crosstabs: Statistics** štrikliramo **Chi-square**, **Contingency coefficient** i **Phi and Cramer's V**. Potom klikom na dugme **Cells** uključimo opcije **Row** i **Column** u okviru **Percentages**.



Ispis iz procedure sadrži sledeće tabele:

Velicina porodice porekla prema broju dece ^ Odnos prema abortusu Crosstabulation

			Odnos prema abortusu			Total
			1 pravozeno_ daodlu ci	2 samopododredjeni mokolnostima	3 netrebadozvoliti	
Velicina porodice porekla prema broju dece	1 samojednodete	Count	20	18	9	47
		% within Velicina porodice porekla prema broju dece	42.6%	38.3%	19.1%	100.0%
		% within Odnos prema abortusu	39.2%	24.0%	16.7%	26.1%
	2 dvoje dece	Count	18	16	16	50
		% within Velicina porodice porekla prema broju dece	36.0%	32.0%	32.0%	100.0%
		% within Odnos prema abortusu	35.3%	21.3%	29.6%	27.8%
	3 trojedece	Count	5	27	17	49
		% within Velicina porodice porekla prema broju dece	10.2%	55.1%	34.7%	100.0%
		% within Odnos prema abortusu	9.8%	36.0%	31.5%	27.2%
	4 cevoroivisece	Count	8	14	12	34
		% within Velicina porodice porekla prema broju dece	23.5%	41.2%	35.3%	100.0%
		% within Odnos prema abortusu	15.7%	18.7%	22.2%	18.9%
Total	Count		51	75	54	180
	% within Velicina porodice porekla prema broju dece		28.3%	41.7%	30.0%	100.0%
	% within Odnos prema abortusu		100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)
Pearson Chi-Square	16.315 ^a	6	.012
Likelihood Ratio	17.811	6	.007
Linear-by-Linear Association	7.413	1	.006
N of Valid Cases	180		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.83.

Symmetric Measures

	Value	Approximate Significance
Nominal by Nominal Phi	.301	.012
Cramer's V	.213	.012
Contingency Coefficient	.288	.012
N of Valid Cases	180	

Najbolji pokazatelj asocijacije u tabeli **Symmetric measures** je Kramerov V-koeficijent čija je vrednost u ovom slučaju 0.213.

U tabeli **Chi-Square Tests** vidimo da je verovatnoća da Pirsonov hi-kvadrat statistik (ako je nulta hipoteza tačna) bude 16.315 ili veći jednaka 0.012 što je manje od 0.05 i stoga zaključujemo da se

nulta hipoteza može odbaciti . Dakle, iz toga sledi (što vidimo i iz kolone **Approximate Significance** u tabeli **Symmetric Measures**) da je Kramеров V-koeficijent statistički značajan.

*Napomena: Budući da se statistička značajnost C-koeficijenta i Kramеровog V-koeficijenta utvrđuje Pirsonovim hi-kvadrat testom, verovatnoće u koloni **Approximate Significance** tabele **Symmetric measure** nužno su jednake verovatnoći u koloni **Asymptotic Significance (2-sided)** za Pirsonov hi-kvadrat statistik u tabeli **Chi-Square Tests**.*

Prema tome, u populaciji postoji povezanost između veličine porodice porekla i odnosa prema abortusu.