Characteristics of categorical data analysis

Zuvic-Butorac, Marta

Kako su rezultati svih mjerenja međusobno zavisni (u međuvremenu postoji djelovanje terapije), odlučujemo se za primjenu Cochraneovog Q testa. U danom primjeru (iako zbog kratkoće navoda nisu navedeni konkretni rezultati svakog mjerenja), rezultat analize bio je Q = 79,3, a pripadni $P < 0,001$. Interpretaciju dobivene statističke značajnosti obavili bismo po konkretnom uvidu u status promijenjenih nalaza.

Literatura:

1. Glantz SA. Primer of Biostatstics. McGraw-Hill, 2002.
2. Riffenburgh RH. Statistics in Medicine, Academic Press, 1999.
3. Motulsky H. Intuitive Biostatistics.Oxford University Press, 1995.
4. Arsham A. University of Baltimore. Dostupno na URL adresi: http://home.ublat.edu/ntsbarh/stat-data/Topics.htm. Datum pristupa informaciji: 4. studenog 2005.

CHARACTERISTICS OF CATEGORICAL DATA ANALYSIS

Marta Žuvić-Butorac
Technical Faculty University of Rijeka, Rijeka, Croatia

Summary

This work presents characteristics of categorical data, their presentation and possible models of statistical analysis. There are two types of categorical data; nominal, where categories are equally valued (i.e. we can measure them only in terms of whether individual items belong to some distinctively different categories) and ordinal, where categories allow to rank order on some scale of measurement. Categorical data are non-numeric by nature, but could be numerically presented and analyzed anyway. Although the information value of categorical data is well below the respective information value of numerical, practically there's no study where their analysis wouldn't be of importance. Numerically, the categorical data can be presented via frequencies (absolute number of items belonging to the category) or their proportion (percentage) in the sample. Adequate graphical presentation goes with pie charts or percentage stacked bars charts. The statistical analysis of categorical data most often is done on contingency tables. The type of the analysis depends on the relation between samples from which the data are drowning. If the samples are independent, the analysis would be performed using difference of proportion test, χ^2 test or Fisher exact test. The limitations and suitability for application of the three is discussed. If the samples are dependent, the choice goes to McNemar χ^2 test (2 samples) or Cochrane's Q test 8more than 2 samples). The conclusions from the aforementioned analyses could be drowned only in terms of significant or nonsignificant relations between the rows and columns in the contingency tables. In the need to measure the level of relations between categorical data, two types of measures are defined: relative risk and odds ratio, which can both be calculated only in 2x2 contingency tables. Relative risk is a ratio of two proportions (suitable only in prospective studies), whether odds ratio measures ratio between odds in two groups (could be calculated both in prospective and retrospective studies). All the aforementioned analyses are well documented with calculations on data collected in biomedical studies.

Key words: categorical data, frequency, proportion, contingency table, χ^2 test, Fisher exact test, McNemar χ^2 test, Cochrane's Q test, relative risk, odds ratio

OSOBITOSTI USPOREDBE NEBROJČANIH PODATAKA

Marta Žuvić-Butorac,
Akademija informatičkih tehnologija, Tehnički fakultet Sveučilišta u Rijeci, Rijeka, Hrvatska

Sažetak

U radu se obrazlažu osobitosti nebrojčanih podataka, njihov prikaz te mogući modeli statističkih analiza primjerenih određenim situacijama. Nebrojčani ili kategorički podaci izražavaju se opisnim kategorijama, koje mogu biti jednakovrijedne (nominalni podaci) ili rangirane na određenoj ljestvici vrijednosti (ordinalni podaci). Iako je informacijska vrijednost nebrojčanih (kategoričkih) podataka manja od informacijske vrijednosti numeričkih, gotovo da ne postoji istraživanje u kojem ne bismo bili suočeni s neophodnošću njihove analize. Nebrojčani se podaci numerički prikazuju frekvencijom (apsolutnim brojem jedinki određene kategorije u uzorku) ili njihovim udjelom (proporcijom, postotkom) kojim se javljaju u uzorku. Grafički ih možemo predstaviti kružnim ili stupčastim dijagramima koji veličinom odsječaka ili udjela u stupcu prikazuju odnose kategorija. Analiza nebrojčanih podataka najčešće se radi proučavanjem kontingencijske tablice. Vrsta statističkog modela za analizu podatka ovisi o odnosu uzoraka iz kojih dobivamo nizove podataka. Radi li se o nezavisnim uzorcima, nebrojčane podatke analiziramo testom proporcija, χ^2 testom ili Fisher egzaktnim testom. Radi li se o zavisnim uzorcima, upotrijebit će se McNemarov χ^2 test ili Cochraneov Q test. Trebamo li analizirati i mjere za veličinu povezanosti određenih kategorija (a ne samo razinu statističke značajnosti povezanosti), upotrijebit će se izračun relativnog rizika ili omjera izgleda, ovisno o tipu istraživanja. Sve statističke metode potkrijepljene su analizom konkretnih primjera, na podacima prikupljenim u biomedicinskim istraživanjima.

Ključne riječi: kategorički podaci, frekvencija, proporcija, kontingencijska tablica, χ^2 test, Fisherov egzaktni test, McNemarov χ^2 test, Cochraneov Q test, relativni rizik, omjer izgleda

Adresa autora:

dr. sci. Marta Žuvić-Butorac
Akademija informatičkih tehnologija
Tehnički fakultet Sveučilišta u Rijeci,
Vukovarska 58
51000 Rijeka
Hrvatska

E-mail: martaz@riteh.hr

Od podataka prema znanju – zašto trebamo statističku analizu podataka?

Statistika je znanost koja potpomaže proces donošenja odluka (primjene znanja) u uvjetima neodređenosti, koristeći određen skup metoda za prikupljanje, analizu, prikaz i interpretaciju podataka. No prikupljeni podaci tek su ogoljene informacije koje ne predstavljaju znanje. Kako od podataka stići do znanja? Temeljni je kvalitativni niz: podatak → informacija → činjenica → znanje. Informacija je temeljni element komunikacije znanja. *Podaci* postaju *informacije* tek u trenutku kada su relevantni za proces odlučivanja. *Informacija* postaje *činjenica*, kada je potkrijepljena zaključcima analize podataka. *Činjenice* postaju *znanje* kada su iskorištene u uspješnom dovršenju procesa odlučivanja, potkrijepljenom pravilnom interpretacijom rezultata primijenjenog statističkog modela. Pri tom je konačno primijenjeno znanje izraženo zajedno s uvjetom neodređenosti (razinom statističke pouzdanosti). Kako se povećava razina točnosti (egzaktnosti) statističkog modela, tako se poboljšava kvaliteta procesa donošenja odluka (slika 1.). Zbog toga je u nizu *od podataka prema znanju* potrebno primijeniti statističku analizu, koja omogućava da se znanje temelji na dokazima potkrijepljenim podacima, a ne na uvjerenjima, mišljenjima ili vjerovanjima.

Slika 1. Ilustracija elemenata i kvalitete procesa konstrukcije statističkog modela temeljenog na podacima koji služi odlučivanju u uvjetima neodređenosti (varijabilnosti podataka).

Tri su temeljne faze procesa u kojem se od prikupljenih podataka kroz statističku analizu dolazi do znanja o procesima ili mehanizmima koje istražujemo:

1. *opis (deskripcija) skupa podataka*, na ovom stupnju pokušavamo iz prikupljenih podataka u uzorku doći do opisa procesa koji je generiran tim podacima (opisivanje i prikaz raspodjela podataka koje su od interesa i njihovih međusobnih relacija).

2. *objašnjenje procesa analizom podataka*; na ovom stupnju pokušavamo zaključiti o općim karakteristikama procesa, iako nemamo na raspolaganju sve moguće podatke o procesu (testiranje hipoteze, zaključivanje od uzorka na populaciju).
3. *predviđanje karakteristika budućeg skupa podataka*; u ovoj fazi, temeljem rezultata prethodnih faza, tražimo mogućnost predviđanja karakteristika procesa na novoprikupljenim podacima. Ovakvo predviđanje omogućuje poboljšanje procesa odlučivanja i temeljeno je na egzaktnim relacijama uz kvantificirane uvjete neodređenosti.

Pri korištenju statističke analize potrebno je podacima pristupati kritično, kako bi argumentirano mogli uposliti statističko modeliranje realnih podataka. Analitičko razmišljanje zahtijeva jasnoću prikaza, konzistentnost, dokaz i sljedbeno fokusiranje na problem koji se istražuje.

Vrste podataka - kakvi su to *nebrojčani* podaci?

Podatke možemo razlikovati temeljem njihove informacijske vrijednosti ili kvalitete, koja ovisi o pouzdanosti i preciznosti mjerenja kao i njihovoj mjernoj ljestvici. Specifično, podaci se mogu klasificirati u dvije temeljne skupine: *brojčani* (numerički) i *nebrojčani* (kategorički). Njihovo imenovanje proizlazi iz jednostavnog razloga što se brojčani podaci prirodno prikupljaju i zapisuju u obliku brojeva, dok se nebrojčani izražavaju opisnim kategorijama. Razlikovanje ovih skupova podataka omogućava lakše snalaženje u izboru statističkih modela koji će se primijeniti u analizi te je stoga ovdje dan njihov kratak opis.

Brojčani (numerički) podaci mogu biti *kontinuirani* ili *diskretni*. Kontinuirani podaci mogu imati bilo koju vrijednost (u nekom rasponu) i izraženi su kao decimalni brojevi, dok diskretni podaci imaju cjelobrojne vrijednosti. Tipični primjeri kontinuiranih podataka koje, primjerice, skupljamo na nekom uzorku bolesnika jesu vrijednosti indeksa tjelesne mase bolesnika, vrijednosti koncentracije šećera u krvi, vrijednosti krvnog tlaka, itd. Diskretni su pak podaci prebrojavanja – za određenog ispitanika to može biti broj djece, broj napadaja, broj primljenih transfuzija krvi i slično.

Nebrojčani (kategorički) podaci mogu biti *ordinalni* i *nominalni*. Ordinalni podaci raspoređeni su na ljestvici vrijednosti i među njima se može uspostaviti odnos. Primjeri ordinalnih podataka su stupanj karcinoma, socioekonomski status ispitanika, itd. Važno je uočiti da među ordinalnim podacima postoji uređenje, no odnosi pojedinih vrijednosti na ljestvici ne mogu se kvantificirati (ljestvica ne omogućava izražavanje primjerice dvostruko manjeg podatka). Nominalni podaci su podaci koje uobičajeno zapisujemo riječima i koji u primjeru skupine ispitanika, dijele skupinu na podskupine s obzirom na neko izraženo svojstvo. To mogu biti, primjerice, spol, pripadnost rasi, krvna grupa, kvalitativna vrijednost nalaza (pozitivan/negativan), odjel na kojem je hospitaliziran bolesnik, itd. Valja uočiti da među nominalnim podacima nema uređenja, već su svi jednakovrijedni.

Nebrojčane podatke lako možemo kvantificirati uvodeći određeno pravilo ili dogovor. Tako, primjerice, možemo dogovoriti da stupanj karcinoma označimo brojevima od 1 do 4, pozitivan i negativan nalaz s 1 i 0.

Informacijska vrijednost podataka opada od brojčanih prema nebrojčanim podacima. Brojčani podaci mogu se prevesti u kategoričke (npr. krvni tlak kao skup kontinuiranih brojčanih podataka možemo prevesti u ordinalne, definiranjem kategorija: snižen, normalan, povišen), dok obrnut slijed nije moguć. Zato brojčani podaci imaju veću informacijsku vrijednost, iako samo snižavanje njihove vrijednosti kategorizacijom može biti vrlo korisno kod sumarnih analiza podataka. Imajući ovo na umu jasno je da je pri prikupljanju podataka od kritične važnosti pohraniti ih u obliku koji ima najveću moguću informacijsku vrijednost, koja se kasnije po potrebi lako prevodi u druge oblike.

Ovdje prikazana sistematizacija vrsta podataka nije jedinstvena, niti su pojedine vrste podataka međusobno isključive. Ovakva je podjela dogovorna i služi samo kao pomoć pri odlukama o odabiru prikaza i analize podataka.

Prikazivanje nebrojčanih podataka

Budući da su nebrojčani podaci zapravo kategorije unutar promatrane skupine koje imaju neko zajedničko odabrano obilježje, njihovu kvantifikaciju izražavamo *udjelom* (*proporcijom*) ili *postotkom* kojim je obilježje zastupljeno u uzorku.
Udio ili proporciju računamo kao količnik broja jedinki s odabranim obilježjem i ukupnog broja jedinki. Može poprimiti vrijednosti iz intervala [0,1]. Primjerice, prebrojavanjem je utvrđeno da u skupini 211 ispitanika ima 136 žena i 75 muškaraca. Pripadni udjeli (proporcije) spolova su:

$$P_{žene} = 136/211 = 0{,}645$$
$$P_{muškarci} = 75/211 = 0{,}355$$

Postotak kojim je obilježje zastupljeno u uzorku određujemo kao udio pomnožen brojem 100. Na istom primjeru,

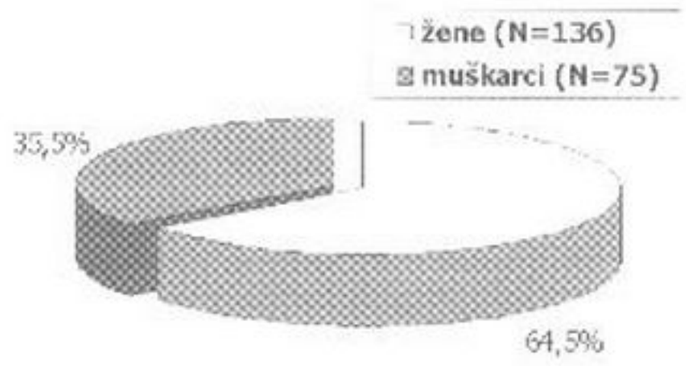
$$P_{žene} = 0{,}645 \cdot 100 = 64{,}5\ \%$$
$$P_{muškarci} = 0{,}355 \cdot 100 = 35{,}5\ \%.$$

Ovi podaci bit će čitljiviji prikažemo li ih tablicom (tablica 1.)

Tablica 1. Raspodjela ispitanika prema spolu u ispitivanoj skupini.

Spol	N	P	P/%
žene	136	0,645	64,5
muškarci	75	0,355	35,5
ukupno	211	1,000	100,0

Dakako, odnos između udjela muškaraca i žena bit će uočljiviji na grafičkom prikazu. Za ovaj je primjer prikladan tzv. *kružni* (torta) dijagram. Uočite da je u grafičkom prikazu izgubljen podatak o brojčanom odnosu, koji se uobičajeno ne navodi na kružnom dijagramu. No, odlučimo li se za grafički, a ne tablični prikaz, brojčane odnose možemo navesti opisno u pripadnoj legendi (slika 2.).



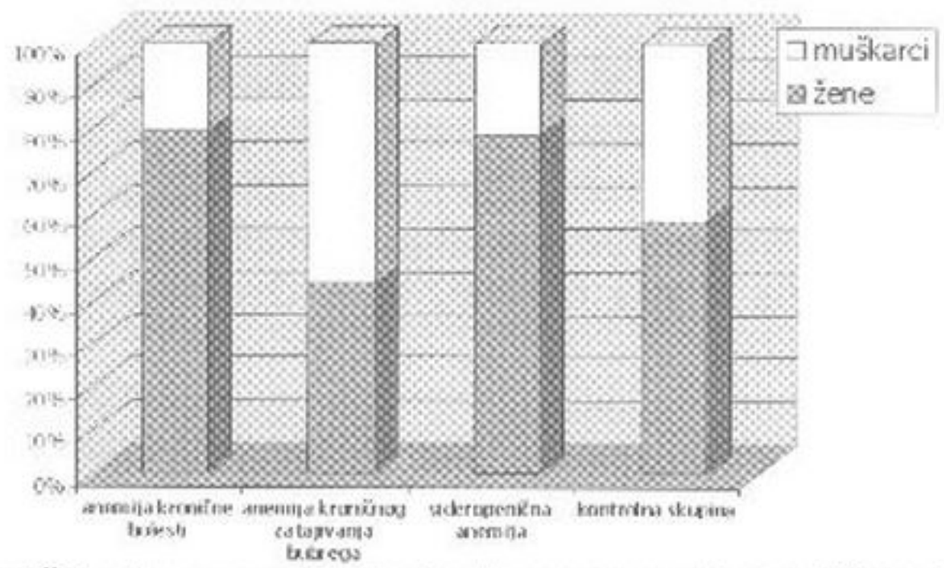
Slika 2. Grafički prikaz raspodjele ispitanika prema spolu u ispitivanoj skupini.

Vrlo često unutar cijele ispitivane skupine javljaju se različita kategorička obilježja prema kojima želimo prikazati podatke. Na istom je primjeru, cijela skupina od 211 ispitanika bila sastavljena od četiri nezavisne skupine: tri ispitivane skupine s različitom dijagnozom (anemija kronične bolesti, anemija kroničnog zatajivanja bubrega, sideropenična anemija) i kontrolne skupine zdravih ispitanika. Želimo li prikazati udjele spolova u svakoj od skupina, ponovno to možemo učiniti tablično ili grafički (tablica 2.).

Tablica 2. Raspodjela ispitanika prema spolu u pojedinim skupinama.

skupina	**žene**		**muškarci**		**ukupno**
	N	P/%	N	P/%	
Anemija kronične bolesti	35	79,5	9	20,5	44
Anemija kroničnog zatajivanja bubrega	24	44,4	30	55,6	54
Sideropenična anemija	41	78,8	11	21,2	52
Kontrola	36	59,0	25	41,0	61
ukupno	136	64,5	75	35,5	211

Odnosi spolova u pojedinim skupinama bit će uočljiviji na grafičkom prikazu pomoću stupčastog dijagrama postotaka (slika 3.).



Slika 3. Grafički prikaz raspodjele ispitanika prema spolu u različitim skupinama.

Kako analizirati prikupljene podatke?

Pri istraživanju postavlja se hipoteza kojoj želimo istražiti vrijednost. Istraživanje vrijednosti postavljene hipoteze provodi se analizom prikupljenih podataka, koja pak podrazumijeva pronalaženje odnosa među nizovima podataka. U biomedicinskim istraživanjima uobičajeno odnose među nizovima podataka istražujemo u četiri osnovna koraka:

1) Formulacija hipoteze. Hipotezu možemo postaviti na dva načina:

   a) nul-hipoteza H_0: opažene razlike među podacima su slučajne. Ova se formulacija u smislu statističkog izražavanja odnosa daje formulacijom da uzorci (iz kojih potječu nizovi podatka) pripadaju istoj populaciji.
   b) alternativna hipoteza H_a: opažene razlike su stvarno postojeće. Ova se formulacija u smislu statističkog izražavanja odnosa daje formulacijom da uzorci (iz kojih potječu nizovi podatka) pripadaju različitim populacijama.

2) Odabir statističkog modela/testa za analizu odnosa među podacima.

3) Izračun *granica pouzdanosti* (engl. *confidence limits,* CL) dobivenog rezultata i *razine statističke značajnosti P* koju daje statistički model za usporedbu odnosa među nizovima podataka.

4) Usporedba izračunate razine statističke značajnosti P sa zadanom razinom statističke značajnosti α. Ako je $P < \alpha$ odbacuje se nul-hipoteza, odnosno za $P > \alpha$ prihvaća se alternativna hipoteza. Standardno se u biomedicinskim istraživanjima vrijednost α postavlja na 0,05 (ova vrijednost odgovara razini mogućeg pogrešnog zaključka od 5 %).

Nizove podataka koje bilježimo, najčešće želimo uspoređivati među skupinama za koje pretpostavljamo da ne dolaze iz iste populacije, odnosno predstavljaju različite uzorke. Prema svojem odnosu, takvi uzorci mogu biti *nezavisni* ili *zavisni.* Najjednostavnije praktično pravilo za određivanje odnosa uzoraka koje analiziramo, jest traženje odgovora na pitanje: da li postoji uzročno posljedična veza između vrijednosti podataka različitih uzoraka? U biomedicinskim istraživanjima odgovor se najčešće svodi na razmatranje dvije moguće situacije:

1. dvije ili više neovisnih skupina jedinki (ljudi, životinja, stanica,...) na kojima se prikupljaju podaci. Ove skupine smatramo *nezavisnim* uzorcima.
2. dvije ili više skupina mjerenja na istim jedinkama (ljudima, životinjama, stanicama,...) na kojima se u međuvremenu desila neka promjena (djelovanje lijeka, napredak bolesti, promjena uvjeta okoline, itd.). Ove skupine smatramo *zavisnim* uzorcima.

Kada odredimo odnos između uzoraka čije nizove podataka želimo uspoređivati, grubi odabir statističkog modela analize ili vrste statističkog testa lako ćemo učiniti pomoću orijentacijskog prikaza u sljedećoj tablici.

Tablica 3. Orijentacijski prikaz za odabir statističkog modela analize podataka

VRSTA PODATAKA	ODNOS UZORAKA				
	NEZAVISNI UZORCI		ZAVISNI UZORCI		POVEZANOST PODATAKA
	2 skupine različitih jedinki	3 ili više skupina različitih jedinki	prije i poslije određenog postupka na istim jedinkama	višestruko ponavljana mjerenja na istim jedinkama	
BROJČANI	t –test za nezavisne uzorke	analiza varijanci ANOVA	t-test za zavisne uzorke	ANOVA za ponavljana mjerenja	Linearna/nelinearna regresija, Pearson korelacija
ORDINALNI	Mann-Whitney U-test	Kruskal-Wallis test	Wilcoxon test	Friedman ANOVA	Spearman korelacija rangova
NOMINALNI	Test razlike proporcija ili χ^2 test / Fisher egzaktni test		McNemar test	Cochrane Q test	Omjer izgleda (OR) Relativni rizik (RR)

Budući da se u ovom radu raspravlja o osobitostima usporedbe nebrojčanih podataka, sljedeće će poglavlje govoriti o nekim specifičnostima usporedbi nominalnih podataka.

Kako analizirati i što zaključivati iz analize nebrojčanih podataka?

Dvije su veličine temeljni rezultat svake statističke analize: granice pouzdanosti (CL) i razina statističke značajnosti (P). Ove dvije veličine su komplementarne i vrlo često se računaju u paru. Granice pouzdanosti omogućavaju da se rezultat izrazi pomoću granica pogreške, koje se standardno u biomedicinskim se istraživanjima računaju kao rubne vrijednosti 95 % intervala pouzdanosti. 95 % interval pouzdanosti predstavlja kontinuirani niz vrijednosti koje bi izračunat statistički rezultat mogao poprimiti kada bi uzorak bio dobiven slučajnim uzorkovanjem iz iste populacije. Razina statističke značajnosti P pak govori mnogo preciznije o sigurnosti s kojom donosimo zaključke; izrazimo li P u postotcima (P), 100 % P daje sigurnost s kojom donosimo zaključak o odbacivanju/prihvaćanju hipoteze. Također, slično kao i u slučaju granica pouzdanosti, P daje postotak istraživanja koji bi dali suprotan zaključak kada bi uzorak bio dobiven slučajnim uzorkovanjem iz iste populacije.

Test razlike proporcija

Promotrimo sljedeći primjer. U randomiziranom dvostruko slijepom prospektivnom istraživanju na skupini od 328 bolesnika, 168 dobiva novi lijek, a 160 placebo.

Promatra se napredovanje bolesti. U "placebo" skupini, kod 45 bolesnika se zamjećuje daljnji napredak bolesti, dok se u skupini "novi lijek" napredak bolesti zamjećuje kod 27 bolesnika. Postavlja se pitanje: zaustavlja li novi lijek napredovanje bolesti?

Postavljamo nul-hipotezu: *Razlika u udjelima* bolesnika kod kojih se zamjećuje napredak bolesti je *slučajna* (alternativna hipoteza glasila bi: udio bolesnika kod kojih se zamjećuje napredak bolesti u skupini "novi lijek" značajno je manja od istovrsnog udjela u skupini "placebo").

1. Odabiremo test kojim ćemo provjeriti hipotezu. Budući da ove dvije skupine predstavljaju nezavisne uzorke, upućeni smo na testiranje razlike proporcija (tablica 3). Odrede se proporcije ($p_1 = 45/160 = 0,28$; $p_2 = 27/168 = 0,16$; razlika proporcija $\Delta p = 0,12$).
2. Test razlike proporcija (kojeg možemo odraditi u bilo kojem od dostupnih aplikacija za statističku obradu podataka unošenjem p i N za svaku od skupina[2]) daje nam izračunate granice pouzdanosti razlike proporcija te vrijednost razine statističke značajnosti razlike proporcija. Granice pouzdanosti razlike proporcija iznose 0,075-0,165, a razina statističke značajnosti razlike proporcija P = 0,009.
3. Usporedbom s $\alpha = 0,05$ ($P < \alpha$) dolazimo do zaključka da je nul-hipotezu potrebno odbaciti, odnosno prihvatiti alternativnu hipotezu.

Važno je uočiti da je već sam izračun granica pouzdanosti govorio o statistički značajnoj razlici u udjelima bolesnika kod kojih napreduje bolest u dvije skupine. Budući da 95 % interval pouzdanosti razlike udjela ima donju granicu 0,075, pa *ne sadržava nulu* (0 = "nema razlike"), očigledno je da je razlika značajna na razini manjoj od 5 %. Iz izračunate vrijednosti razine statističke značajnosti (P = 0,009) zaključujemo da je opažena razlika u proporcijama značajna na razini od 0,009. Interpretacija ovog zaključka bila bi da sa sigurnošću od 99,1 % možemo tvrditi da je udio bolesnika kod kojih napreduje bolest manji u skupini koja dobiva novi lijek.

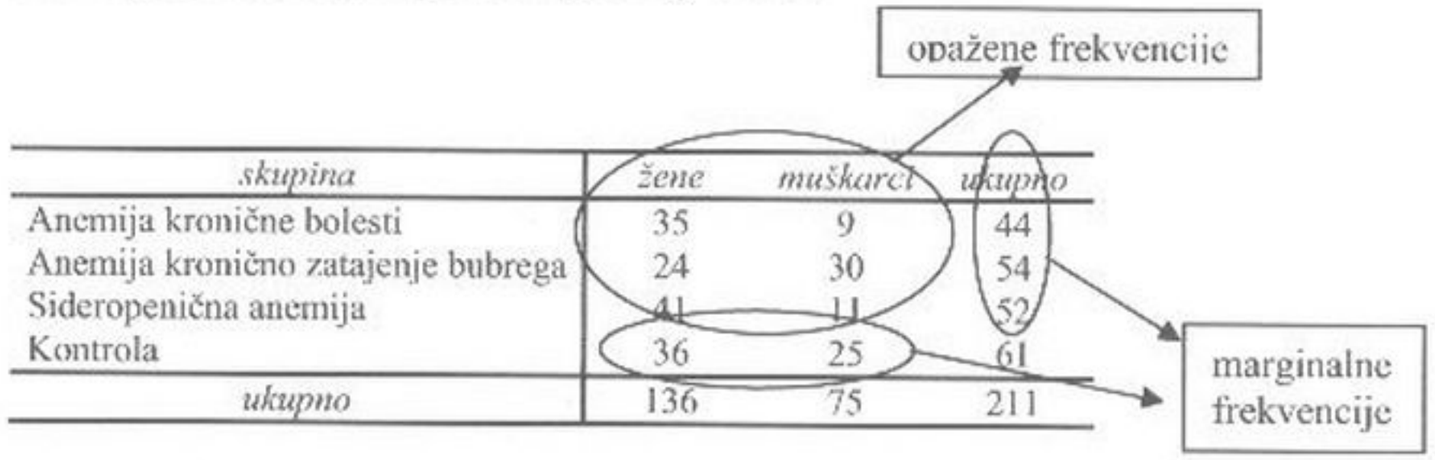
Izračun razine statističke značajnosti dozvoljava još jednu interpretaciju, s obzirom na broj mogućnosti u kojima bi se mogao javiti suprotan zaključak. Tako u ovom slučaju vrijednost P = 0,009 možemo interpretirati na sljedeći način: u 0,9 % jednakih istraživanja može se očekivati da se udio bolesnika kod kojih napreduje bolest u skupini koja dobiva novi lijek ne razlikuje od udjela u skupini koja prima placebo.

χ^2 test

Iz tablice 3. vidimo da je za testiranje razlike udjela iz prethodnog primjera bilo moguće upotrijebiti i χ^2 test. No testom razlike proporcija mogli smo uspoređivati samo vrijednosti udjela u dvije skupine. χ^2 test pruža mogućnost testiranja razlika među više skupina istovremeno, što ćemo pokazati na primjeru tablice 2.

[2] Računski dio testa razlike proporcija provodi se u koracima: 1) traženje razlike proporcija $\Delta p = p_1 - p_2$; 2) računanje standardne devijacije razlike proporcija kao $S_{\Delta p} = \text{sqrt}(p_1 q_1/N_1 + p_2 q_2/N_2)$, gdje su $q_1 = 1-p_1$ i $q_2 = 1-p_2$; 3) računanje 95 % granica pouzdanosti razlike proporcija kao $\Delta p \pm 1,96 \cdot S_{\Delta p}$ i računanje razine statističke značajnosti kao $p = \Delta p / S_{\Delta p}$.

χ^2 testom provodi se testiranje razlike *opaženih* i *očekivanih frekvencija* u *kontingencijskoj tablici*. Kontingencijska je tablica ona kod koje su kategorije u redovima i stupcima međusobno *isključive*, što znači da jedinka iz uzorka može biti pribrojana u samo jednu od ćelija tablice, a ukupni zbroj redaka i stupaca mora biti jednak broju jedinki u uzorku. Iz navedenog slijedi da je tablica 2. kontingencijska, čime je zadovoljen osnovni uvjet primjene χ^2 testa.



Slika 4. Obilježavanje opaženih i marginalnih frekvencija u kontingencijskoj tablici.

Broj žena i muškaraca u pojedinim skupinama predstavljaju *opažene frekvencije* (učestalosti) promatranih obilježja. Što su očekivane frekvencije? Očekivane frekvencije se računaju temeljem nul-hipoteze, koja bi u ovom slučaju glasila:

Raspodjela spolova ne razlikuje se među navedenim skupinama.

Očekivane frekvencije tada možemo izračunati kao frekvencije koje bismo očekivali u istom uzorku kada ne bi bilo razlika u učestalosti žena i muškaraca među pojedinim skupinama. Tablica očekivanih frekvencija izračuna se temeljem *marginalnih frekvencija* u kontingencijskoj tablici (slika 4.). Primjerice očekivane frekvencije žena i muškaraca u skupini anemija kronične bolesti računamo tako da uzimamo iste udjele muškaraca i žena koji se pojavljuju u cijeloj skupini (136 naprama 75, odnosno kao postotni udjeli 64,5 % i 35,5 %). Dobivamo za te ćelije (u retku anemija kronične bolesti):

$$f_{očekivano}\,(\textit{žene}) = 0{,}645 \cdot 44 = 28{,}4 \text{ i } f_{očekivano}\,(\textit{muškarci}) = 0{,}355 \cdot 44 = 15{,}6.$$

Na isti način popunjavaju se i ostale ćelije očekivanih frekvencija. Uočite da su u tablici očekivanih frekvencija marginalne frekvencije istih vrijednosti kao i u tablici opaženih frekvencija.

Tablica 4. Opažene i očekivane frekvencije za kontingencijsku tablicu 2.

skupina	$f_{opažene}$		*ukupno*	$f_{očekivane}$		*ukupno*
	žene	*muškarci*		*žene*	*muškarci*	
Anemija kronične bolesti	35	9	44	28,4	15,6	44
Anemija kronično zatajenje bubrega	24	30	54	34,8	19,2	54
Sideropenična anemija	41	11	52	33,5	18,5	52
Kontrola	36	25	61	39,3	21,7	61
ukupno	136	75	211	136	75	211

Izračunava se vrijednost $\chi^2 = \Sigma [(f_{opažena} - f_{očekivana})^2 / f_{očekivana}]$, što će u ovom primjeru biti prikazano tablicom.

Tablica 5. Izračun χ^2 vrijednosti za podatke iz tablice 4.

$f_{opažena}$	$f_{očekivana}$	$f_{opažena} - f_{očekivana}$	$(f_{opažena} - f_{očekivana})^2 / f_{očekivana}$
35	28,4	6,6	1,53
9	15,6	-6,6	2,79
24	34,8	-10,8	3,35
30	19,2	10,8	6,08
41	33,5	7,5	1,68
11	18,5	-7,5	3,04
36	39,3	-3,3	0,28
25	21,7	3,3	0,50
			$\chi^2 = 19,25$

Za izračunati χ^2 određuje se pripadna razina statističke značajnosti iz raspodjele χ^2 vrijednosti (statističke tablice, odnosno izračun programske aplikacije za statističku obradu) koja u gornjem primjeru iznosi $P = 0,0002$, odnosno $P < 0,001$. Zaključujemo, sa sigurnošću od 99,9 % da se nul-hipoteza mora odbaciti, što znači da s istom sigurnošću tvrdimo da spolovi nisu jednoliko raspodijeljeni u skupinama.

Uočite da sama konačna vrijednost χ^2 i njegove pripadne razine statističke značajnosti (koju nam najčešće aplikacija za statističku obradu podataka daje kao gotov rezultat), ne upućuje nas na nikakve zaključke o tome gdje su razlike u raspodjeli spolova najveće, a gdje su moguće slučajne. Takve je zaključke moguće donositi uočavanjem vrijednosti u tablici 5. Najveći doprinos vrijednosti χ^2 dolazi iz ćelija skupine anemija s kroničnim zatajivanjem bubrega, pa možemo ukazati da se ova raspodjela značajno razlikuje od ostalih raspodjela spolova, pri čemu je u ovoj skupini značajno manje žena. Najmanja razlika dolazi iz kontrolne skupine, što navodi na zaključak da je kontrolna skupina dobro uravnotežena s obzirom na zastupljenost spolova, u odnosu na ukupan broj ispitivanih bolesnika, no same se skupine bolesnika značajno razlikuju s obzirom na zastupljenost spolova. Pojedinačne usporedbe udjela spolova među skupinama i njihove razine statističke značajnosti mogu se dodatno odrediti usporedbom proporcija.

Iako χ^2 test izgleda vrlo široko primjenjiv (može se primijeniti na kontingencijske tablice bilo koje veličine pa čak i u više dimenzijskom prostoru kategorija) i jednostavan kao statistički model istraživanja povezanosti u kontingencijskim tablicama (rezultat analize lako je interpretirati), ipak je pri njegovoj upotrebi potrebno imati na umu da se za pouzdanost rezultata moraju poštovati *osnovni uvjeti primjene*, a to su:

- podaci moraju biti dobiveni slučajnim izborom
- podaci moraju činiti kontingencijsku tablicu
- frekvencije ne smiju biti malene (preporuka $N > 20$, sve $f_{očekivane} > 5$)

- uzorci iz kojih se uspoređuju podaci moraju biti nezavisni.

Rješenje problema zbog premalenog broja podataka u pojedinim ćelijama daje tzv. Yates korekcija χ^2 vrijednosti[3], no u uvjetima malenih frekvencija u ćelijama, preporuka je umjesto χ^2 testa upotrijebiti *Fisherov egzaktni test*.

Fisherov egzaktni test

Fisherov egzaktni test primjenjiv je na kontingencijske tablice koje čak u pojedinim ćelijama mogu imati i frekvencije jednake nuli, no moguće ga je primijeniti samo na tablice veličine 2 x 2 (dva retka i dva stupca). Da bismo vidjeli kako se povezanost redaka i stupaca testira ovim statističkim modelom, pogledajmo sljedeći primjer. U istraživanju rezultata primjene nove terapije kod neke bolesti, bilježene su nuspojave koje su u skupini od 40 bolesnika uz primjenu nove terapije uočene kod 1 bolesnika, dok je u skupini od 60 bolesnika liječenih standardnom terapijom zabilježeno 4 bolesnika s nuspojavama. Postavlja se pitanje: da li nova terapija značajno smanjuje učestalost nuspojava?

Nul-hipoteza jest da nema razlike u učestalosti nuspojava kod dvije vrste terapija. Za istraživanje vrijednosti nul-hipoteze, ponovno sastavljamo kontingencijsku tablicu, no i bez sastavljanja tablice jasno je da se radi o vrlo malim frekvencijama, koje ne dozvoljavaju upotrebu χ^2 testa. Fisherov test upošljava konstrukciju svih mogućih tablica koje imaju jednake marginalne frekvencije kao i tablica koju testiramo. Uočite da je *broj mogućih tablica istih marginalnih frekvencija = najmanja marginalna frekvencija + 1.* U ovom je primjeru najmanja marginalna frekvencija 5, pa ćemo uz zadanu kontingencijsku tablicu pokušati konstruirati još 6 tablica koje će imati jednake marginalne frekvencije kao i osnovna. Redoslijed sastavljanja i numeriranja tih tablica (numeriranje ide od 0 nadalje) mora slijediti pravilo da Tablica 0 započinje s frekvencijom 0 u gornjoj lijevoj ćeliji (tablica 6.)

Tablica 6. Kontingencijska tablica (za primjer u tekstu) i šest pripadnih tablica jednakih marginalnih frekvencija.

Terapija	*Nuspojave*		*ukupno*
	da	ne	
Nova	1	39	40
Standardna	4	56	60
ukupno	5	95	100

Tablica 0			Tablica 1			Tablica 2		
0	40	40	1	39	40	2	38	40
5	55	60	4	56	60	3	57	60
5	95	100	5	95	100	5	95	100
Tablica 3			Tablica 4			Tablica 5		
3	37	40	4	36	40	5	35	40
2	58	60	1	59	60	0	60	60
5	95	100	5	95	100	5	95	100

Za svaku od tablica može se izračunati vjerojatnost njezina pojavljivanja p_x kao

$$p_x = \frac{r_1!s_1!r_2!s_2!}{N!a!b!c!d!}$$

[3] Yates korekcija χ^2 vrijednosti određuje se primjenom formule $\chi^2 = \Sigma\,[(f_{opažene} - f_{očekivane}|- 0{,}5)^2 / f_{očekivane}]$.

gdje su *a, b, c, d* frekvencije, r_1, s_1, r_2, s_2 marginalne frekvencije, a *N* ukupni broj podataka u tablici[4].
Razina statističke značajnosti Fisherova testa određuje se kao zbroj svih vjerojatnosti pojavljivanja tablica redom do opažene tablice:

$$p=p_0+p_1+\ldots+p_{OT}$$

što je u našem primjeru zbroj vjerojatnosti pojavljivanja Tablice 0 i Tablice 1. Dobivamo vrijednost P = 0,645.

Slijedom standardnog načina zaključivanja, budući da je izračunata razina statističke značajnosti veća od 0,05, zaključujemo da se nul-hipoteza prihvaća, odnosno da nema razlike u učestalosti nuspojava pri korištenju nove terapije u odnosu na standardnu.
Fisherov test dobio je naziv egzaktni, zbog izravnog računanja razine statističke značajnosti, koja se u svim ostalim testovima određuje posredno, iz raspodjele vrijednosti parametra korištene statistike. Iako ovaj test zahtijeva mnogo više vremena za računanje, zbog brzine današnjih osobnih računala i pouzdanosti rezultata, Fisherov egzaktni test preporučljivo je provoditi za sve slučajeve tablica 2x2 gdje je $N \leq 100$.

Povezanost podataka u tablicama 2x2

Testovi kao što su χ^2 ili Fisherov egzaktni test daju nam kao rezultat razinu statističke značajnosti za povezanost kategorija u recima i stupcima, no ne govore ništa o veličini te povezanosti. Kao mjere za povezanost podataka u tablicama 2x2 često se određuju *relativni rizik* (engl. *relative risk*, RR) i *omjer izgleda* (engl. *odds ratio*, OR). Obje mjere zapravo predstavljaju *omjere* relativnih udjela, za razliku od prijašnjih testova u kojima su se promatrale *razlike* udjela pojedinih kategorija.

Relativni rizik

Relativni rizik je mjera za povezanost kategoričkih podataka koja se definira kao omjer vjerojatnosti događaja u ispitivanoj skupini i vjerojatnosti događaja u kontrolnoj skupini.

$$RR = \frac{p_i}{p_k} = \frac{\textit{vjerojatnost (incidencija) događaja u } \textbf{\textit{ispitivanoj}} \textit{ skupini}}{\textit{vjerojatnost (incidencija) događaja u } \textbf{\textit{kontrolnoj}} \textit{ skupini}}$$

Iako bismo matematički ovakve omjere mogli izračunati za velik broj situacija, relativni rizik ima smisla računati samo na podacima dobivenim iz prospektivnih kliničkih istraživanja (primjerice istraživanja tijeka bolesti uz primjenu novog lijeka u ispitivanoj skupini i placeba u kontrolnoj) ili epidemioloških istraživanja (primjerice usporedbi incidencija bolesti u skupini koja je izložena nekom rizičnom faktoru i skupini koja nije).

[4] Značenje matematičke oznake a! (čitamo a faktorijela) je a!= a·(a-1)·(a-2)...3·2·1

Vrijednost relativnog rizika lako se izračuna prema navodu uz tablicu 7. Kako interpretirati vrijednost relativnog rizika? Ako je *RR*=1 znači da su vjerojatnosti ili incidencije događaja jednake u obje skupine, pa je i rizik za događaj jednak u obje skupine. Ako je *RR* < 1, rizik događaja u ispitivanoj skupini smanjen je u odnosu na kontrolnu skupinu, odnosno ako je *RR* > 1, rizik je u ispitivanoj skupini povećan. Sama vrijednost *RR* govori nam o tome *koliki će postotak* ispitanika u ispitivanoj skupini imati isti ishod kao u kontrolnoj skupini, te *koliko puta* je rizik u ispitivanoj skupini manji/veći od rizika u kontrolnoj. Da bismo ustvrdili da li je vrijednost *RR*≠1 statistički značajna na razini manjoj od 5 %, potrebno je izračunati i 95 % granice pouzdanosti za *RR*. Računanje 95 % granica pouzdanosti relativnog rizika nije jednoznačno definirano – postoji nekoliko metoda računanja kojima se one mogu približno odrediti[5]. Važno je da interval pouzdanosti *ne sadrži* broj 1, jer je tada osigurano da s 95 % sigurnošću možemo tvrditi da rizik u ispitivanoj skupini nije jednak kao u kontrolnoj.

Tablica 7. Opći format kontingencijske tablice za određivanje relativnog rizika kao $RR = p_i/p_k$, gdje su p_i= A/A+B, a p_k = C/C+D.

skupina	REZULTAT (bolest, reakcija,...)		*ukupno*	*Vjerojatnost razvoja bolesti, reakcije,...*
	+	-		
ispitivana	A	B	A+B	p_i = A/A+B
kontrolna	C	D	C+D	p_k = C/C+D
ukupno	A+C	B+D	A+B+C+D	$RR = P_i/P_k$

Pokušajmo odrediti *RR* i njegove granice pouzdanosti na primjeru koji je obrađen kod testa razlike proporcija. Sastavimo kontingencijsku tablicu u kojoj možemo odmah računati relativni rizik.

Tablica 8. Kontingencijska tablica rezultata prospektivnog istraživanja utjecaja novog lijeka na napredak bolesti.

liječenje	Napredak bolesti		*ukupno*	*P*
	da	ne		
novi lijek	27	141	168	p_i =0,16
placebo	45	115	160	p_k = 0,28
ukupno	72	256	328	***RR* =0,57**

Zaključujemo:

1. Izračunati *RR* < 1 (*RR* = 0,57), dakle rizik napredovanja bolesti *smanjen je* liječenjem novim lijekom.
2. *RR* = 0,57 što znači da će 57 % bolesnika liječenih novim lijekom imati isti ishod kao da su primali placebo.

[5] Kako relativni rizik ne može biti manji od 0, a može poprimiti vrlo velike vrijednosti, za očekivati je da granice pouzdanosti nisu simetrične s obzirom na vrijednost *RR*. Međutim, logaritamska vrijednost granica pouzdanosti daje približno simetričan interval, pa je jedan od najjednostavnijih načina za ovaj izračun korištenje Katzove jednadžbe ln(*RR*) ± 1,96·sqrt [(B/A)/(A+B) + (D/C)/(C+D)] iz koje se antilogaritmiranjem dobivaju 95 % granice pouzdanosti.

3. Omjer vjerojatnosti skupine "placebo" prema skupini "novi lijek" je $1/RR = 1,75$ što znači da će bolesnici koji primaju placebo imati 1,75 puta vjerojatniji napredak bolesti nego bolesnici liječeni novim lijekom.

Granice pouzdanosti za *RR* računate Katzovom metodom u ovom primjeru iznose 0,37-0,87, što znači da postoji barem 95 % sigurnost u pouzdanost zaključaka 1.-3., jer interval pouzdanosti ne uključuje broj 1.

Važno je uočiti da zaključci koji se izvode iz dobivene vrijednosti *RR* ovise o postavljanju omjera vjerojatnosti. Zato pri izračunu *RR* treba voditi računa o smislenosti postavljenog omjera i interpretaciji njegove vrijednosti.

Omjer izgleda

Za razliku od relativnog rizika, koji se računa u prospektivnim istraživanjima, omjer izgleda primjereno je računati u retrospektivnim istraživanjima u kojima se analiziraju odnosi broja *događaja* u *izloženoj* i *neizloženoj* skupini. *Izloženost* podrazumijeva izlaganje nekom utjecaju, rizičnom čimbeniku, postupku, a *događaj* podrazumijeva slučaj (jedinku) koji razvija određeni status (bolest, reakciju...).
Omjer izgleda je mjera za povezanost kategoričkih podataka koja se definira kao omjer izgleda za događaj u izloženoj skupini i omjer izgleda za događaj u neizloženoj skupini.

$$OR = \frac{i_i}{i_n} = \frac{\textit{izgledi za događaj u } \textbf{\textit{izloženoj}} \textit{ skupini}}{\textit{izgledi za događaj u } \textbf{\textit{neizloženoj}} \textit{ skupini}}$$

Izgled za događaj je omjer broja slučajeva u skupini koji su razvili određeni status i onih koji nisu. Jednostavniji prikaz vidimo u kontingencijskoj tablici 9.

Tablica 9. Opći format kontingencijske tablice za određivanje *omjera izgleda* kao $OR = i_i / i_k$, gdje su i_i= A/B, a i_k = C/D.

UTJECAJ, IZLOŽENOST	REZULTAT (bolest, reakcija,...)		*ukupno*	*Izgledi* *za razvoj bolesti, reakcije...*
	+	-		
da	A	B	A+B	i_i=A/B
ne	C	D	C+D	i_k=C/D
ukupno	A+C	B+D	A+B+C+D	$OR = i_i / i_k$

Kako u retrospektivnim istraživanjima broj događaja ovisi o odabiru istraživača, jasno je da u ovakvom istraživanju ne bi bilo korektno računati relativni rizik. Situacija je upravo obrnuta u prospektivnim kliničkim ili epidemiološkim kohortnim istraživanjima, gdje istraživač odlučuje koliko će jedinki/ispitanika biti u skupinama koje se promatraju. Stoga se omjer izgleda smije računati u oba tipa istraživanja (i prospektivnim i retrospektivnim), no relativni rizik isključivo u prospektivnim

istraživanjima. Omjer izgleda po iznosu je približno jednak relativnom riziku kada su incidencije događaja ("+" rezultat) malene u odnosu na ukupni broj ispitanika[6].
Kako interpretirati vrijednost omjera izgleda? Ako je *OR* = 1 znači da su izgledi za razvoj događaja jednaki u obje skupine. Ako je *OR* < 1, izgledi za događaj u izloženoj skupini manji su od izgleda za događaj u neizloženoj skupini i obratno ako je *OR* > 1, izgledi za događaj su u izloženoj skupini veći nego u neizloženoj skupini. Sama vrijednost *OR* govori nam o tome koliko je događaja u izloženoj skupini na jedan događaj u neizloženoj. Da bismo ustvrdili da li je vrijednost *OR* ≠ 1 statistički značajna na razini manjoj od 5 %, potrebno je izračunati i 95 % granice pouzdanosti za *OR*. Kao i kod relativnog rizika, za računanje 95 % intervala pouzdanosti postoji nekoliko metoda računanja[7]. Važno je da interval pouzdanosti *ne sadrži* broj 1, jer tada s 95 % sigurnošću možemo tvrditi da su izgledi u izloženoj skupini različiti od izgleda u neizloženoj.

Tablica 10. Kontingencijska tablica rezultata retrospektivnog istraživanja incidencije karcinoma pluća kod pušača i nepušača.

liječenje	Karcinom pluća		*ukupno*	*izgledi*
	da	ne		
pušači	153	347	500	i_1 =0,44
nepušači	15	685	700	i_k = 0,02
ukupno	168	1032	1200	***OR* = 22**

Zaključujemo:
1. *OR* > 1 omjer izgleda za razvoj karcinoma pluća je velik kod pušača prema nepušačima
2. *OR* = 22 na 22 oboljela pušača dolazi 1 oboljeli nepušač
Račun 95 % intervala pouzdanosti prema Woolfovoj metodi[6] daje za ovaj primjer 11,6 - 34,5 pa zaključke 1. i 2. izražavamo s barem 95 %-nom sigurnošću, jer interval pouzdanosti ne sadrži broj 1.

McNemarov i Cochraneov Q test

U svim dosadašnjim razmatranjima promatrane skupine bile su *nezavisne*. Kao što ćemo vidjeti u ovom razmatranju, do sada obrađeni modeli statističkih analiza dali bi potpuno pogrešan rezultat u slučaju *zavisnih* uzoraka iz kojih dobivamo podatke. Promotrimo to na primjeru. U istraživanju utjecaja hiposenzibilizacijske terapije, 50 osoba podvrgnuto je alergijskom testu prije i poslije terapije. Dobiveni su rezultati prikazani tablicom 11. Postavlja se pitanje: da li terapija utječe na rezultat alergijskog testa?

---

[6]Matematički: Za A<<A+B i C<<C+D, vrijedi A+B≈B i C+D≈D, pa je *RR*= (A/A+B)/(C/C+D) ≈ (A/B)/(C/D) = *OR* . U istraživanju opisanom u tablici 8. mogli smo izračunati i omjer izgleda koji bi iznosio *OR*=0,19/0,39=0,49. Ta je vrijednost dosta blizu vrijednosti relativnog rizika *RR*=0,57, iako incidencije događaja nisu bile jako malene u odnosu na ukupni broj ispitanika u odgovarajućoj skupini.
[7]Kao i kod relativnog rizika, omjer izgleda ne može biti manji od 0, a može poprimiti velike vrijednosti, pa granice pouzdanosti nisu simetrične s obzirom na vrijednost *OR*. Međutim, logaritamska vrijednost granica pouzdanosti daje približno simetričan interval, pa je jedan od najjednostavnijih načina za ovaj izračun korištenje Woolfove metode: ln(*OR*) ± 1,96·sqrt [(1/A + 1/B + 1/C + 1/D] iz koje se antilogaritmiranjem dobivaju 95 % granice pouzdanosti.