

PRIMJENA HI-KVADRAT TESTA U SOCIOLOŠKIM ISTRAŽIVANJIMA

Mihajlo Mijanović
Filosofski fakultet Nikšić

Veliki broj socioloških istraživanja podrazumijeva adekvatnu statističku obradu. Statistička obrada podataka je neminovnost ako želimo postići egzaktnost koju u principu eksplicitno definišemo u cilju istraživanja i postavljenim hipotezama. Ovom prilikom nemamo pretenzije da iznosimo sve poteškoće koje prate jedan istraživački rad, već da pomognemo istraživačima i naučnicima kako da dobivene rezultate mukotrpno rada što bolje osmisle i učine ih razumljivijim i pristupačnijim onima koji žele da se koriste dobivenim rezultatima. Sigurno je da manjkavost nekog istraživanja, posebno ako je u pitanju sumnjiva valjanost kod sirovih podataka ne možemo nadoknaditi statističkom obradom, ma kako kvalitetna bila. Statističke metode često upućuju na pomenutu manjkavost, što je takođe bitno za istraživače. Valjani podaci prikupljeni na osnovu nekog istraživanja ostaju nedorečeni i bez mogućnosti pune primjene ukoliko je odsutna adekvatna statistička obrada. Pod pojmom adekvatne statističke obrade podrazumijevamo primjenu statističkih metoda koje će omogućiti donošenje ispravnih zaključaka, koji su dalje zasnovani na postavljenim hipotezama i ciljevima istraživanja. Skloni smo konstataciji da nema dobrih i loših metoda već adekvatnih i neadekvatnih za konkretan slučaj. Ekspanzijom računarske tehnike i informatike, bitno se ubrzala statističko-matematička procedura, međutim ostao je problem odabira i primjene odgovarajućih statističkih metoda. Pored pomenutog problema evidentan je problem konšćenja obrađenih podataka tj. kvantitativnih pokazatelja. U istraživanju ova faza se pominje pod pojmom interpretacija dobivenih rezultata. Sociološka i mnoga druga istraživanja često su upućena na primjenu raznih anketa i upitnika, koji su po sadržini vrlo različita, a imaju za cilj da ispituju konkretnu problematiku npr. stavove javnog mnjenja prema rukovodećem kadru ili mišljenje tzv. pretplatnika o visini pretplate i kvalitetu programa, pristrasnost nekog informisanja o političkoj i ekonomskoj situaciji itd. Kako vidimo teoretski beskonačno veliki broj mogućnosti gdje dobivene podatke možemo dobiti uz pomoć upitnika, anketa, razgovora i sl. Ovakve i slične probleme dakako potrebno je na izvjestan način kvantificirati tj. podvrgnuti adekvatnoj statističkoj obradi, kako bi ponovo došli do nepristrasnih i valjanih kvalitativnih pokazatelja. U svakom slučaju cilj je dosta jasan, a put do cilja je često vrlo težak pod pretpostavkom da smo došli do valjanih sirovih podataka. U statističkoj teoriji poznato je više metoda koje se bave obradom ovakve problematike, a jedna od najčešćih je metoda koja se naziva

HI-KVADRAT TEST, oznaka (χ).

Šta je HI-kvadrat, kada i kako ga primijeniti ?

U najkraćem Hi-kvadrat je statistička metoda pomoću koje se ispituju-testiraju razlike između opaženih i očekivanih frekvencija. Hi-kvadrat uključuje proporcije i vjerovatnoće pod uslovom da su transformisane u apsolutne frekvencije. Jedna značajna osobina Hi-kvadrata je aditivnost, koja omogućuje kombinaciju većeg broja statistika ili drugih vrijednosti u istom testu. Na taj način mogu se testirati hipoteze koje imaju više od jednog niza podataka tj. više različitih serija i obilježja.

Hi-kvadrat je pogodan u slučajevima kada su pokazatelji kvantitativne i kvalitativne prirode. Sve vrijednosti kvalitativne prirode izražavaju se u apsolutnim frekvencijama i pretvaraju se u kvantitativne pokazatelje. Vrijednosti koje su kvantitativne prirode mogu se podvrgnuti obradi preko Hi-kvadrat testa

u slučajevima kada pojava nije u skladu sa normalnom Gaussovom funkcijom, dakle za primjenu ove metode nije neophodna normalna raspodjela niti kvantitativni pokazatelji, što je velika prednost u odnosu na brojne parametrijske statističke metode. Sama metoda je dosta jednostavna sa matematičko-statističkog aspekta i vrlo komplikovana sa aspekta procjene kada je smisleno primijeniti. Takođe često nije jednostavno napraviti korektnu interpretaciju rezultata dobijenu uz pomoć Hi-kvadrata. Sa statističkog aspekta najlakše je prihvatiti ili odbaciti postavljenu hipotezu sa faktorom greške koju unaprijed utvrdimo. Čin prihvatanja ili odbacivanja postavljenih hipoteza je nedvosmislen, a problem zbog čega je to tako često nije jednostavan i u svakom slučaju ostaje da se razjasni od strane naučnika-istraživača konkretne oblasti. Već smo rekli da Hi-kvadrat ima velikih pogodnosti - ne samo što nas ne obavezuje na normalnu raspodjelu i kvantitativne pokazatelje, već i zato što uz pomoć ove metode možemo ispitivati - testirati razlike ili sličnosti kada imamo jednu ili više serija sa različitim obilježjima, gdje broj statističkih jedinica odnosno apsolutnih frekvencija ne mora da bude isti u serijama. Dakle možemo testirati jedan ili više zavisnih ili nezavisnih uzoraka po više obilježja ili po istom obilježju sa više modaliteta.

U radu sa hi-kvadrat testom susrećemo se sa novim terminima koji su već upotrebljeni i za koje smatramo da su uglavnom jasni i zbog jednog broja čitalaca, izvinjenjem onima koji poznaju nešto bolje ovu metodu, detaljnije ćemo pojasniti termine: **opažene frekvencije, očekivane frekvencije, stepeni slobode, vjerovatnoća, granična vrijednost, upotreba tablica za prihvatanje ili odbacivanje postavljene hipoteze**. Sve pomenute termine pojasnićemo na primjerima koje susrećemo u istraživanjima socioloških i antropoloških nauka.

Opažene frekvencije (f_o)

Ako višimo neko mjerenje više puta ili su u pitanju različiti modaliteti istog mjerenja dobićemo neki niz ili statističku seriju. Prikupljeni podaci bilo kojim mjernim instrumentom nazivamo **opažene frekvencije**. Opažene frekvencije u daljem tekstu označavamo sa **(f_o)**.

Primjer:

Ispitali smo mišljenje studenata po pitanju ličnog standarda. U anketi su učestvovali studenti sa tri fakulteta slučajnim izborom.

Pitanje je glasilo: Da li ste zadovoljni sa ličnim standardom?

A) jesam B) nisam.

Rezultati su formirani u statističkoj tabeli br.1 i imaju slijedeći izgled.

Tabela br.1

fakulteti	odgovor jesam	odgovor nisam	ukupno
Filosofski	30	70	100
Pravni	40	100	140
Ekonomski	70	200	270
Ukupno	140	370	510

Kako vidimo sve numeričke vrijednosti prikazane u tabeli br.1 su opažene frekvencije.

Očekivane frekvencije (ft)

Ove frekvencije su u uskoj vezi sa postavljenom hipotezom. Možemo pretpostaviti da će se nešto dogoditi ili se nešto dogodilo sa nekim frekvencijama, koje prema našim predviđanjima imaju sasvim drugi raspored u odnosu na opažene frekvencije. Prema tome možemo postaviti bilo kakvu logičku hipotezu, a zatim je testirati prema opaženim frekvencijama. Hi-kvadrat upravo polazi od toga da ispita postoje li statistički značajne razlike između opaženih i očekivanih frekvencija. Opažene frekvencije često se nazivaju empirijske vrijednosti, a očekivane frekvencije - teoretske vrijednosti.

Hipoteze se mogu postaviti na različite načine, pa od toga zavisi kakve će biti očekivane frekvencije. Možemo postaviti hipotezu da će se događaji ravnomjerno rasporediti po svim modalitetima - takozvani **uniformni raspored**. Pod uslovom naprijed postavljene hipoteze imali bismo slijedeću situaciju.

Tabela br.1

opaž.fr.(fo)	30; 40; 50; 60; 20	200
oče.fr.(ft)	40; 40; 40; 40; 40	200

Treba skrenuti pažnju da suma opaženih frekvencija mora biti jednaka sumi očekivanih frekvencija. Opažene i očekivane frekvencije možemo prikazati grafički uz pomoć poligona frekvencija. U konkretnom slučaju možemo postaviti hipotezu da ne postoji statistički značajna razlika između opaženih i očekivanih frekvencija.

Kada govorimo o hipotezama moramo napomenuti da postoje dvije vrste hipoteza i to:

1. Nulta hipoteza (H0)

2. Radna hipoteza (H1)

Radnu hipotezu prihvatamo ukoliko testom utvrdimo da postoje statistički značajne razlike između opaženih i očekivanih frekvencija. U tom slučaju odbacujemo nultu hipotezu.

Nultu hipotezu prihvatamo ukoliko testom utvrdimo da ne postoji statistički značajna razlika između opaženih i očekivanih frekvencija. U tom slučaju odbacujemo radnu hipotezu.

U svakom slučaju uvijek jednu hipotezu prihvatimo a drugu odbacimo, zavisno od toga da li postoje ili ne postoje statistički značajne razlike između opaženih i očekivanih frekvencija.

Hi-kvadrat test najčešće upotrebljavamo u slijedećim slučajevima:

1. Kad imamo frekvencije jednog uzorka po različitim modalitetima, pa želimo ustanoviti da li te frekvencije odstupaju od frekvencija koje očekujemo uz neku hipotezu.
2. Kad imamo frekvencije dva ili više nezavisnih uzoraka, te želimo ustanoviti da li se uzorci razlikuju u opaženim frekvencijama odnosno modalitetima.
3. Kad imamo frekvencije dva ili više zavisnih uzoraka koji imaju dihotomna svojstva, te želimo ustanoviti razlike u mjerenim svojstvima.

Osnovni obrazac za izračunavanje Hi-kvadrat testa glasi:

$$\chi^2 = \sum_{i=1}^K \frac{(f_o - f_t)^2}{f_t}$$

Primjer Hi-kvadrat testa kad imamo frekvencije jednog zavisnog uzorka:

Ispitali smo mišljenje studenata o kvalitetu predavanja tokom osam semestara. Vrijednosti su transformisane u opažene frekvencije koje su prikazane u tabeli br.2.kolona (fo).

Pošto smo izračunali vrijednost Hi-kvadrata po opstem obrascu dobili smo vrijednost koja iznosi 64.43.

Tabela br.2

(fo - ft)² : ft

semestar	fo	ft	fo-ft	(fo-ft)	(fo-ft) ² : ft
prvi	70	90	-20	400	4.44
drugi	80	90	-10	100	1.11
treći	100	90	10	100	1.11
četvrti	120	90	30	900	10.00
peti	140	90	50	2500	27.77
šesti	60	90	-30	900	10.00
sedmi	90	90	0	0	0.00
osmi	60	90	-30	900	10.00
ukupno	720	720	0	-	64.43

Naš zadatak je da dobijenu vrijednost Hi-kvadrata interpretiramo. Da bismo upotrebili dobijenu vrijednost potrebno je prethodno razjasniti slijedeće:

1. Stepene slobode (df).
2. Vjerovatnoću (p).
3. Graničnu vrijednost (gv).
4. Upotrebu tablica za očitavanje Hi-kvadrat testa.

1. Stepene slobode (df)

Ako pretpostavimo da je zadato da se odredi 5 brojeva koji će imati zadatu sumu, prva četiri broja mogli bismo odmah odrediti uzimajući slobodno bilo koji broj u obzir do zadate sume. Tek kod određivanja petog broja nismo više slobodni, već prinuđeni da uzmemo tačno određeni broj.

Primjer:

Ako je suma 5 brojeva 30, neka su brojevi bili:

$10 + 5 + 3 + 4 + x = 30$. Kod prva četiri broja bili smo slobodni u izboru do zadate sume, a u petom slučaju ta sloboda je izgubljena. Tako da za ovakve primjere kažemo da smo izgubili jedan stepen slobode. Ako imamo slučajeve sa dvije računске operacije i sličan primjer gubimo dva stepena slobode itd. Veliki engleski statističar Fisher kaže da je broj stepena slobode ravan onoj veličini koja pokazuje koliko se klasa neka distribucija može odrediti proizvoljno. U našem primjeru broj stepena slobode zavisi od broja intervala odnosno broja semestara. Broj intervala označavamo sa (k). U primjeru imamo:

$$df = k - 1; df = 8 - 1 = 7.$$

2. Vjerovatnoća (p)

Određivanja stepena vjerovatnoće je u vezi sa kriterijumom prihvatanja ili odbacivanja postavljene hipoteze. Ovaj kriterijumi se mora unaprijed odrediti. U tablicama za očitavanje značajnosti Hi-kvadrat testa, taj kriterijumi dat je na nivou vjerovatnoće od:

$p = 0.05$; $p = 0.025$ i $p = 0.01$. Postavljeni kriterijumi su ustvari pouzdanost sa kojom ulazimo u prihvatanje i odbacivanje hipoteze.

Svakako da postoje i drugi nivoi kriterijuma, blaži i strožiji. Kad je riječ o ovom tipu istraživanja hipoteze se uglavnom testiraju na ova tri nivoa.

3. Granična vrijednost (gv)

Kada nam je poznat broj stepena slobode i kada smo odredili stepen vjerovatnoće sa kojom prihvatamo ili odbacujemo postavljenu hipotezu, tada iz tablica za Hi-kvadrat očitavamo graničnu vrijednost.

4. Upotreba tablica Hi-kvadrata

Tablice sadrže granične vrijednosti Hi-kvadrata za stepene slobode od 1 do 30 na pomenutim nivoima značajnosti. Stepene slobode su upisani u prvoj koloni, a vjerovatnoća u prvom redu. Granične vrijednosti za određenu vjerovatnoću i određeni stepen slobode nalaze se u presjeku zadatog reda i kolone. U prilogu ćemo dati samo dio tablice kako bi je mogli demonstrirati na našim primjerima.

Kako smo izračunali Hi-kvadrat koji u primjeru iznosi 64.43, tabela br.2, možemo pristupiti testiranju dobijene vrijednosti. U praksi je uobičajeno da se elementi za interpretaciju Hi-kvadrata daju pregledno slijedećim redoslijedom:

Hi-kvadrat = 64.43; df = 7; p = 0.01; gv = 18.5; S.

Tabela br.3

st.sl. (df)	p = 0.050	p = 0.025	p = 0.010
1	3.83	5.02	6.63
2	5.99	7.38	9.21
3	7.81	9.35	11.3
4	9.49	11.1	13.3
5	11.1	12.8	15.1
6	12.6	14.4	16.8
7	14.1	16.0	18.5
8	15.5	17.5	20.1
.	.	.	.
30	43.8	47.0	50.9

Znak (S) podrazumijeva statističku značajnost ili signifikantnost na određenom nivou pod oznakom (p).

Interpretacija na osnovu dobijenih rezultata iz tabele br.2 mogla bi se svesti na to da postoji statistički značajna razlika između opaženih i očekivanih vrijednosti. Ta razlika utvrđena je na nivou greške od 0.01. Kako je pomenuti nivo najstrožiji, što se vidi iz priloženih tablica, besmisleno je testirati hipotezu na preostala dva nivoa.

Kada je riječ o Hi-kvadrat testu u principu hipoteza se postavlja u nultom obliku tj. da nema razlike između opaženih i očekivanih frekvencija: $(f_o - f_t) = 0$, odnosno

Ho : Hi-kvadrat = 0.

Nultu hipotezu u konkretnom slučaju nedvosmisleno odbacujemo zato što je vrijednost Hi-kvadrata od 64.43 znatno iznad granične vrijednosti koja iznosi 18.5. U uvodu ovog rada napomenuli smo da je često neophodno analizovati strukturu Hi-kvadrat testa odnosno utvrditi koje su kategorije, klase ili grupe uticale najviše na veličinu Hi-kvadrata. Primjer iz tabele br.2 koncipiran je tako da se radi o zavisnom uzorku tj. jednim te istim studentima čije smo mišljenje testirali tokom osam semestara. Očekivali smo da studenti imaju nepromijenjeno mišljenje po pomenutoj problematici tokom svih osam semestara. Kako vidimo to se nije ostvarilo, a pokazatelji u zadnjoj koloni tabele br.2 ukazuju da su najveće razlike ispoljene u petom semestru, a najmanje u sedmom semestru, gdje u stvari nema razlika između opaženih i očekivanih frekvencija. Sustinu problema zbog čega je to tako, svakako da bi mogli dati profesori koji rade na pomenutom fakultetu. U svakom slučaju radi se o fiktivnom primjeru pa nas dobijeni rezultati sa tog stanovišta ne interesuju.

Hi-kvadrat test u tabelama kontigencije (2 x 2)

U praktičnom radu istraživači se često susreću sa situacijom kada treba modalitete neke varijable svesti na dva pola npr: da - ne, ima - nema, slažem - ne slažem, muško - žensko, dobro - loše itd. Ovakvu podjelu nazivamo **dihotomizacijom**.

Primjer:

Anketirano je 130 studenata slučajnim izborom. Od toga je bilo 60 studentkinja i 70 studenata. Pitanje je glasilo:

Da li ste zadovoljni svojim materijalnim položajem. A) da B) ne.

Rezultati ankete su bili da je 20 studentkinja odgovorilo sa (da), a 40 sa (ne), 20 studenata je odgovorilo sa (da), a 50 sa (ne).

Postavljena je nulta hipoteza tj. ne postoji statistički značajna razlika između studentica i studenata u pogledu mišljenja o ličnom standardu.

Postavljenu hipotezu testirati na nivo značajnosti $p = 0.05$ i nivou značajnosti $p = 0.01$.

Rezultati ankete prikazani su u tabeli br.4

Tabela br.4

odgovori	studentkinje	studenti	ukupno
da	20 (A)	20 (B)	40 (A+B)
ne	40 (C)	50 (D)	90 (C+D)
ukupno	60 (A+C)	70 (B+D)	130 N

Iz kontigencijskih tabela (2 x 2) Hi-kvadrat test se može izračunati po skraćenom postupku na osnovu slijedeće formule:

$$\chi^2 = \frac{N(IAD - BCI)^2}{(A+B)(C+D)(A+C)(B+D)}$$

Nakon uvrštavanja vrijednosti u formulu dobijamo da je vrijednost Hi-kvadrat testa = 0.34.

Stepeni slobode za tabele (2 x 2) i veće izračunava se po formuli: **(broj redova - 1) x (broj kolona - 1)**. U primjeru imamo:

$$df = (2-1) \times (2-1) = 1$$

Na osnovu prethodnog slijedi:

Hi-kvadrat = 0.34; df = 1; gv(0.05) = 3.84; gv(0.01) = 6.63; NS

S obzirom da su granične vrijednosti na oba nivoa veće od vrijednosti Hi-kvadrata, u konkretnom slučaju prihvatamo nultu hipotezu, koja je glasila da ne postoji statistički značajna razlika u mišljenjima o ličnom standardu između studentkinja i studenata.

Hi-kvadrat test u slučajevima kada imamo više nezavisnih uzoraka

U velikom broju istraživanja susrećemo se sa oblicima testiranja hipoteza kad nas dihotomizacija ne zadovoljava iz prostog razloga što podjele na da ili ne, odnosno ima ili nema, dobro ili loše su vrlo grube. Uvođenjem više od dvije mogućnosti odgovora po istom pitanju dobijamo tabele veće od (2 x 2).

Primjer:

Pitanje za roditelje i učenike neke škole bilo je slijedeće:

Da li smatrate da u školi treba proučavati vjeronauku ?

A) da, B) ne, C) ne znam.

Pretpostavljamo da se mišljenja roditelja i učenika statistički značajno ne razlikuju. Rezultati istraživanja prikazani su u tabeli br.5

Tabela br.5

odgovori	da	ne	ne znam	ukupno
roditelji	200	100	100	400
učenici	50	250	100	400
ukupno	250	350	200	800

Postupno izračunavanje Hi-kvadrata prikazano je u tabeli br.6

Tabela br.6

$(fo - ft)^2 : ft$

fo	ft	(fo-ft)	(fo-ft) ²	(fo-ft) ² :ft
200	125	75	5625	45
100	175	-75	5625	32
100	100	0	0	0
50	125	-75	5625	45
250	175	75	5625	32
100	100	0	0	0
800	800	0	-	154

Hi-kvadrat = 154, df = 2, gv(0.01) = 9.21, S.

Na osnovu dobijenih rezultata odbacujemo nultu hipotezu sa faktorom greške manjim od 1 % odnosno $p < 0.01$. Dakle komentar bi bio kratak: mišljenja roditelja i učenika se bitno razlikuju po pitanju izučavanja vjeronauke u školama. U dalje komentare se ne bi upuštali jer se radi o fiktivnim vrijednostima. Iz priložene tabele da se primijetiti da su mišljenja učenika i roditelja podudarna u odgovoru ne znam.

Očekivane frekvencije u tabelama većim od (2 x 2) izračunavaju se po formuli:

$$(\text{suma reda} \times \text{suma kolone}): N$$

Primjer:

Ako imamo opaženu frekvenciju u prvoj ćeliji tabele br.6 koja iznosi:

$f_o=200$, tada je $f_t=125$ jer je :

$$(400 \times 250): 800 = 125$$

Istom metodologijom moramo izračunati očekivane frekvencije za svaku opaženu frekvenciju.

Napomena:

Hi-kvadrat test ima i izvjesnih ograničenja posebno kada je riječ o kontingencijskim tabelama gdje broj frekvencija u nekoj ćeliji pada ispod 10. U takvim slučajevima uvode se dodatne korekcije takozvane Yatesove korekcije kako bi umanjili ukupnu veličinu HI-kvadrata, a time povećali vjerovatnoću da nulta hipoteza ostane na snazi. Suviše mali broj opaženih odnosno očekivanih frekvencija u nekoj ćeliji dovodi u pitanje smislenosti izračunavanja Hi-kvadrat testa npr: ako se frekvencije kreću ispod 5.

Hi-kvadrat test ima primjenu u specijalnim oblicima korelacije kao što je Kontingencijski korelacijski koeficijent (C), zatim u testiranju normalnosti distribucije, testiranju razlike medijana, takozvani medijan-test, itd. Upravo zbog širokog spektra upotrebe treba strogo voditi računa da ne bi ušli u greške kod primjene ovog testa. Posebna opreznost je potrebna u slučajevima donošenja zaključaka na osnovu Hi-kvadrata. Kako ova metoda zauzima bitno mjesto u praksi i teoriji, detaljnije informacije o ovoj metodi mogu se naći u poznatim udzbenicima statistike od: Andersona, Guilforda, Edwardsa, Krkovića, Ivanovića, Mijanovića, Pecca, Snedekora, Sigela, Žižićke itd.