UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI
FAKULTET
DEPARTMAN ZA MATEMATIKU I
INFORMATIKU

NEPARAMETARSKI TESTOVI U BIOSTATISTICI SA PRIMENOM NA JAVNO ZDRAVLJE

-MASTER RAD-

Mentor:

Prof. dr Zagorka Lozanov Crvenković

Student:

Vesna Živanović

Novi Sad, 2019.

Sadržaj

Predgovor

Potreba za neparametarskom statistikom proizilazi iz činjenice da dobijeni podaci nemaju uvek normalnu raspodelu tako da se metode parametarske statistike ne mogu koristiti.

Sa podacima u numeričkom obliku često se srećemo kako u svakodnevnom životu, tako i u kliničkim i epidemiološkim studijama, u obliku učestalosti (broj pušača,broj oboljenja, broj slučajeva sa simptomima i bez njih i slično). Srećemo se i sa situacijama gde izvorno izmerene kvantitativne podatke, na primer, uzrast, nivo šećera – glikemije u krvi, menjamo u kategorije i zanima nas zastupljenost u kategorijama kao što su starosne kategorije, glikemija- povišena i dotična vrednost.

Obično zbog malog uzorka, suočeni smo sa situacijama kada uslove normalne raspodele ne možemo da ispunimo. U ovom radu baviću se ovakvom vrstom problema i navešću osnovne postupke prilikom njihovog rešenja i interpretacije**.**

Rad je podeljen u tri tematske celine. U prvom delu definisaćemo i pojasniti pojmove statistike i biostatistike.U drugom delu rada upoznajemo se sa osnovnim funkcijama statističkog programa R. U trećoj celini, koja je ujedno i centralni deo rada, bavimo se razlčitim neparametarskim testovima i njihovom upotrebom u biostatistici kroz raznovrsne primere iz prakse.

1. O statistici i biostatistici

Ako se prisetimo uloge statistike kod prepoznavanja uzroka nastanka epidemije kolere i registrovanja obolelih u Londonu, možemo samo doći do zaključka da statistika ima bitnu ulogu i prati problematiku javnog zdravlja već duži period. Navedimo i kao primer značaj statistike u pokušaju određivanja uzroka oboljenja koji su slučajnog karaktera. Slučajnost možemo dokumentovati time što neće svaki kontakt sa faktorom koji izaziva oboljenje i dovesti do manifestacija oboljenja kao u slučaju tuberkuloze, HIV infekcije itd. Veliki broj oboljenja nastaju kombinacijom različitih faktora, a ni jedan od njih ne možemo označiti kao jedini uzrok. Stil života, genetika i životno okruženje zajedno određuju pojavu i razvoj pojedinih oboljenja, kao što su visok krvni pritisak, maligni tumor, šećerna bolest. U svim ovim slučajevima statistika može da odigra ključnu ulogu.

U istraživanjima često radimo sa populacijama[1] koje se sastoje od mnogih pojedinaca, naročito u zdravstvenim. Bez korišćenja određene metode koja ih zajednički opisuje i pomaže da se saznaju njihove sličnosti i razlike, ne možemo ih proučavati,. Ovakve zadatke rešava statistika.

Definicija: **Statistika je nauka koja se bavi rezultatima grupnih posmatranja, njihovo prikupljanje, analiza i primena u donošenju odluka i pretpostavki.**

Statistika se prema oblasti primene deli na:

- matematičku- proučava, razvija statističko znanje i same metode
- primenjenu.

U oblasti javnog zdravlja imamo biomedicinsku statistiku, koja predstavlja primenu statistike u biomedicinskim naukama, kao i u javnom zdravlju. Imamo i zdravstvenu statistiku koja se bavi užom oblašću opisivanja zdravstvenih sistema. Skup statističkih metoda koje se često koriste u epidemiološkim studijama predstavlja epidemiološku statistiku.

Definicija : **Biostatistika koristi instrumente i koncepcije statistike u oblasti medicine i u biološkim naukama.**

Razvoj statistike u javnom zdravlju posebno je podstakla multifaktorijalna priroda nastanka oboljenja, kao i potreba za obradom ne samo jednog para podataka, već celih polja podataka i traženje njihove uzajamne povezanosti. Kako proračune nije bilo moguće raditi ručno, sa razvojem računara, dolazi do njihove sve češće upotrebe u epidemiologiji, kao i u drugim

[1] Predstavlja najveći mogući skup entiteta koji predstavljaju predmet našeg interesovanja u datom momentu

oblastima istraživanja javnog zdravlja. Računari su nam omogućili da sprovedemo studije o povezanosti pušenja i raka pluća[2]. Takođe su omogućili i razvoj logističke regresije, čija je prva primena zabeležena 1961. godine za određivanje povezanosti koronarne bolesti srca, nivoa holesterola i sistolnog krvnog pritiska. [3]

2. Uvod u statistički program R

Program R je jezik i okruženje za statističke proračune i statističko grafičko prikazivanje. Nastao i distribuiran je prema licenci GNU GPL -General Public Licence. To znači da je jezik i okruženje R moguće koristiti, širiti i poboljšavati bez ograničenja.

R je sličan jeziku i okruženju C, odnosno softveru C-plus, koji se prodaje. Sličnost nastaje iz odluke izvornih stvaralaca projekta R, da u njegovom stvaranju pođu od jezika C. Veliki broj zadataka iz jezika C bez promena funkcioniše i u R-u .[4]

Za razliku od drugih programa za statističku analizu, koji broje skup operacija koje su u suštini nepromenljive, R je jako fleksibilan i relativno prilagodljiv za bilo koju specifičnu vrstu analize. Ovaj sistem je zasnovan na paketima, koji po pravilu sadrže komande za statističke operacije u određenoj specifičnoj oblasti. Tako, na primer, paketi *epicalc* sadrže komande namenjene za statističku analizu u epidemiologiji, a paket *survival* sadrži komande namenjene za različite oblike analize preživljavanja u epidemiološkim i zdravstvenim istraživanjima. Instaliranjem programa R, instalira se nekoliko paketa koji sadrže standardna statistička izračunavanja i grafički prikaz – osnovni paket. Svako može napraviti paket komandi i da ga stavi na raspolaganje za korišćenje zajedno sa postojećim paketima, a to sledi na osnovu licence GNU-GPL.

Osnova njegovog funkcionisanja je komandna linija, odnosno red u koji se zadaje komanda, a R izvršava zadato naređenje, i po tome se R razlikuje od klasičnih statističkih programa. Neke od komandi, uprkos svom jednostavnom značenju, će uraditi relativno kompleksna izračunavanja. Kod tipičnih analiza potrebno je zadati sled nadovezujućih komandi i argumenata, da bismo dobilli traženi rezultat. Na ovakav način se napravi script (grupa komandi), na osnovu čega R izvršava tražena naređenja i dolazi do rezultata.

[2] https://www.bmj.com/content/328/7455/1519.full
[3] Vidi [1]
[4] https://www.r-project.org/about.html

Tekstualni editor (R editor) se uključuje za pisanje skripti i omogućava pisanje i ređanje skripti u obliku modela koje možemo bilo kad ponovo otvoriti i koristiti.[5] Softver Tinn-R je grafički i funkcionalno savršenija zamena editoru R. Nastao kao zamena Notepad-u, koji je deo operativnog sistema Windows. Prvenstveno je namenjen da bude editor skripti za okruženje R. Ima mnogo komandi koje pojednostavljuju i čine efektivnijim pisanje skripti. Distribuiran je pod licencom GPL, kao i R.[6]

Program R je za epidemiologiju i istraživanja u javnom zdravlju ili u medicini odlična alternativa za dostupne komercijalne softvere. Postoje mnogo dostupnih paketa komandi napisanih za ovu oblast istraživanja, koje omogućavaju statističku analizu, grafički prikaz podataka i rezultata od jednostavnih deskriptivnih statističkih radnji, pa sve do kompleksnih prognostičkih modelovanja ili postupaka odlučivanja u medicini.[7]

3. Neparametarski testovi

Neparametarski testovi su postupci kod kojih ne pretpostavljamo da su podaci normalno raspodeljeni[8]. Ne pretpostavljamo poznavanje raspodele u populaciji, uopšteno gledano. Zato ih obično nazivamo i statistički testovi bez raspodele (distribution-free statistics). Dve bitne karakteristike po kojima se razlikuju od parametarskih: ne potvrđuju hipotezu o parametrima populacije, a možemo ih koristiti i u slučajevima kada je oblik populacije, iz koje potiče uzorak[9], poznat.

Njihova prednost je sposobnost ispitivanja podataka na nižim skalama – vrednuje se redosled, jednostavnije računanje, što danas, prilikom korišćenja računara i nije tako bitno. Imaju i negativnu stranu, a to je njihova niža snaga. Kada je moguće da biramo između parametarskog i neparametarskog testa, a ujedno postoje uslovi za oba, odlučićemo se za parametarske.[10]

[5] https://www.e-reading.club/bookreader.php/137398/Software_for_Data_Analysis_Programming_with_R.pdf
[6] Vidi [3]
[7] https://tbrieder.org/epidata/course_reading/e_aragon.pdf
[8] Krajem 18. veka nemački matematičar i naučnik Johan Karl Fridrih Gaus (1777–1855) počeo je da koristi krivu u obliku zvona za izražavanje situacije kada su podaci razdeljeni simetrično i ravnomerno oko srednje vrednosti. Krivu možemo da naslikamo tako da pretpostavljamo da je srednja vrednost 0, a standardna devijacija tačno 1. Ova kriva se zove kriva standardne normalne raspodele. Za nju je karakteristično da je simetrična oko srednje vrednosti (označavamo je kao μ (grčko slovo mi), dakle srednja vrednost populacije), kao i da su aritmetička sredina, medijana i mod jednaki.
[9] Iz praktičnih razloga nikad nećemo raditi sa celom populacijom pa iz nje odaberemo samo deo, jedan njen podskup. Bitno da je uzorak izabran na slučajan način.
[10] Vidi [9]

3.1. Skale

Pre nego što pređem na pojedinačne neparametarske statističke testove, potrebno je reći nešto o skalama, jer su neparametarski testovi pogodan instrument za obradu podataka dobijenih na nižim skalama merenja. Postoje četiri osnovne vrste skala, a to su: nominalna, ordinalna, intervalna i odnosna (racio) skala. Ovo je najpoznatija podela, koju je 1946. godine predložio američki psiholog Stanley Smith Stevens (1906 - 1973) u svom radu „On the theory of scales of measurement".

Nominalna skala je najjednostavnija i nastaje priključivanjem imena posmatranja. Ime može biti broj, odnosno cifra. Primer je Međunarodna klasifikacija bolesti, gde svaka dijagnoza ima jedinstvenu kombinaciju slova i brojeva. Dihotomne skale su, na primer zdrav – bolestan, muškarac - žena, pušač – nepušač. Može da bude i više podela, na primer, pušač – bivši pušač – nepušač. Granice između kategorija ne moraju da budu ravnomerne i ne moraju kvantitativno da se upoređuju u smislu veći ili manji.

Ordinalna ili redna skala omogućava svrstavanje objekata prema tome koji imaju više, a koji manje kvaliteta. Ove skale nam ne omogućavaju da kažemo za koliko je to više kvaliteta. Primer je rezultat lečenja: izlečeni bez posledica, izlečeni sa trajnim posledicama, neizlečeni, umrli. Ovu skalu koristimo za vrednovanja socio-ekonomskih karakteristika, kao i zdravstvenog stanja, na primer Glazgovska skala kome (Glasgow Coma Scale) ili skala težine povrede (Injury Severity Scale).

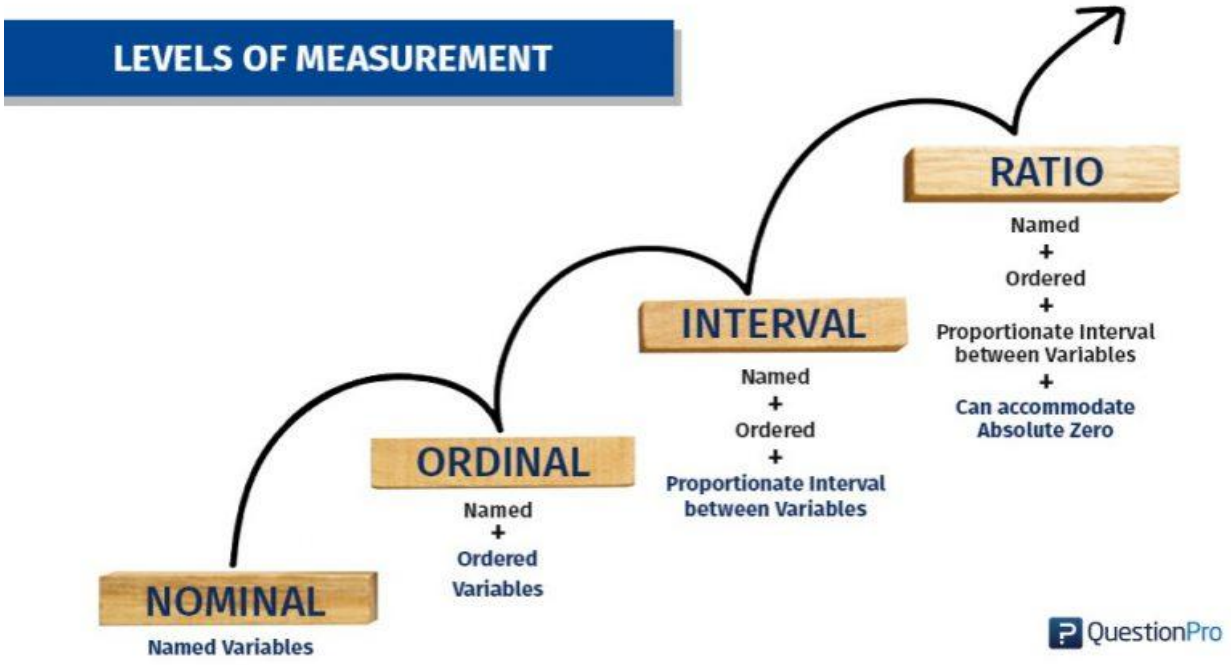
Intervalna skala je preciznija od ordinalne skale jer je poznata udaljenost između dva merenja skale. Skala merenja temperature je tipičan primer ovakve skale. Celzijusova skala počinje sa nulom, što je dogovorena vrednost, isto kao 100 stepeni. Tačke između njih su rastuće temperature, a njihova udaljenost je jednaka i jedinica je jedan stepen Celzijusa. Kako postoje vrednosti manje od 0°C, znamo da i ova vrednost sadrži određen kvantitet i nije krajnja tačka na skali.

Odnosne skale su najinformativnije skale. Sadrže sve karakteristike nominalne, ordinalne i intervalne skale. Sastoje se ne samo od jednako udaljenih tačaka, već sadrže i nultu tačku koja ima značenje. Ako pitamo ispitanike statističkog ispitivanja za uzrast, onda će razlika između dve proizvoljne uzastopne tačke sa sledećim godinama starosti uvek biti jednaka. Nula prikazuje tačku kada se čovek rodio. Možemo da upoređujemo, zato što je sedamdesetogodišnji čovek tačno dva puta starijiod čoveka koji ima 35 godina. Slično je i kod merenja telesne mase i visine.

Ako imamo na raspolaganju male uzorke (n je manje od 30, odnosno 20) i ne možemo da obezbedimo da podaci dolaze iz normalno raspodeljene populacije onda smo prinuđeni da koristimo neparametarske testove.

U praktičnom životu možemo da, na primer, uzrast izrazimo ne samo u nizu pojedinih godina – na intervalnoj skali, već to može da bude i po dekadama. Zato je značajno da se intervali ne poklapaju i da budu identični, a definicija dekada mora da bude jednoglasno usvojena. Na primer, godine od 20 do 29, od 30 do 39 i tako dalje. U ovakvom slučaju smo transformisali naše merenje na nižu skalu – rednu, i ne moramo koristiti neparametrijske testove.

Redosled kojim su navedene skale merenja nije proizvoljno odabran. Naime, svaki naredni tip skale merenja zadržava osobine prethodnog tipa, ali donosi i neke nove osobine.[11]



Poređenja različitih mernih skala

[11] https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/

3.2. Tabele kontigencije

Tabele kontigencije (contingency table) omogućavaju tabelarnu ukrštenu klasifikaciju podataka. To su kombinacije dve (ili više) tabela frekvencije[12]. Omogućavaju saznanje frekencije, broja ispitanika koji odgovara specifičnoj kategoriji za više od jedne promenljive. Ispitivanje ovih frekvencija omogućava saznanje odnosa između promenljivih. Nominalnim promenljivim ili brojčanim promenljivim, koje dostižu relativno mali broj skupova vrednosti, odgovara samo tabela kontigencije. U slučaju da je neophodno koristiti brojčanu promenljivu sa većim brojem dobijenih vrednosti, potrebno je najpre je prekodirati, gde će vrednost promenljive biti jednoznačno dodeljena u neku kategoriju (napr. nizak, srednji, visok).

3.2.1. Tabele kontigencije 2 × 2 tabele

Najjednostavnije tabele kontigencije su 2 × 2 tabele, gde su obe promenljive binarne. Konstrukciju tabele kontigencije možemo videti na primeru (Primer 1).

Tabela 1: Tabela kontigencije 2x2

	C	**D**	**Σ**
A	**a**	**b**	**a+b**
B	**c**	**d**	**c+d**
Σ	**a+c**	**b+d**	**a+b+c+d**

Tabela ima dva reda i dve kolone, obe promenljive su tipa ili/ili. Ili učesnik nesreće nije imao na glavi kacigu ili je imao, ili je završio u komi ili van kome. Kako niko ne može da ima i nema kacigu, to su slučajevi koji se uzajamno isključuju. Imamo sve moguće kombinacije koje stvarno mogu da se dese: sa kacigom i van kome, sa kacigom i u komi, bez kacige i van kome, bez kacige i u komi.

Primer 1: Primer pravljenja tabele kontigencije

Traži se učinak zaštitne kacige kod motociklista koji su imali ozbiljnu saobraćajnu nezgodu. Saznalo se da je 157 od 243 učesnika nezgoda na motorima koristilo kacigu, , a od njih je 33 preživelo ozbiljnu povredu praćenu komom, dok je 65 osoba koje nisu koristile kacigu posle

[12] Tabela frekvencija se konstruiše tako što se raspoređene vrednosti podataka raspoređuju u rastućem redosledu sa odgovarajućim frekvencijama.

nezgode upalo u komu. Da li možemo reći da kaciga štiti učesnike od ozbiljnih povreda mozga praćenih komom?

Tabelu kontigencije pravimo na osnovu gore datih podataka:

	U komi	Van kome	Σ
Kaciga		33	157
Bez kacige		65	
Σ			243

Dopunićemo računanje u redovima i kolonama i napravićemo tabelu kontigencije:

	U komi	Van kome	Σ
Kaciga	124	33	157
Bez kacige	21	65	86
Σ	145	98	243

Iz prethodne tabele još ne znamo odgovor na pitanje iz primera.

Koliko je bilo svih onih koji su nosili kacigu, predstavlja zbir prvog reda, to je 157. Zbir drugog reda, odnosno 86, govori koliko je bilo onih bez kacige. U prvoj koloni su svi koji nisu završili u komi, a u drugoj su svi koji su završili u komi. Pravilo šta treba da bude u redu, a šta u koloni, ne postoji, jedino je potrebno obratiti pažnju na redosled prilikom interpretacije rezultata.

Da bi saznali da li kaciga stvarno čuva onog ko je nosi prilikom udesa na motoru, počinjemo od toga kakva je verovatnoća da učesnik nesreće ima kacigu. Rezultat se predstavlja odnosom onih koji su imali kacigu prema ukupnom broju,[13] a to je $\frac{157}{243} = 0{,}65$. $\frac{145}{243} = 0.60$ je verovatnoća za stanje van kome posle povrede i jednaka je broju svih koji su van kome prema svima u uzorku. Sada se pitamo koja je verovatnoća nositi kacigu i ne upasti u komu. Pošto je kombinacija dve verovatnoće jednaka njihovom proizvodu, u ovom slučaju to bi bilo $0{,}65 \cdot 0{,}60 = 0{,}39$. Koliko je to slučajeva iz celog uzorka saznaćemo tako što broj slučajeva pomnožimo verovatnoćom

[13] Definicija verovatnoće (Pjer Simon de Laplas)

nositi kacigu i ne upasti u komu, odnosno $243 \cdot 0{,}39 = 94{,}77$. Ako bi važile izračunate verovatnoće, imali bismo u prvom kvadratu 94,77 slučajeva, a ne posmatranih 124 slučaja. Ovaj postupak vodi ka očekivanim vrednostima. One se izračunaju ili na ovaj način, ili možemo da koristimo skraćen postupak, gde pomnožimo zbirove u odgovajajućem redu i koloni i podelimo sa celokupnim zbirom.

Formula za izračunavanja očekivane vrednosti za prvu ćeliju, označenu kao (zadržavamo oznake iz tabele 1):

$$a\text{:}\ \frac{(a+b)\cdot\ (a+c)}{a+b+c+d}$$

Uopštena formula glasila bi: očekivana vrednost ćelije jednaka je količniku proizvoda zbira reda i zbira kolone sa ukupnim zbirom.

3.2.2. Tabele kontigencije $m \times n$ tabele

Situacija u kojoj tabela kontigencije ima više od dve kolone ili dva reda nije retka. Postupak je sličan kao i u prethodnom slučaju, s tom razlikom da moramo da koristimo drugu proceduru.[14]

3.3. Mere povezanosti

Mere povezanosti između promenljivih predstavljaju način sumiranja jačine veze između dve ili više promenljive. Za dve promenljive se kaže da su povezane kada poznavanje informacija o jednoj promenljivoj može pomoći u predviđanju vrednosti druge.

Termini korelacija i povezanost se često koriste kao sinonimi, ali neki autori između njih prave razliku. Prema njima, povezanost je širi pojam i obuhvata bilo kakvu vrstu veze između promenljivih, dok se pod korelacijom obično podrazumeva samo linearna veza između promenljivih.

Metode za otkrivanje povezanosti se mogu bazirati na grafičkom prikazu odnosa ili nekom vrstom koeficijenta. Najčešće se u praksi koriste dijagram rasturanja i koeficijent korelacije.

Dijagram rasturanja (eng. scatter plot) je tip matematičkog dijagrama koji se koristi da bi prikazao tipične vrednosti dve promenljive u koordinatnom sistemu. Kreirao ga je britanski statističar Francis Galton 1888. godine da bi prikazao vezu između dve promenljive. Podaci su

[14] Vidi [7]

prikazani kao kolekcija tačaka u koordinatnom sistemu gde su vrednosti promenljivih prikazane na x i y osi.

Dijagram rasturanja omogućava brzo uočavanje povezanosti između promenljivih, jer vizuelno prikazuje jačinu i smer veze, ali ne daje precizne numeričke pokazatelje te povezanosti. S druge strane, koeficijenti su precizniji u kvantifikaciji veze, a njihovo tumačenje je složenije. Zato se ove dve metode često kombinuju, tako što se prvo nacrta dijagram rasturanja, a zatim se ide u precizniju analizu putem nekog od koeficijenata. Izbor metoda identifikacije povezanosti zavisi i od vrste promenljivi čija se povezanost ispituje.

Koeficijenti korelacije uzimaju vrednost između -1 i 1, pri čemu:

- 1 označava snažnu pozitivnu vezu
- -1 označava jaku negativnu vezu.
- 0 ukazuje da nema veze između promenljivih

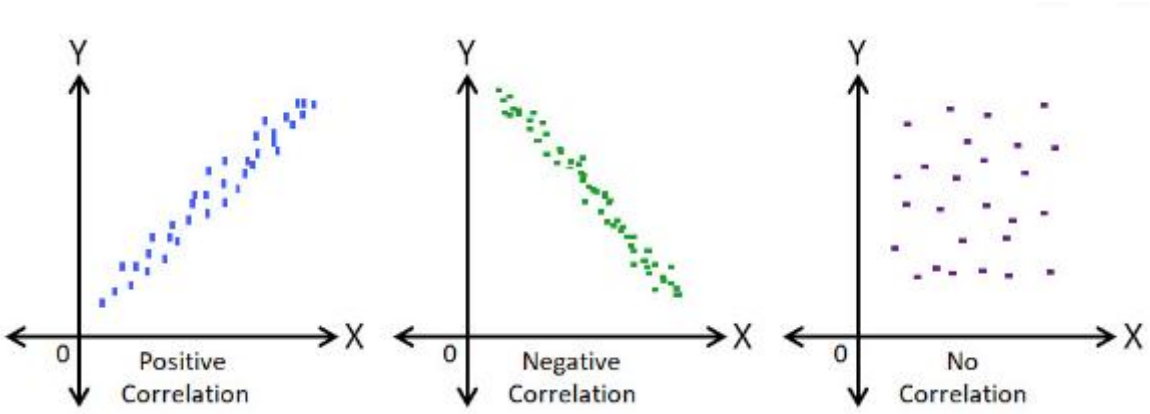
Predznak pokazuje da li je korelacija pozitivna (obe promenljive zajedno i opadaju i rastu) ili negativna (jedna promenljiva opada kada druga raste i obrnuto). Apsolutna vrednost tog koeficijenta (kada zanemarimo njegov predznak) pokazuje jačinu veze.

Koeficijent korelacije 1 znači da za svaki pozitivni porast jedne promenljive, u drugoj postoji pozitivno povećanje određenog udela. Na primer, incidencija ishemijske bolesti srca je u pozitivnoj korelaciji sa mekoćom vode za piće, kako su to pokazale brojne epidemiološke studije. Dakle, što je voda tvrđa, time je veća pojava IBS.

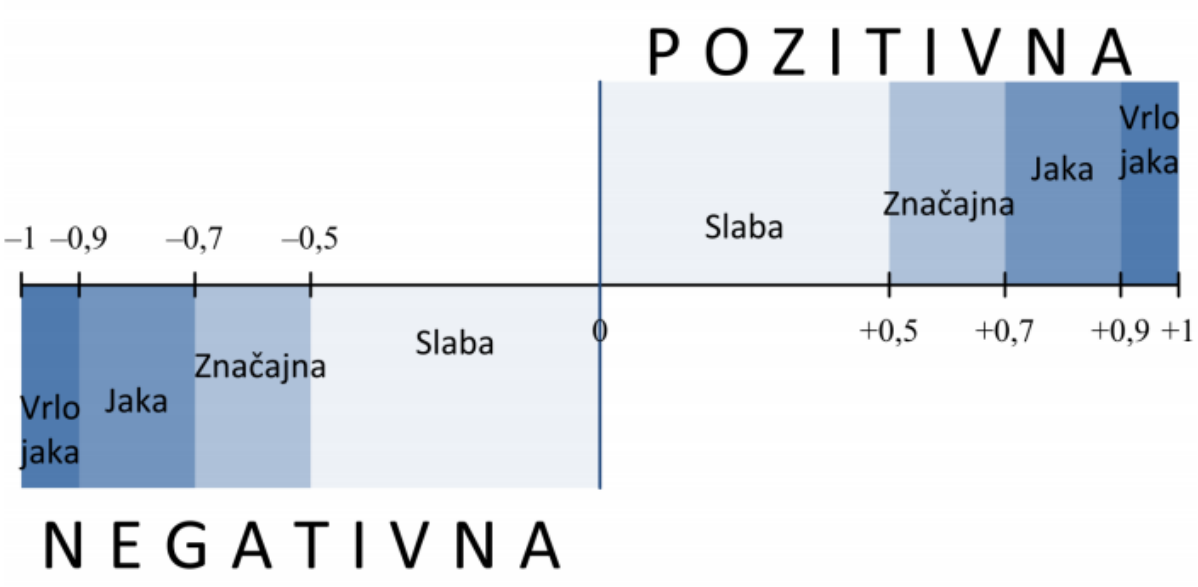
Koeficijent korelacije -1 znači da za svaki pozitivni porast jedne promenljive, ima negativan pad fiksnog udela u drugoj. Na primer, količina infuzione tečnosti u boci smanjuje se u (skoro) savršenoj korelaciji sa brzinom.

Ako izračunat koeficijent korelacije nije jednak ± 1, ali je dovoljno blizu, možemo da kažemo da ova korelacija ukazuje na jaku vezu između dve promenljive. Ako se koeficijent korelacije približava $\pm 0{,}5$, onda će se pre raditi o srednje jakom odnosu. U slučajevima kada se približava nuli, govorimo da je odnos slab ili da odnosa nema. Dijagram rasturanja tada izgleda kao oblak nasumično raspoređenih tačaka.[15]

[15] Vidi [1]



Dijagrami rasturanja



Interpretacija vrednosti koeficijenta povezanosti

$0 < |r| \leq 0{,}5$ Slaba korelaciona veza

$0{,}5 < |r| \leq 0{,}7$ Značajna korelaciona veza

$0{,}7 < |r| \leq 0{,}9$ Jaka korelaciona veza

$0{,}9 < |r| \leq 1$ Vrlo jaka korelaciona veza

3.3.1. Phi koeficijent

Ispitivanje postojanja korelacione veze između dve nominalne promenljive za tabele veličine 2×2 vršimo pomoću Phi koeficijenta korelacije.

Phikoeficijent korelacije izračunava se primenom sledeće formule:

$$\phi = \pm\sqrt{\frac{\chi^2}{N}}$$

gde je χ^2 je izračunat kao u Pirsonovom hi-kvadrat testu, N je ukupan broj posmatranih subjekata.

Jednostavna mera, primenljiva samo na slučaju 2×2 tabela kontigencije, gde $a, b, c\ i\ d$ predstavljaju frekvencije iz tabele 1, tada formula za Phi koeficijent je:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

ϕ se kreće od 0 do 1 ili -1. Znak zavisi od znaka proizvoda glavnih dijagonalnih elemenata tabele tj ad umanjen za proizvod van dijagonalnih elemenata a to je bd.

Primer: Uporedimo ocene (iznad proseka ili ispod proseka) rada 50 apoteka koje daju dve rivalske kompanije za proizvodnju lekova (Galenika i Hemofarm). Hoćemo da utvrdimo da li su njihove ocene međusobno povezane? Da li visoka ocena Galenike za neku od apoteka predviđa visoku ocenu Hemofarma za istu apoteku? Ili je visoka ocena Galenike povezan sa suprotnim ocenom Hemofarma za istu apoteku (što dovodi do negativne fi korelacije)?

		Hemofarm	
		Iznad proseka	Ispod proseka
Galenika	Iznad proseka	6	2
	Ispod proseka	6	36

Rešenje:

```
> Input =("
+  Galenika      Iznad Ispod
+    Iznad         6       2
+    Ispod         6       36
+ ")
> Matrix.2 = as.matrix(read.table(textConnection(Input),
+                                 header=TRUE,
+                                 row.names=1))
> Matrix.2
      Iznad Ispod
Iznad     6     2
Ispod     6    36
>
> library(DescTools)
>
> Phi(Matrix.2)
[1] 0.5211684
```

Vrednost fi koeficijenta je 0,5211684 što znači da veza između ocena dve rivalske kompanije Galenike i Hemofaram je značajna, pozitivna. Visoka ocena Galenike za neku od apoteka predviđa visoku ocenu Hemofarma za istu apoteku.

3.3.2. Kramerov V koeficijent

Ovo je mera povezanosti između dve nominalne promenljive. Kramerov V koeficijent je proširenje Phi koeficijenta na tabele kontingencije koje su veće od 2×2. Najčešće se koristi kada imamo različit broj vrsta i kolona kao što je 2×3, 3×5, itd. U slučaju tabela kontigencije 2×2 Kramerov V je jednak Phi koeficijentu. Ovaj koeficijent je nazvan po švedskom matematičaru i statističaru Haraldu Krameru.

Definisan je kao:

$$V = \sqrt{\frac{\Phi^2}{t}} = \sqrt{\frac{\chi^2}{Nt}}$$

gde je

$t = minimum\,(r - 1, c - 1)$, r je broj vrsta, a c broj kolona

$N -$ veličina uzorka

Ovaj koeficijent uzima vrednosti od 0 (ne postoji veza između dve promenljive) do 1 (potpuna povezanost) i može dostići 1 samo kada su dve promenljive jednake jedna drugoj.

Interpretacija vrednosti Kramerovog koeficijenta:

V=0 – ne postoji veza između promenljivih

V=1 –potpuna povezanost promenljivih

V<0,25 – slaba veza promenljivih

V>0,75 – jaka veza promenljivih

0,25<V<0,75 – značajna veza promenljivih

Sledeći primer nam pokazuje da su Phi koeficijent i Kramerov V koeficijent jednaki za tabele kontigencije 2×2.

Primer: Od 26 studenata medicine koji su pristupili praktičnom delu ispita iz anatomije položilo ih je 10, a od 14 studentkinja polozilo ih je 5.

Pol	Nisu položili	Položili	Ukupno
Muškarci	16	10	26
Žene	9	5	14
Ukupno	25	15	40

Kolika je povezanost pola i uspešnosti studenata medicine na praktičnom delu ispita iz anatomije?

Rešenje:

```
> Input =("
+ Pol      Polozio Pao
+  Musko     16     10
+  Zensko    9      5
+ ")
> Matrix.2 = as.matrix(read.table(textConnection(Input),
+                                 header=TRUE,
+                                 row.names=1))
> Matrix.2
       Polozio Pao
Musko       16  10
Zensko       9   5

> library(DescTools)
>
> Phi(Matrix.2)
[1] 0.0270666
> library(rcompanion)
>
> cramerV(Matrix.2)
Cramer V
 0.02707
```

Primer: Ispitivanje je sprovedeno na uzorku od sedamdesetoro dece osnovnoškolskog uzrasta čiji su roditelji klasifikovani na bogate, srednja klasa i siromašne da bi se utvrdila njihova poseta bolnici (privatna i državna).Podaci su dati u tabeli:

Društveni slojevi	Privatni zdravstveni centar	Državni zdravstveni centar
Siromašni	0	32
Srednja klasa	7	35
Bogati	63	3

Na osnovu prikupljenih podataka želimo da saznamo da li je poseta vrsti zdravstvenog centra nezavisna od društvenog sloja roditelja?

Rešenje:

```
library(lsr)


pb <- c(0, 7, 63)
db <- c(32, 35, 3)
X <- cbind( pb, db )
rownames(X) <- c( 'Siromasni', 'Srednja klasa', 'Bogati' )
print(X)


cramersV( X )
```

```
             pb db
Siromasni     0 32
Srednja klasa 7 35
Bogati       63  3
[1] 0.8668997
```

Vrednost koeficijenta je 0,8668997 to znači da jen veza između društvenog sloja roditelja i poseta vrsti zdravstvenog centra jaka.

3.3.3. Kendalov tau

Ime je dobio po Mauriceu Kendallu, koji ga je uveo 1938. godine. Kendalov tau je neparametarska mera rangiranih (redoslednih) podataka. Koeficijent uzima vrednost od -1 do 1.

Postoji nekoliko verzija Tau:

- Tau-A i Tau-B se obično koriste za kvadratne tabele (sa jednakim kolonama i vrstama).
- Tau-C se obično koristi za pravougaone tabele.

Većina statističkih paketa ima ugrađen Tau-B, ali možemo iskoristiti sledeću formulu da biste ga izračunali ručno:

$$\tau = \frac{C - D}{C + D}$$

gde je C broj skladnih parova i D je broj neskladnih parova.[16]

Za izračunavanje vrednosti Kendalove korelacije koriste se i formule:

$$\tau = \frac{C-D}{\frac{n(n-1)}{2}} \quad \text{ili} \quad \tau = \frac{2(C-D)}{n(n-1)}$$

gde su C i D gore definisani

$n -$ veličina uzorka

Znak od $C - D$ određuje smer veze. Dakle, pozitivna veza će biti kada je vrednost $C - D$ pozitivna, dok se negativna veza javlja kada je vrednost $C - D$ negativna. Kendalov tau koeficijent se kreće od -1 (savršena negativna veza kada je C=0) do 1 (savršena pozitivna veza kada je D=0).

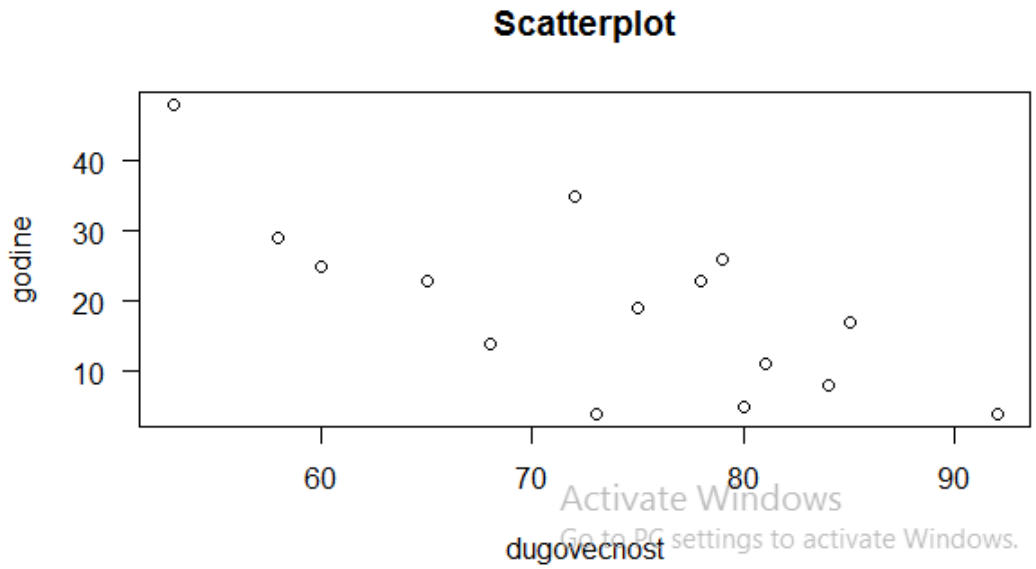
Primer: Želimo da proverimo odnos između pušenja i dugog života. Uzet je uzorak od 15 muškaraca starijih od 50 godina, a prosečan broj pušenih cigareta dnevno i zabeležene su godine smrti. Da li iz uzorka možemo zaključiti da dugovečnost ne zavisi od pušenja?

[16] Neka postoji par zapažanja $-(x_i, y_i)$ i (x_j, y_j):
Par je skladan ako su $x_i > x_j$ i $y_i > y_j$ ili $x_i < x_j$ i $y_i < y_j$
Par je neskladan ako su $x_i > x_j$ i $y_i < y_j$ili $x_i < x_j$ i $y_i > y_j$

Cigarete	Dugovečnost
5	80
23	78
25	60
48	53
17	85
8	84
4	73
26	79
11	81
19	75
14	68
35	72
29	58
4	92
23	65

Rešenje:

```
> plot(dugovecnost, godine, main="Scatterplot", las=1)
```



Dijagram rasturanja prosečanog broja pušenih cigareta dnevno u zavisnosti od godina smrti

```
> godine<-c(5, 23, 25, 48, 17, 8, 4, 26, 11, 19, 14, 35, 29, 4, 23)
> dugovecnost<-c(80, 78, 60, 53, 85, 84, 73, 79, 81, 75, 68, 72, 58, 92, 65)
> cor(dugovecnost, godine, method = "kendall")
[1] -0.5096389
> cor(dugovecnost, godine, method = "spearman")
[1] -0.6744197
```

Kendalov koeficijent ima vrednost -0,5096389 što znači da je korelacija između godina i dugovečnosti značajna i negativna. Muškarci koji su pušili više cigareta dnevno imaju kraći zivotni vek.

3.3.4. Gama koeficijent

Gudman-Kruskalov gama (ili ukratko gama) koeficijent je mera povezanosti između dve ordinalne promenljive. Iako postoje drugi koeficijenti koji mogu izračunati odnose za ove vrste promenljivih, poput Kendalovog tau, uglavnom se koristi kada imamo više redova i kolona. U pogledu svoje interpretacije i računanja sličan je Kendalovom tau.

Gama koeficijent se kreće između -1 i 1, gde je:

- 1 savršena pozitivna korelacija (ako jedna vrednost poraste, onda i druga).
- -1 savršena obrnuta korelacija (kako jedna vrednost raste, tako i druga opada).
- 0 ne postoji povezanost između promenljivih

Izračunavanje gama koeficijenta:

$$\gamma = \frac{N_c - N_d}{N_c + N_d}$$

Gde je N_c je broj skladnih parova, a N_d je broj neskladnih parova.

Primer: Želimo da utvrdimo da li postoji povezanost između fizičke aktivnosti i stepena gojaznosti. 85 ljudi je učestvovalo u istraživanju. Učesnici su svrstani u jedan od četiri nivoa fizičke aktivnosti: nizak, umeren, visok i veoma visok. Učesnike je takođe ocenila medicinska sestra da bi utvrdila njihovu klasifikaciju telesnih masti. Na osnovu ove procene, učesnici su razvrstani u jedan od četiri nivoa: veoma gojazni, gojazni, normalani i mršavi.

		Klasifikacija	telesne	masti	
		veoma gojazan	gojazan	normalan	mršav
	nizak	3	8	10	2
Fizicka aktivnost	umeren	0	11	13	3
	visok	0	3	18	24

Rešenje:

```
> library(readxl)
> Pusenje <- read_excel("Vesna/Pusenje.xlsx",
+     sheet = "Sheet6")
> View(Pusenje)
> myTable<-table(Pusenje$`Fizicka aktivnost`,Pusenje$`Klasifikacija telesne masti`)
>
> gkgamma(myTable)

        Goodman-Kruskal's gamma for ordinal categorical data

data:  myTable
Z = -0.19694, p-value = 0.8439
95 percent confidence interval:
 -0.2968674  0.2426591
sample estimates:
Goodman-Kruskal's gamma
            -0.02710414
```

Gama koeficijent (- 0,02710414) nam govori da postoji slaba, negativna veza između nivoa fizičke aktivnosti i zdravstvenog stanja. To ukazuje da se sa porastom fizičke aktivnosti smanjuju nivo telesne masti.

3.3.5. Spirmanov koeficijent korelacije rangova

Ime je dobio po Čarlsu Spirmanu. Spirmanova korelacija je je neparametarska verzija Pirsonovog koeficijenta korelacije. Označen je grčkim slovom ρ.

Podaci moraju biti ordinalni. Ako se Spirmanovim koeficijentima korelacije želi utvrditi povezanost između dve kvantitativne (numeričke) promenljive prethodno ih je potrebno svesti na ordinalnu skalu, odnosno pripadajuće rezultate transformisati u rangove.

Formula kojom izračunavamo Spirmanov koeficijent korelacija rangova:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

d_i- je razlika rangova

n - broj ispitanika

Spearman vraća vrednost od -1 do 1, gde:

- 1 – savršena pozitivna korelacija
- -1 – savršena negativna korelacija
- 0 – nema korelacije[17]

[17] Vidi [6] i [5]

Primer: U tabeli podaci obuhvataju 23 uzorka podzemnih voda koji su prikupljeni tokom snimanja koncentracija uranijuma i ukupno rastvorene čvrste supstance (mg/L).

	Koncentracija uranijuma	Rastvorena čvrsta supstanca
1	678,1	0,8
2	818,93	1,93
3	302,38	0,97
4	1149,6	11,8
5	573,14	1,41
6	1034,55	2,41
7	633,23	3,4
8	1095,42	0,98
9	1122,58	2,46
10	686,51	0,26
11	1172,84	9,97
12	593,7	0,37
13	1247,95	6,7
14	533,99	0,09
15	605,51	1,72
16	696,96	6,76
17	1282,95	10,27
18	531,16	0,13
19	788,36	2,87
20	956,06	3,1
21	1149,38	0,96
22	1069,82	3,77
23	1124,17	7,09

Da li su dve promenljive povezane?

Rešenje:

Metode za saznanje povezanosti između promenljivih pored dijagrama rasturanja je i boxplot dijagram.

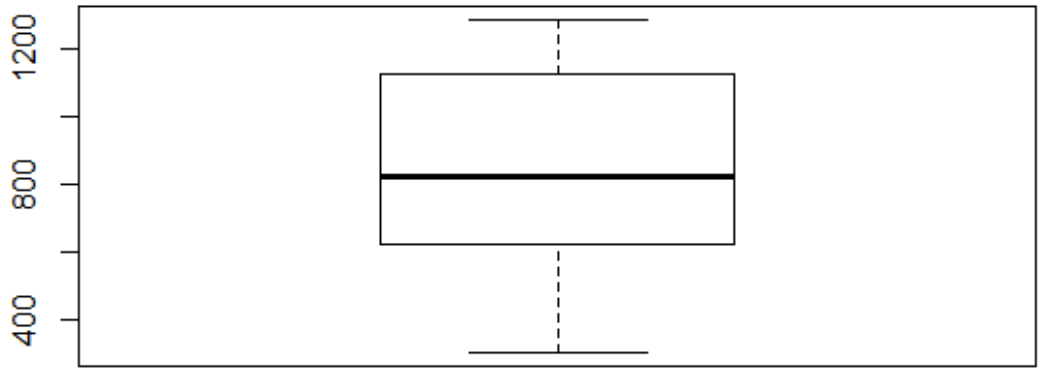
Kvantil: Za p ∈ (0,1) definišemo p kvantil kao

$$q_p = x_{p(n+1))}$$

Donji kvartil je onda 0,25-kvantil, a gornji kvartil je 0,75-kvantil. Medijana je 0,5-kvantil. Kutijasti ili boks dijagram (eng. box and whisker plot) je dijagram koji se sastoji od pravougaonika koji prikazuje podatke od donjeg do gornjeg kvartila. Horizontalna linija po pravougaoniku označava medijanu. Donje i gornje horizontalne linije se nazivaju ”whisker”.

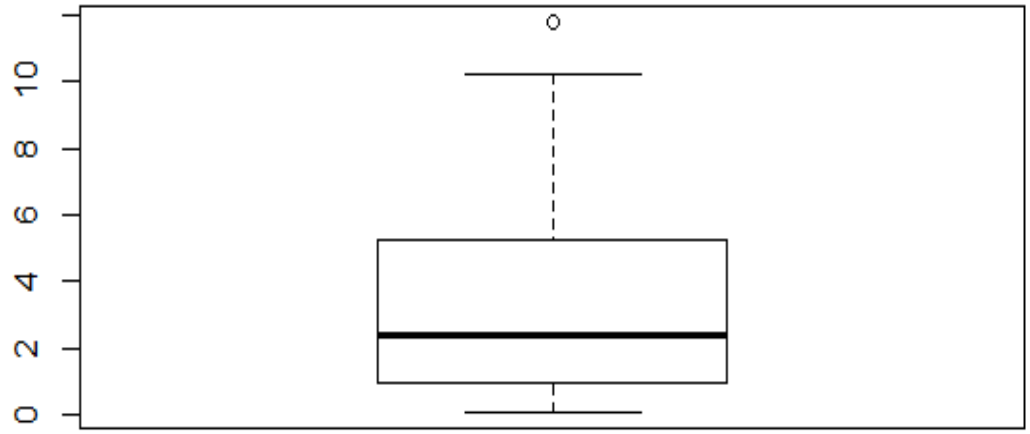
```
> kon.ur<-c(678.1, 818.93, 302.38, 1149.6, 573.14, 1034.55, 633.23, 1095.42, 1122.58, 6
86.51, 1172.84, 593.7, 1247.95, 533.99, 605.51, 696.96, 1282.95, 531.16, 788.36, 956.0
6, 1149.38, 1069.82, 1124.17)
> rcs<-c(0.8, 1.93, 0.97, 11.8, 1.41, 2.41, 3.4, 0.98, 2.46, 0.26, 9.97, 0.37, 6.7, 0.0
9, 1.72, 6.76, 10.27, 0.13, 2.87, 3.1, 0.96, 3.77, 7.09)
> cor.test(kon.ur, rcs, method = "spearman", alternative = "greater")

        Spearman's rank correlation rho

data:  kon.ur and rcs
S = 590, p-value = 0.0001147
alternative hypothesis: true rho is greater than 0
sample estimates:
     rho
0.708498
```

```
> boxplot(kon.ur)
```



Boxplot dijagram za koncentraciju uranijuma

```
> boxplot(rcs)
```



Boxplot dijagram za rastvorenu čvrstu supstancu

Sa boxplot dijagrama za koncentraciju uranijuma vidimo da je medijana prilično blizu centra kutije, a whisker-i su otprilike jednakih dužina, dok sa boxplot dijagrama za rastvorenu čvrstu supstancu imamo medijanu blizu donjeg kvartila i donji whisker je kraći od gornjeg.

Vrednost Spirmanovog koeficijenta korelacije od 0,708498 potvrđuje, što je bilo vidljivo iz boxplot dijagrama, da postoji pozitivna značajna povezanost između koncentracije uranijuma i rastvorene čvrste supstance.

3.3.6. Koeficijent entropije

Koeficijent entropije je nominalna mera povezanosti. Prvi put ga je uveo Henri Teil i zasnovan je na konceptu entropije informacija. Neka imamo uzorke dve diskretne slučajne promenljive[18], X i Y. Izgradnjom zajedničke raspodele, $P(X,Y)$, iz koje možemo izračunati uslovne raspodele, $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$ i $P(Y \mid X) = \frac{P(X,Y)}{P(X)}$, i izračunavanjem različitih entropija, možemo odrediti stepen povezanosti dve promenljive. Entropiju jedne raspodele možemo izračunati:

$$H(X) = -\sum_{x} P_X(x) log\, P_X(x)$$

dok je uslovna entropija data kao:

$$H(X|Y) = -\sum_{x,y} P_{X,Y}(x,y) log\, P_{X|Y}(x|y)$$

Koeficijent entropije definisan je kao

$$U(X|Y) = 2\frac{H(X) + H(Y) - H(X|Y)}{H(X) + H(Y)}$$

Uzima sve vrednosti od 0 do 1.[19]

Primer: Podaci su dati u tabeli:

	Grupa savetovanja za trudnice			
		Jutro	Poslepodne	Veče
Radni status	Zaposlen	3	2	40
	Nezaposlen	12	20	13

[18] Diskretne (prekidne, diskontinuirane) slučajne promenljive se karakterišu prekidima, pauzama, koje indikuju nedostatak vrednosti među određenim vrednostima.
[19] https://www.statisticshowto.datasciencecentral.com/

Da li su dve promenljive povezane?

Rešenje:

```
library(DescTools)

m <- as.table(cbind(c(3,12), c(2,20), c(40,13)))
dimnames(m) <- list(paste("A", 1:2), paste("B", 1:3))
m

# direction default is "symmetric"
UncertCoef(m)
UncertCoef(m, conf.level=0.95)

UncertCoef(m, direction="row")
UncertCoef(m, direction="column")
```

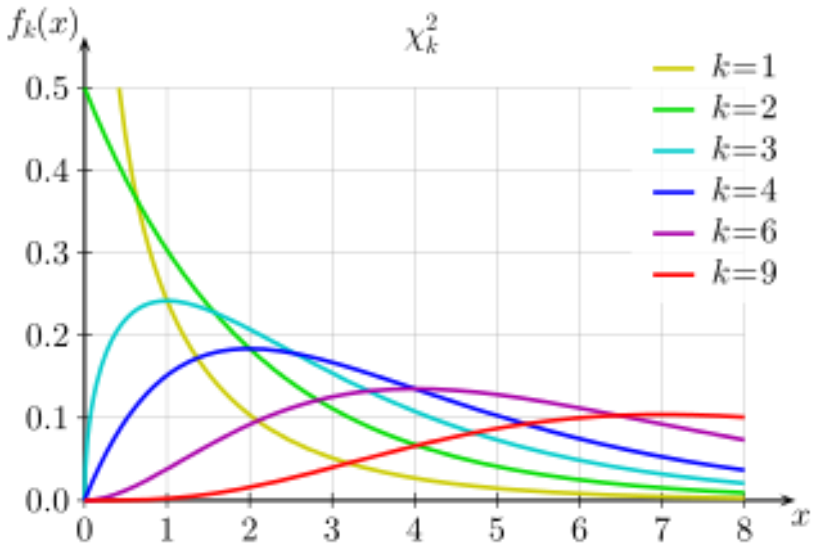
```
    B 1 B 2 B 3
A 1   3   2  40
A 2  12  20  13
[1] 0.2514785
       uc    lwr.ci    upr.ci
0.2514785 0.1135966 0.3893604
[1] 0.2989467
[1] 0.2170192
```

Na osnovu rezultata zaključujemo da se radi o slaboj vezi između promenljivih.

3.4. Hi-kvadrat test

Iz normalne raspodele možemo da izvedemo hi-kvadrat raspodelu, u oznaci χ^2.Za svaki stepen slobode raspodela ima drugačije umeren oblik krive.

Sa slike vidimo da je oblik raspodele za prva dva stepena slobode (k=1 i k=2) značajno različit od ostalih.



Hi kvadrat raspodela za različite stepene slobode

Hi-kvadrat test je zasnovan na upoređivanju testiranih podataka koji imaju hi-kvadrat raspodelu. Kako bi ustanovili da li postoji saglasnost između opaženog (Observed -O) i očekivanog (Expected - E), u praksi se koristi Pirsonovo saznanje da hi-kvadrat raspodelu možemo da koristimo kao test slaganja između posmatranja i hipoteze. Vrednost hi-kvadrat računamo pomoću sledeće formule:

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i}$$

$O_i -$ opažena vrednost i-te ćelije

$E_i -$očekivana vrednost i-te ćelije

Broj stepena slobode, označava se sa df, se izračunava kao proizvod broja vrsta smanjen za jedan i broja kolona isto tako smanjenih za jedan, ili zapisano formulom:

$$df = (r - 1)(k - 1)$$

$r -$ broj vrsta

$k -$ broj kolona

3.4.1. Hi-kvadrat test nezavisnosti

Koristi se za utvrđivanje da li postoji povezanost između dve nominalne promenljive. Hi-kvadrat potvrđuje hipotezu da su dva kriterijuma nezavisna, nasuprot alternative da dva kriterijuma nisu

nezavisna. Ako izračunata vrednost testa dobije ili prekorači kritičnu vrednost za dati broj stepena slobode i verovatnoće, odbacujemo hipotezu o nezavisnosti.

Primer 2: Hi-kvadrat test nezavisnosti i tabele kontigencije $2x2$ u programu R (podaci iz primera 1):

Rešenje:

```
c(124, 33, 21,65)
kaciga <-c(124, 33, 21,65) # ulazni podaci svrstani po redovima
x <- matrix(kaciga, nrow=2, byrow=T) # transformacija na matricu i njeno ispisivanje
 x
# dozivanje komande prop.test()
prop.test(x, alternative = c("two.sided"), conf.level = 0.95, correct =
FALSE)
```

```
[1] 124  33  21  65
     [,1] [,2]
[1,]  124   33
[2,]   21   65

        2-sample test for equality of proportions without continuity
        correction

data:  x
X-squared = 68.738, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.4346911 0.6565546
sample estimates:
   prop 1    prop 2
0.7898089 0.2441860
```

Rezultat testiranja hipoteze na nivou značajnosti 0,95 i 0,99 (p< 0,05 kao i p< 0,01) nam omogućava odbacivanje hipoteze o nezavisnosti. Rezultat hi-kvadrat testa je dovoljno visok (68,738), što govori da je očigledan uticaj kacige na pojavu kome pri teškim povredama mozga.

Primer 3: Tabele kontigencije $m\ x\ n$ i hi-kvadrat test

Niska porođajna masa je definisana kao masa novorođenčeta manja od 2500g. Hoćemo da ispitamo da li porođajna masa zavisi od pušenja majke. Imamo tri grupe: nepušače, pušače i bivše pušače. Pitamo se da li pušačka navika utiče na porođajnu masu novorođenčeta.

	Nepušač	Bivši pušač	Pušač
Porođajna masa < 2500g	143	160	26
Porođajna masa ≥ 2500 g	576	104	476

Rešenje:

```
c(143, 160, 26, 576, 140, 476)
masa <- c(143, 160, 26, 576, 140, 476)
masa<- matrix(masa, ncol=2) # transformacija na matricu sa dve kolone
masa
chisq.test(masa) # test
```

```
[1] 143 160  26 576 140 476
     [,1] [,2]
[1,]  143  576
[2,]  160  140
[3,]   26  476

        Pearson's Chi-squared test

data:  masa
X-squared = 259.3, df = 2, p-value < 2.2e-16
```

Možemo da zaključimo da, sa verovatnoćom moguće greške $p < 0,001$, porođajna masa zavisi od pušenja majke (odbacili smo nultu hipotezu nezavisnosti efekata pušenja i porođajne mase).

3.4.2. Hi-kvadrat test homogenosti

Hi kvadrat test homogenosti je produžetak hi kvadratnog testa nezavisnosti. Testovi homogenosti su korisni da se utvrdi da li se dva ili više nezavisnih slučajnih uzoraka uzimaju iz iste populacije ili iz različitih populacija. Umesto jednog uzorka - kao što koristimo sa problemom nezavisnosti, ovde imamo dva ili više uzoraka. Obe vrste testova koriste istu statistiku testiranja. Međutim, one se razlikuju jedna od druge. Nultom hipotezom tvrdimo da su proporcije u populacijama jednake.

Primer 4: Nutricionista želi da zna da li doba dana utiče na sklonost ka konzumiranju kafe. Dati podaci predstavljaju broj kupljenih napitaka na slučajnom uzorku korisnika usluga kafeterije.

	Rano ujutru	Kasno ujutru	Rano popodne	Kasno popodne
Broj kafa	3	5	8	11
Broj drugih pića	52	48	51	47

Rešenje:

```
> table2=matrix(c(3,52,5,48,8,51,11,47),ncol=4)
> colnames(table2)=c("Rano ujutru","Kasno ujutru","Rano popodne", "Kasno popodne")
> rownames(table2)=c("Broj kafa","Broj drugih pića")
> table2
                Rano ujutru Kasno ujutru Rano popodne Kasno popodne
Broj kafa                 3            5            8            11
Broj drugih pića         52           48           51            47
> chisq.test(table2)

        Pearson's Chi-squared test

data:  table2
X-squared = 5.3626, df = 3, p-value = 0.1471
```

Zaključak: Prihvatamo nultu hipotezu i zaključujemo da su proporcije kupovine kafe tokom dana identične.

Slučaj dva zavisna uzorka: uzorci su zavisni kada upoređujemo objekte merenja pre i posle nekakvog delovanja na njih kako bi otkrili ima li to delovanje neki značaj na ishod merenja.

Slučaj dva nezavisna uzorka: kada se ne mogu osigurati zavisni uzorci tj. upareni uzorci koriste se nezavisni. Za razliku od zavisnih uzoraka koji dolaze iz iste populacije, nezavisni uzorci se slučajno biraju iz dve populacije.

3.5. Mek Nemarov test

Mek Nemarov test se koristi za utvrđivanje da li postoji značajna promena nominalnih podataka pre i posle događaja. Koristi se samo za tabelu 2 × 2, dok hi-kvadrat testovi se koriste za veće tabele.

Tabela za Mek Nemarov test:

		Posle testa 2		
Pre testa 1		Test 2 pozitivan	Test 2 negativan	Ukupno
	Test 1 pozitivan	a	b	a+b
	Test 1 negativan	c	d	c+d
	Ukupno	a+c	b+d	n=a+b+c+d

Primer 4: Želimo da saznamo uticaj video snimka na pušenje. U studiji je učestvovalo 50 učesnika, 25 pušača i 25 nepušača. Svi učesnici pogledali su emotivan video snimak koji prikazuje uticaj smrti od raka povezanih sa pušenjem na njihove porodice. Dve nedelje nakon ovog video snimka, isti učesnici su upitani da li ostaju pušači ili prestaju da puše.

		Posle video snimka		Ukupno
		Pušač	Nepušač	
Pre video snimka	Pušač	9	16	25
	Nepušač	5	20	25
	Ukupno	14	36	50

Rešenje:

Kako se iz tabele vidi, razlike između pre i posle odgledanog video snimka nalaze se u ćelijama b (16) i c (5), dok su u ćelijama a (9) i d (20) navedeni samo oni koji su ili pušačii ili nepušači u oba slučaja. Iz tabele, gornja desna ćelija, možemo videti da je bilo 16 učesnika koji su prvobitno pušili, ali nakon video snimka postali su nepušači. Kako je video snimak zamišljen da smanji pušenje, ovo se može smatrati uspehom. Međutim, u donjoj levoj ćeliji, možemo videti da je 5 nepušača nastavilo da puši nakon video snimka. Možemo da zaključimo da se udeo nepušača povećao nakon video snimka i želeli bismo znati da li je ta razlika statistički značajna. Da bismo to saznali, koristimo Mek Nemarov test.

```
> Performance
            2nd Survey
1st Survey   Approve Disapprove
  Approve          9         16
  Disapprove       5         20
> Performance <-
+     matrix(c(9, 5, 16, 20),
+            nrow = 2,
+            dimnames = list("Pre video snimka" = c("Pušač", "Nepušač"),
+                            "Posle video snimka" = c("Pušač", "Nepušač")))
> Performance
                Posle video snimka
Pre video snimka Pušač Nepušač
         Pušač       9      16
         Nepušač     5      20
> addmargins(Performance)
                Posle video snimka
Pre video snimka Pušač Nepušač Sum
         Pušač       9      16  25
         Nepušač     5      20  25
         Sum        14      36  50
> mcnemar.test(Performance)

        McNemar's Chi-squared test with continuity correction

data:  Performance
McNemar's chi-squared = 4.7619, df = 1, p-value = 0.0291
```

Mek Nemarov test nam je utvrdio da udeo nepušača pre i posle video snimka statistički značajno različit.

3.6. Fišerov test tačne verovatnoće

Ime je dobio po svom izumitelju Ronaldu Fišeru i to je statistički test koji se koristi kada želimo da vidimo da li se odnos jedne promenljive razlikuje od vrednosti druge promenljive. Koristi se umesto hi-testa za tabele $2x2$, posebno u slučaju malih uzoraka.

Unos podataka za Fišerov test:

	Karakteristika prisutna	Karakteristika odsutna	Ukupno
Ishod pozitivan	a	b	a+b
Ishod negativan	c	d	c+d
Ukupno	a+c	b+d	n=a+b+c+d

Primer 5: Želimo da saznamo da li postoji značajna razlika u dve terapije za lečenje zavisnosti od kokaina (ne uzimaju kokain najmanje 6 meseci). Testiramo 21 pacijenata, rezultati su dati u tabeli:

	Terapija 1	Terapija 2	Ukupno
Izlečen	2	7	9
Nije izlečen	9	3	12
Ukupno	11	10	21

Rešenje:

```
> Convictions <- matrix(c(2, 9, 7, 3), nrow = 2,
+                       dimnames =
+                           list(c("Izlečen", "Nije izlečen"),
+                                c("Terapija 1", "Terapija 2")))
> Convictions
             Terapija 1 Terapija 2
Izlečen               2          7
Nije izlečen          9          3
> fisher.test(Convictions, alternative = "less")# one-tail

        Fisher's Exact Test for Count Data

data:  Convictions
p-value = 0.02417
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.7569535
sample estimates:
odds ratio
 0.1093739

> fisher.test(Convictions, conf.int = FALSE) # two tail

        Fisher's Exact Test for Count Data

data:  Convictions
p-value = 0.02997
alternative hypothesis: true odds ratio is not equal to 1
sample estimates:
odds ratio
 0.1093739
```

Zaključak: Odbacujemo nultu hipotezu i zaključujemo da postoji značajna razlika između stope izlečenja za dve terapije.

Funkciju $fisher.test(Convictions, alternative = "less")$ koristili bismo samo ako bi pre eksperimenta odlučili da li nas zanima rezultat da je terapija 1 bolja od terapije 2, inače koristimo funkciju $fisher.test(Convictions, conf.int = FALSE)$.[20]

3.7. Znakovni test

Ovaj test bavi se medijanom kao testom centralne tendencije. Znamo da je vrednost medijane[21] i aritmetičke sredine[22] jednaka u slučaju normalne raspodele populacije. Pri drugačijoj raspodeli mogu se međusobno znatno razlikovati. Umesto brojeva koriste se znakovi $+$ i $-$ pa se zato ovaj test zove znakovni. Pomoću ovog testa možemo da potvrđujemo hipotezu o jednakosti ili većem ili manjem u odnosu na medijanu, kao i hipotezu o razlici medijana među posmatranim parovima.

Primer 6: Na odeljenju intenzivne nege hospitalizovani su bolesnici sa različitim nivoom poremećaja svesti, merenim Glazgovskom skalom kome (GCS) sa rezultatima navedenim u tabeli.

	Bolesnik									
	1	2	3	4	5	6	7	8	9	10
Skor	13	8	6	9	11	7	13	10	12	9

Pitamo se da li je moguće reći da je medijana skora 9.

Rešenje:

Ovde je reč o skoru, kao i o malom broju poređenja na skali. Zato nije opravdano da koristimo t-test[23], već moramo da tražimo znakovni test. Potvrđujemo hipotezu da je medijana posmatranog skora jednaka 9. U prvom koraku prevešćemo vrednosti na znakove, koristeći $+$ ako je vrednost veća od potvrđivane medijane i $-$ ako je manja. Sa 0 označavamo slučaj jednakosti i dobijamo tabelu.

[20] Vidi [4], [8] i [10]

[21] Medijana konačnog zbira je vrednost koja deli skupove na dva jednaka dela. Kad je broj vrednosti neparan, onda će medijana biti srednja vrednost poređanih merenja. Ako je njihov broj paran, onda su dve vrednosti u sredini, a medijana je njihova aritmetička sredina.

[22] Aritmetička sredina se dobija sabiranjem svih vrednosti u uzorku i deljenjem tog zbira sa brojem vrednosti.

[23] U situaciji kad ne poznajemo ni srednju vrednost populacije, ni njenu varijansu, a imamo dovoljno veliki uzorak – između 20 i 30 posmatranja, moramo da koristimo raspodelu koja se zove Studentova t raspodela. t-raspodela ima srednju vrednost 0, a ujedno je simetrično raspodeljena oko srednje vrednosti, disperziju veću od 1.

Znakovi odstupanja GCS od potvrđivane vrednosti medijane 9.

	Bolesnik									
	1	2	3	4	5	6	7	8	9	10
Skor	13	8	6	9	11	7	13	10	12	9
Znakovi	+	-	-	0	+	-	+	+	+	0

Broj znakova „+“ je 5, a „–„ samo 3. Jednaka broju 9 su samo dva slučaja. Imamo više merenja u plusu, nego u minusu, ako ne uzimamo u obzir merenja koja su jednaka potvrđivanoj vrednosti 9 . Pretpostavljamo, ako bi bilo tačno da je medijana jednaka 9, onda bi trebalo da bude isti broj odstupanja u smislu plus i minus. Potvrđena nulta hipoteza i alternativna hipoteza biće:

$H_0: median\ GCS\ =\ 9$

$H_A: median\ GCS\ \neq\ 9$

Kako pretpostavljamo da je broj pozitivnih i negativnih razlika od potvrđivane vrednosti jednak, hipotezu možemo da formulišemona sledeći način:

$$H_0: P(+) = P(-) = 0{,}5 \qquad H_A: P(+) \neq P(-) \neq 0{,}5$$

Nulta hipoteza govori da je verovatnoća pozitivnih i negativnih odstupanja jednaka, odnosno da je verovatnoća 0,5.

Za izračunavanje ćemo koristiti komandu $SIGN.test()$, koja je deo paketa BSDA. Njega je neophodno najpre potražiti u CRAN-u i instalirati da bismo koristili ovu komandu. Komanda zahteva više argumenata: $SIGN.test(x, y = NULL, md = 0, alternative = "two.sided", conf.level = 0.95)$. Argument x sadrži vrednost testirane promenljive u formi brojeva (ne znakova). Argument md je onda jednak vrednosti medijane koju potvrđujemo. Argument $alternative = "two.sided"$ je podešen na obostran test, a moguće ga je podesiti i na manji „less“ ili veći „greater“. Na kraju možemo da odredimo nivo poverenja, pretstavljen vrednošću 0,95

```
> c(13, 8, 6, 9, 11, 7, 13, 10, 12, 9)
 [1] 13  8  6  9 11  7 13 10 12  9
> GCS <- c(13, 8, 6, 9, 11, 7, 13, 10, 12, 9)
> SIGN.test(GCS, y = NULL, md = 9, alternative = "two.sided", conf.level =0.95)

        One-sample Sign-Test

data:  GCS
s = 5, p-value = 0.7266
alternative hypothesis: true median is not equal to 9
95 percent confidence interval:
  7.324444 12.675556
sample estimates:
median of x
        9.5

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.8906 8.0000 12.0000
Interpolated CI       0.9500 7.3244 12.6756
Upper Achieved CI     0.9785 7.0000 13.0000
```

Pozvali smo komandu $SIGN.test$ sa podacima zamenjenim u promenljivu GCS i proveravali jednačinu medijane merenja prema vrednosti 9. Iskoristili smo interval poverenja 0.95 i dvostrani test. U rezultatu se pojavilo s = 5, što nam govori da smo imali pet posmatranja u smislu plusa. Verovatnoća odbacivanja alternativne hipoteze je p-value = 0.7266, čime moramo da prihvatimo nultu hipotezu o jednakosti medijana vrednosti 9.

Znakovni test možemo koristiti i za potvrđivanje razlika između dve medijane pri parnom testu, to pokazujemo u sledećem primeru.

Primer 7: Izvršeno je istraživanje nove dekongestivnog sprejakako bi se videlo koliko je sprej bio efikasan u uklanjanju tegoba nazalne kongestije (zapušenost nosa). Intenzitet fleka pre i posle upotrebe spreja prikazan je na slici:

	A	B	C	D
1	Osobe	Pre	Posle	Znak
2	1	210	197	-
3	2	205	195	-
4	3	193	191	-
5	4	182	174	-
6	5	271	236	-
7	6	239	226	-
8	7	164	157	-
9	8	197	206	+
10	9	222	215	-
11	10	251	196	-
12	11	187	181	-
13	12	175	164	-
14	13	186	198	+
15	14	243	233	-
16	15	246	240	-

Pitamo se da li je sprejefikasan u smanjenju zapušenosti nosa na osnovu ovih podataka o uzorku.

Rešenje:

Radeći isto kao i u prethodnom slučaju, saznaćemo kolika je razlika u smislu plusa, a koliko u smislu minusa. Hipotezu ćemo formulisati tako da nas zanima da li je medijana odgovora posle uvođenja modifikacija veća nego na početku.

$$H_0: medijana\ Pred\ =\ medijana\ Po \qquad H_A: medijana\ Po\ <\ medijana\ Pred$$

ili je možemo formulisati i

$$H_0: medijana\ razlike\ je\ jednaka\ 0 \qquad H_A: medijana\ razlike\ je\ veća\ od\ 0.$$

Koristićemo istu komandu, samo u ovom slučaju imaćemo dve promenljive. Kao prvu promenljivu komande $SIGN.test()$ uvodimo promenljivu $posle$, a posle nje će slediti promenljiva pre.

```
> c(210, 205, 193, 182, 271, 239, 164, 197,222, 251,187, 175, 186,243, 246)
 [1] 210 205 193 182 271 239 164 197 222 251 187 175 186 243 246
> pre<- c(210, 205, 193, 182, 271, 239, 164, 197,222, 251,187, 175, 186,243, 246)
>
> c(197, 195, 191, 174, 236, 226, 157, 206, 215, 196, 181,  164, 198, 233, 240)
 [1] 197 195 191 174 236 226 157 206 215 196 181 164 198 233 240
> posle<- c(197, 195, 191, 174, 236, 226, 157, 206, 215, 196, 181,  164, 198, 233, 240)
>
>
> SIGN.test(posle, pre, alternative = "two.sided", conf.level = 0.05)

        Dependent-samples Sign-Test

data:  posle and pre
S = 2, p-value = 0.007385
alternative hypothesis: true median difference is not equal to 0
5 percent confidence interval:
 -8.254608 -7.872696
sample estimates:
median of x-y
           -8

Achieved and Interpolated Confidence Intervals:

                  Conf.Level   L.E.pt  U.E.pt
Lower Achieved CI     0.0000  -8.0000 -8.0000
Interpolated CI       0.0500  -8.2546 -7.8727
Upper Achieved CI     0.3928 -10.0000 -7.0000
```

Rezultat je pokazao da postoje dva merenja gde je razlika pozitivna. Test daje p-vrednost = 0 .007385, što ukazuje da postoji značajna razlika između merenja pre i posle. Dakle, sprej je efikasan u smanjenju zapušenosti nosa.[24]

3.8. Test medijane

Znakovni test se zasniva na pretpostavci da su dva uzorka u uzajamnom odnosu i predstavlja neparametarsku varijantu parnog t-testa. Često se srećemo sa situacijom kada dva uzorka nisu uzajamno povezana, a broj merenja nije uparen(zavisan) i može da bude različit. Kod normalno raspodeljenih populacija to nije problem jer možemo da koristimo t-test. Ukoliko ne možemo da očekujemo normalnu raspodelu populacije iz kojih biramo uzorak, onda moramo da radimo sa medijanama. Test medijane koristimo za potvrđivanje da li postoji ili ne postoji razlika između medijana dve populacije. Poznat je kao i Vestenberg-Mudov test medijane, po njegovim autorima. Statističko okruženje R nudi u osnovnom paketu $stats$ komandu $mood.test(x, y, alternative = c("two.sided", "less", "greater"), \dots)$. Postupak je zasnovan na računanju medijana za oba uzorka. Onda se frekvencija slučajeva većih od zajedničke medijane i manjih od zajedničke medijane stave u tabelu 2×2. Postupak je prikazan u tabeli:

[24] Vidi [8], [9] i [10]

Tabela $2x2$ za izračunavanje testa medijane:

	Promenljiva 1	Promenljiva 2	Ukupno
Broj slučajeva većih od zajedničke medijane	a	b	a+b
Broj slučajeva manjih od zajedničke medijane	c	d	c+d
Ukupno	a+c	b+d	n=a+b+c+d

Dalje računanje je jednostavan hi-kvadrat test.

Primer 8: Da bismo istražili uticaj novog leka protiv polenske groznice na vozačke sposobnosti, proučavaćemo 24 osobe sa polenskom groznicom: 12 koji su uzimali lekove i 12 koji nisu. Svi učesnici su tada ušli u simulator i prošli kroz vozački test koji je dodelio rezultat svakom vozaču kako je dato na slici.

	A	B
1	Nisu uzimali lek	Uzimali lek
2	11	34
3	15	31
4	9	35
5	4	29
6	34	28
7	17	12
8	18	18
9	14	30
10	12	14
11	13	22
12	26	10
13	31	29

Rešenje:

```
> c(11,15,9,4,34,17,18,14,12,13,26,31)
 [1] 11 15  9  4 34 17 18 14 12 13 26 31
> con<- c(11,15,9,4,34,17,18,14,12,13,26,31)
> c(34,31,35,29,28,12,18,30,14,22,10,29)
 [1] 34 31 35 29 28 12 18 30 14 22 10 29
> drug<- c(34,31,35,29,28,12,18,30,14,22,10,29)
> mood.test(con, drug)

        Mood two-sample test of scale

data:  con and drug
Z = 0.14262, p-value = 0.8866
alternative hypothesis: two.sided
```

Na osnovu rezultata vidimo da test nije doneo statistički bitnu razliku između medijana oba uzorka. Zbog njegove male snage, danas se ovaj test koristi samo retko. Za rešenje ove vrste problema prednost se daje Vilkoksonovom testu za dva uzorka.[25]

3.9. Vilkoksonov test

Frenk Vilkokson[26] je 1945. godine po prvi put uveo test sume rangova. Dve godine kasnije, 1947. statističari Man i Vitni su uveli U test, koji je u sustini isti Vilkoksonovom testu sume rangova. Često se u literaturi ovaj test naziva Vilkokson-Man-Vitnijev test.Vilkoksonov test ekvivalentnih parova (dva zavisna uzorka) nije isto što i test sume rangova, iako su oba testa neparametarska i koriste se kod ordinalnih ili intervalnih podataka.

Kada radimo test sume rangova, vodimo računa da nezavisni uzorci potiču iz istog osnovnog skupa. Pod pojmom test ranga podrazumeva se neparametarska procedura upoređivanja razlika, odnosno povezanost zasnovana na posmatranju poretka testiranih podataka. Ponaša se isto kao t-test. Kad nisu zadovoljeni zahtevi za t-test za dva uzorka, često se može koristiti neparametarski test Vilkoksona za dva uzorka.

$wilcox.test(x, y = NULL, alternative = c("two.sided","less","greater"), mu = 0, paired = FALSE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95, \ldots)$ je komanda koja ga realizuje. Ako želimo da utvrdimo jednakost, odnosno nejednakost prema određenoj vrednosti aritmetičke sredine populacije, možemo da koristimo i varijantu za parni test, kada navodimo $paired = TRUE$. Izračunavanje intervala poverenja pri nivou verovatnoće se određuje na osnovu ostala dva argumenta conf.int = TRUE, conf.level = 0.99.

[25] Vidi [10]
[26] Frank Wilcoxon (1892 – 1965)

3.9.1. Vilkoksonov test ekvivalentnih parova

Primer 9: Potrebno je da utvrdimo da li se sposobnost subjekata koji imaju normalan vid da identifikuju objekte desnim okom razlikuje od njihove sposobnosti identifikacije levim okom. Ukupno 16 ispitanika učestvuju u ovom testiranju. Njima je predstavljen niz slika i ocenjena je njihovoj sposobnosti da prepoznaju predmete uz pomoć oba oka. Rezultati su prikazani na slici.

	A	B	C
1	Osoba	Desno oko	Levo oko
2	1	50	47
3	2	45	45
4	3	33	31
5	4	22	24
6	5	99	78
7	6	79	76
8	7	4	13
9	8	36	46
10	9	62	45
11	10	27	23
12	111	15	14
13	13	26	34
14	14	83	79
15	15	86	81
16	16	51	44

Na osnovu ovih podataka odredićemo da li postoji razlika između dva oka.

Rešenje:

```
> c(50,45,33,22,99,79,4,36,62,27,15,26,83,86,51)
 [1] 50 45 33 22 99 79  4 36 62 27 15 26 83 86 51
> desnooko<- c(50,45,33,22,99,79,4,36,62,27,15,26,83,86,51)
> c(47,45,31,24,78,76,13,46,45,23,14,34,79,81,44)
 [1] 47 45 31 24 78 76 13 46 45 23 14 34 79 81 44
> levooko<- c(47,45,31,24,78,76,13,46,45,23,14,34,79,81,44)
> wilcox.test(desnooko, levooko, paired = TRUE,)

        Wilcoxon signed rank test with continuity correction

data:  desnooko and levooko
V = 69.5, p-value = 0.2999
alternative hypothesis: true location shift is not equal to 0
```

Na osnovu dobijenih rezultata možemo zaključiti da nema značajne razlike između dva oka.

3.9.2. Vilkoksonov test sume rangova

Primer 10: Načelnik odeljenja traumatologije u većem gradu tvrdi da su povrede pacijenta koji su hospitalizovane na njegovom odeljenju, teže od onih koji su hospitalizovane kod njegovog kolege u bolnici u manjem mestu. Odlučili su da uporede težinu povreda na oba odeljenja primenom ISS, što je skor težine povrede zasnovan na anatomskom vrednovanju.

Bolnica	ISS skor											
Gradska	45	52	37	62	75	35	25	45	49	25	75	9
Prigradska	36	42	49	75	75	36	25	9	49	50		

Da li je moguće na osnovu prikazanih podataka reći da je težina povreda kod primljenih pacijenata različita?

Rešenje:

```
>  grad <- c(45,52,35,37,62,75,35,25,45,49,25,75,9)
>  selo <- c(36,42,75,49,75,75,36,25,9,49,50)
>  wilcox.test(grad, selo, alternative = c("two.sided"))

        Wilcoxon rank sum test with continuity correction

data:  grad and selo
W = 62.5, p-value = 0.6199
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(grad, selo, alternative = c("two.sided")) :
  cannot compute exact p-value with ties
```

Rezultat nam govori da test nije doneo statistički bitnu razliku između medijana oba uzorka.

3.10. Kruskal-Volisova analiza varijanse

Nekada ne možemo da ispunimo uslove parametarske analize varijanse (ANOVA[27]), najčešće zbog malog broja podataka pojedinih uzoraka, ili u slučajevima kada ne možemo da govorimo o normalnoj raspodeli podataka, zato što su to, na primer, podaci u obliku skora ili niže skale. U ovakvom slučaju moramo da posegnemo za neparametarskom analizom varijanse, koja nosi ime po njenim autorima, dakle Kruskal-Volis. Isto kao i njen ekvivalent i ona potvrđuje hipotezu za više od dve populacije. Kada dobijemo rezultat ovog testa da su najmanje dve populacije različite, tada koristimo Vilkoksonov test za dva uzorka kako bi saznali koje su to dve populacije

[27] ANOVA (Analysis of Variance) koncipirana je na potvrđivanju razlika između varijansi.

različite. Kruskal-Volisov test upoređuje medijane i saznaje da li su, ili nisu jednake, za razliku od ANOVA koja radi sa varijansama populacije. Isto kao Vilkoksonov test, postupak je zasnovan na principu upoređivanja redosleda. Postupak sledi ukoliko su redosledi slični, odnosno različiti.

Test se koristi za ispitivanje nulte hipoteze da k ($k > 2$) nezavisnih uzoraka pripada istom osnovnom skupu. Njegovo korišćenje prikazano je na primeru.

Upoređivanje tri uzorka pomoću Kruskal-Volisovog neparametrijskog testa.

Primer 11: U ovom primeru ćemo saznati kako su pacijenti zadovoljni radom tri lekara, koji se zovu Dejan, Branko i Milan. Svakog od njih dodelili smo njihovim pacijentima, tako da se nije desilo da pacijent jednog lekara vrednuje drugog lekara, već se uvek izjašnjavao samo prema tome koji se lekar za njega direktno brinuo. Vrednovanje je iskazano na skali od 1 do 4, gde je 1 izražavalo veliko nezadovoljstvo, a 4 veliko zadovoljstvo. Pitamo se postoji li razlika među ispitivanim lekarima.

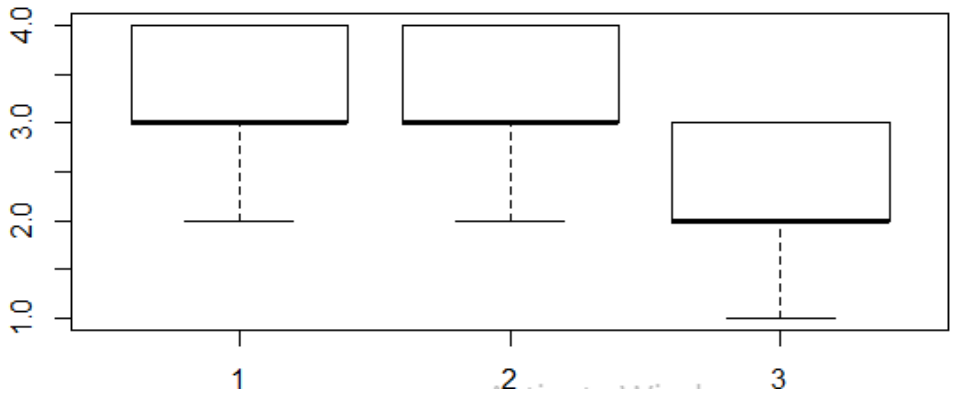
Lekari		
Dejan	Branko	Milan
3	2	2
4	4	3
2	3	1
3	3	4
4	4	2
3	4	2
4	2	3
2	3	3
1	2	4
3	4	2

Rešenje:

```
> c(3, 4, 4, 3, 3, 4, 3, 2, 2, 4)
 [1] 3 4 4 3 3 4 3 2 2 4
> dejan<-c(3, 4, 4, 3, 3, 4, 3, 2, 2, 4)
> c(3, 4, 4, 3, 4, 4, 3, 3, 3, 2)
 [1] 3 4 4 3 4 4 3 3 3 2
> branko<-c(3, 4, 4, 3, 4, 4, 3, 3, 3, 2)
> c(2, 1, 3, 2, 3, 1, 2, 2, 3, 2)
 [1] 2 1 3 2 3 1 2 2 3 2
> milan<-c(2, 1, 3, 2, 3, 1, 2, 2, 3, 2)
> lekari <- list(dejan, branko, milan)
> kruskal.test(lekari) # pozivanje komande

        Kruskal-Wallis rank sum test

data:  lekari
Kruskal-Wallis chi-squared = 10.624, df = 2, p-value = 0.004933
```

Na osnovu rezultata možemo zaključiti da su prisutne razlike kod najmanje dva uzorka na nivou poverenja $p < 0,01$, ali ne znamo koji se od uzoraka međusobno razlikuju. Koristićemo boks plot dijagram i to će nam pomoći da uočimo razliku. Nacrtaćemo ga pozivanjem komande $boxplot()$ sa parametrom lekari, koji sadrži rezultate tri lekara.

```
> boxplot(lekari)
```



Boxplot dijagram za tri lekara

Sa slike vidimo da postoji razlika između trećeg lekara, odnosno Milana i njegovih dvojice kolega Dejana i Branka. Ovu tvrdnju proveravamo računanjem koristeći Vilkoksonov test za svaki par lekara, odnosno za sledeće parove: Dejan-Branko, Dejan-Milan, Branko-Milan.

```
> wilcox.test(dejan, branko)

        Wilcoxon rank sum test with continuity correction

data:  dejan and branko
W = 47, p-value = 0.837
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(dejan, branko) :
  cannot compute exact p-value with ties
> wilcox.test(dejan, milan)

        Wilcoxon rank sum test with continuity correction

data:  dejan and milan
W = 83, p-value = 0.009911
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(dejan, milan) :
  cannot compute exact p-value with ties
> wilcox.test(branko, milan)

        Wilcoxon rank sum test with continuity correction

data:  branko and milan
W = 87, p-value = 0.003675
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(branko, milan) :
  cannot compute exact p-value with ties
```

Zaključak: Između Dejana i Branka nije bilo statistički značajne razlike, ali između Dejana i Milana i Branka i Milana ova razlika je bila na nivou poverenja $p < 0,01$.[28]

3.11. Kohranov Q test

Ime je dobio po William Gemmell Cochran-u. Kohranov Q test je proširenje Mek Nemarovog testa koji uključuje više od dva zavisna uzorka, a kada je k = 2 Kohran Q test će dobiti rezultat koji je ekvivalentan dobijenom rezultatu Mek Nemarovim testom. Na istoj grupi ispitanika vršimo merenje u različito vreme.

Primer 12: Radnici u Domu zdravlja u Kruševcu pokazuju dve vrste ponašanja: energično i umorno. Ovo ponašanje je izmereno nauzorku od 20 radnika u ponedeljak, sredu i petak u toku jedne nedelje, kao što je prikazano na slici (gde 1 predstavlja energičnost i 0 predstavlja umor). Da li postoji značajna razlika u ponašanju između tri vremenska perioda?

[28] Vidi [4], [8], [9] i [10]

	A	B	C	D	E
1	Pacijenti	Ponedeljak	Sreda	Petak	
2	1	1	1	1	3
3	2	0	1	1	2
4	3	0	0	1	1
5	4	0	1	0	1
6	5	1	0	0	1
7	6	0	1	1	2
8	7	0	1	1	2
9	8	0	0	1	1
10	9	0	1	1	2
11	10	0	1	0	1
12	11	1	1	0	2
13	12	1	1	1	3
14	13	0	0	0	0
15	14	1	0	1	2
16	15	0	1	1	2
17	16	0	1	0	1
18	17	0	0	1	1
19	18	0	1	1	2
20	19	1	0	1	2
21	20	0	1	1	2
22		6	13	14	33

Rešenje:

```
> result<-c(1,1,1,0,1,1,0,0,1,0,1,0,1,0,0,0,1,1,0,1,1,0,0,1,0,1,1,0,1,0,1,1,0,1,1,1,0,
0,0,1,0,1,0,1,1,0,1,0,0,0,1,0,1,1,1,0,1,0,1,1)
> length(result)
[1] 60
> methods<-factor(rep(LETTERS[1:3],20))
> dz<-factor(rep(letters[1:20],each=3))
> tapply(result,list(dz,methods), sum)
  A B C
a 1 1 1
b 0 1 1
c 0 0 1
d 0 1 0
e 1 0 0
f 0 1 1
g 0 1 1
h 0 0 1
i 0 1 1
j 0 1 0
k 1 1 0
l 1 1 1
m 0 0 0
n 1 0 1
o 0 1 1
p 0 1 0
q 0 0 1
r 0 1 1
s 1 0 1
t 0 1 1
> cochran.qtest(result~methods|dz)
```

```
        Cochran's Q test

data:  result by methods, block = dz
Q = 6.7059, df = 2, p-value = 0.03498
alternative hypothesis: true difference in probabilities is not equal to 0
sample estimates:
proba in group A proba in group B proba in group C
            0.30             0.65             0.70

        Pairwise comparisons using Wilcoxon sign test

       A B
B 0.1384 -
C 0.1157 1

P value adjustment method: fdr
```

Vidimo da je Q = 6.706 i $p - vrednost = 0.035 < \alpha$ što pokazuje da postoji značajna razlika između ispitivanih dana. Kao što iz ovog primera vidimo, Cohranov Q test nam daje podatak da rezultati različitih ispitivanih uslova ne pripadaju istoj populaciji kao i energičnost radnika u procentima:

Ponedeljak	Sreda	Petak
30%	65%	70%

Postoji značajna razlika između procenta radnika koji su energični u ponedeljak (30%), sreda (65%) i petak (70%).[29]

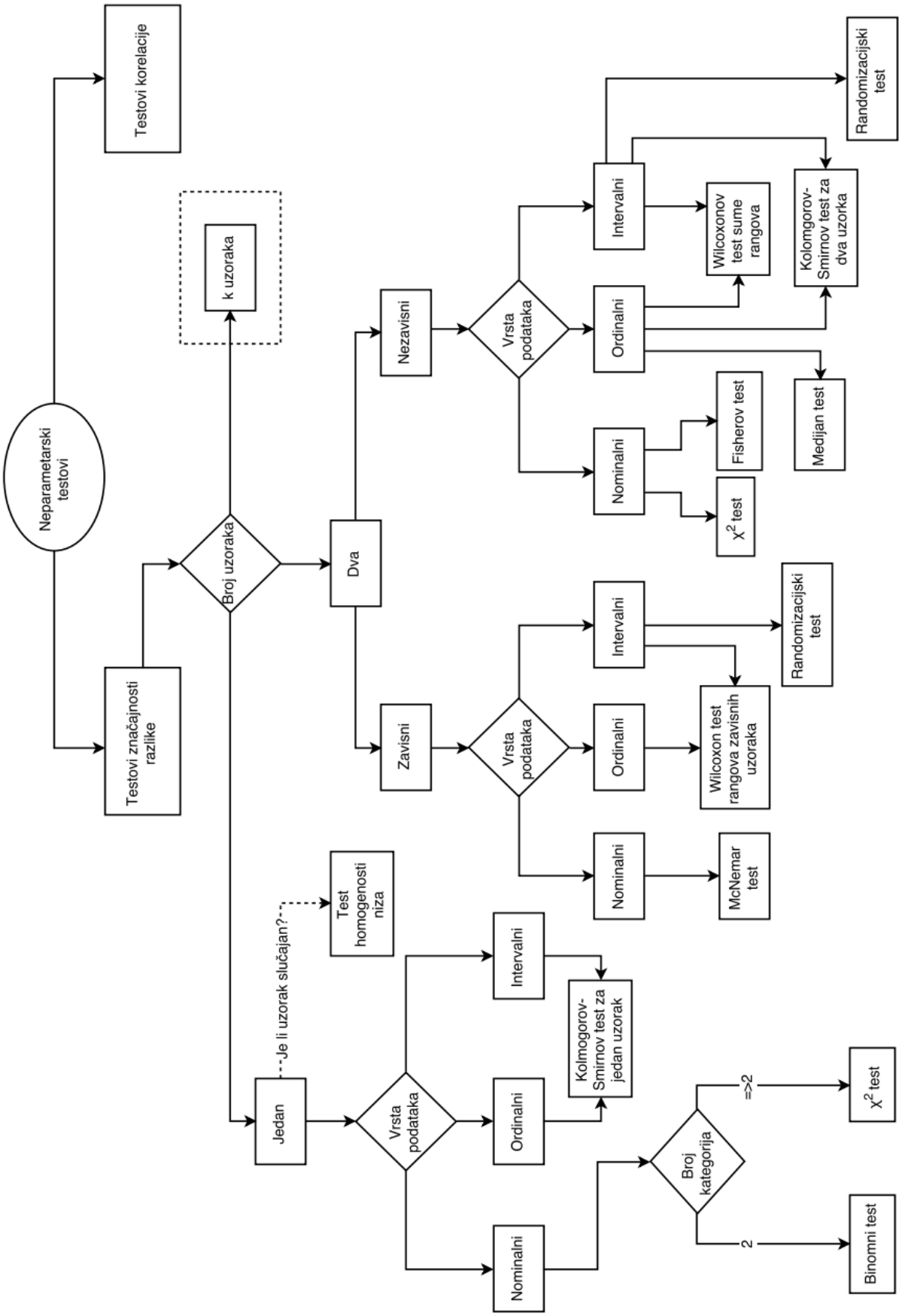
[29] Vidi [4] i [10]

Algoritam odabira metode neparametarske statistike

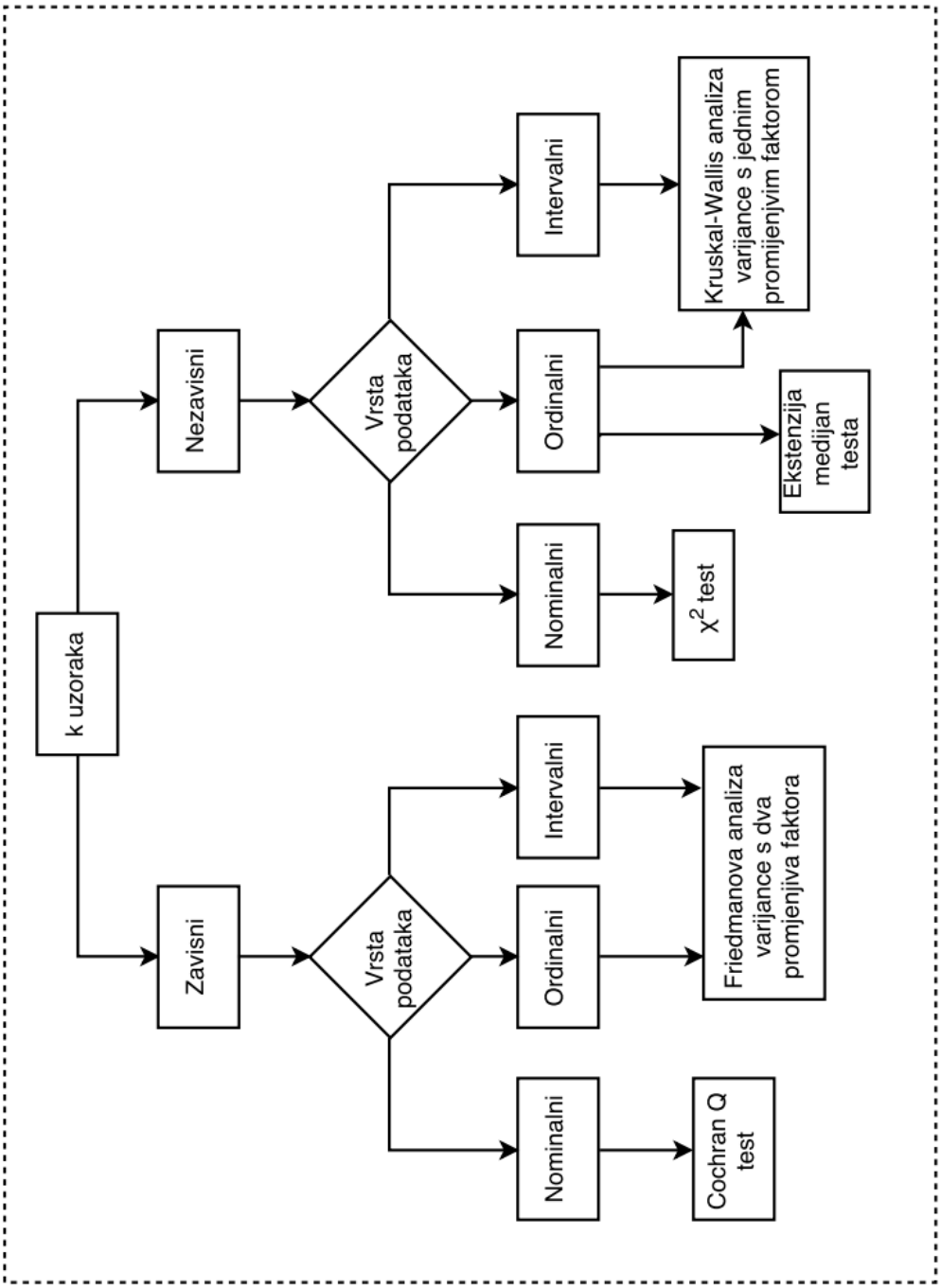
Od neparametarskih statističkih testova potrebno je izabrati onaj koji odgovara svrsi testa i vrsti podataka. Kako bi se jednostavno moglo utvrditi koji je test odgovarajući i primenljiv izrađen je алгоритам njihovog odabira koji je prikazan na slici. Neparametarski testovi se dele na testove značajnosti razlike, koji testiraju jesu li uzorci iz iste populacije, i testove korelacije, koji ispituju postoji li povezanost između promenljivih uzoraka.

Ako želimo da testiramo postoji li značajna razlika između uzoraka, potrebno je primeniti sledeće etape u postupku:

- Prvo treba odrediti broj uzoraka.
- Ako imamo dva ili više uzoraka potrebno je uočiti jesu li zavisni ili nezavisni.
- Nakon toga se određuje vrsta podatka, jesu li u nominalnoj, ordinalnoj ili intervalnoj skali.
- Poslednji korak jest odabir testa.


Neparametarski testovi
Testovi korelacije
Testovi značajnosti razlike
Broj uzoraka
k uzoraka
Dva
Jedan
Je li uzorak slučajan?
Test homogenosti niza
Nezavisni
Zavisni
Vrsta podataka
Intervalni
Ordinalni
Nominalni
Randomizacijski test
Wilcoxonov test sume rangova
Kolomgorov-Smirnov test za dva uzorka
Medijan test
Fisherov test
χ^2 test
Wilcoxon test rangova zavisnih uzoraka
McNemar test
Kolmogorov-Smirnov test za jedan uzorak
Broj kategorija
=>2
2
Binomni test

Algoritam



Algoritam (nastavak)

Zaključak

Korišćenjem neparametarskih metoda analize smanjuje se rizik od pogrešnog zaključivanja, jer pomenute metode ne daju pretpostavke o populaciji iako mogu imati nižu statističku moć od parametarskih metoda. Drugim rečima, neparametarske metode su „uvek validne, ali ne uvek i efikasne“, dok su parametarske metode „uvek efikasne, ali ne uvek važeće“. Zbog toga se parametarske metode preporučuju kada je njihova upotreba opravdana. Iz niza navedenih primera u radu, može se videti njihova svrsishodnost kod malih uzoraka, gde se u praksi pokazuje njihova validnost. Upotreba neparametarskih metoda, kao što je prikazano, je veoma značajna u medicinskoj biostatistici kada su druge metode obrade podataka neprimenjljive.

Literatura:

[1] Martin Rusnák, Viera Rusnákova, Merek Majdan, Bioštatistika pre študentov verejného zdravotníctva, Fakulta zdravotníctva a sociálnej práce, Tŕne 2010.

[2] Peter Dalgaard. Introductory Statistics with R. Springer, 2nd edition, 2008. ISBN 978-0-387-79053-4.

[3] José Cláudio Faria, Philippe Grosjean, Enio Galinkin Jelihovschi and Ricardo Pietrobon, Tinn-R Editor, Rmetrics Association, Zurich 2010.

[4] Gregory W. Corder, Dale I. Foreman, Nonparametric statistics for non-statisticians, John Wiley & Sons, Inc. Hoboken, New Jersey 2009.

[5] David J.Sheskin, Handbook of parametric and nonparametric statistical procedures, Chapman & Hall/CRC, Florida 2000.

[6] Peter Y. Chen,Paula M. Popovich,Correlation parametric and non parametric measures, Sage publication, 2002.

[7] J. Susan Milton, Paul M. McTeer, James J. Corbett, Introduction to statistics, McGraw-Hill Companies, 1997.

[8] A. Petrie, C. Sabin, Medical Statistic at a Glance, Blackwell Science, 2000.

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4754273/

[10] http://www.real-statistics.com/non-parametric-tests/

Biografija

Vesna Živanović je rođena 17.03.1992. u Loznici. Osnovnu školu „Borivoje Ž. Milojević“ završila je 2007. godine u Krupnju kao nosilac Vukove diplome. Iste godine upisala je smer elektrotehničar računara u Srednjoj skoli u Krupnju, koju je završila 2011. godine kao odličan đak. Po završetku srednje škole upisala je osnovne akademske studije na Prirodno-matematičkom fakultetu u Novom Sadu, smer matematika, koje je završila u septembru 2015. godine.Nakon osnovnih studija, upisala je master akademske studije na Prirodno- matematičkom fakultetu u Novom Sadu, master profesor matematike. U februaru 2016. godine počela je da radi u Srednjoj školi u Krupnju na mestu profesora matematike.

Položila je sve ispite predviđene nastavnim planom i programom ove godine u junu.

Vesna Živanović

UNIVERZITET U NOVOM SADU

PRIRODNO-MATEMATIČKI FAKULTET

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR
Identifikacioni broj:
IBR
Tip dokumentacije: *monografska dokumentacija*
TD
Tip zapisa: *tekstualni štampani materijal*
TZ
Vrsta rada: *master rad*
VR
Autor: *Vesna Živanović*
AU
Mentor: *dr Zagorka Lozanov-Crvenković*
MN
Naslov rada: *Neparametarski testovi u biostatistici sa primenom na javno zdravlje*
NR
Jezik publikacije: *srpski (latinica)*
JP
Jezik izvoda: *s/e*
JI
Zemlja publikovanja: *Republika Srbija*
ZP
Uže geografsko područje: *Vojvodina*
UGP
Godina: *2019.*
GO
Izdavač: *autorski reprint*
IZ
Mesto i adresa: *Novi Sad, Trg Dositeja Obradovića 4*
MA
Fizički opis rada: *3 poglavlja, 52 strane, 10 lit. citat, 26 tabela, 10 grafika*
FO
Naučna oblast: *matematika*
NO
Naučna disciplina: *profesor matematike*
ND
Ključne reči: *skale, tabele kontigencije, mere povezanosti, neparametarski testovi, R*
PO

UDK

Č

Čuva se: *u biblioteci Departmana za matematiku i informatiku, Prirodno-matematičkog fakulteta, u Novom Sadu*

ČU

Važna napomena:

VN

Izvod: Potreba za neparametarskom statistikom proizilazi iz činjenice da dobijeni podaci nemaju uvek normalnu raspodelu tako da se metode parametarske statistike ne mogu koristiti. Korišćenjem neparametarskih metoda analize smanjuje se rizik od pogrešnog zaključivanja, jer pomenute metode ne daju pretpostavke o populaciji iako mogu imati nižu statističku moć od parametarskih metoda. Drugim rečima, neparametarske metode su „uvek validne, ali ne uvek i efikasne“, dok su parametarske metode „uvek efikasne, ali ne uvek važeće“. Iz niza navedenih primera u radu, može se videti njihova svrsishodnost kod malih uzoraka, gde se u praksi pokazuje njihova validnost. Upotreba neparametarskih metoda, kao što je prikazano, je veoma zanačajna u medicinskoj biostatistici kada su druge metode obrade podataka neprimenjljive. Svaki od neparametarskih testova je pojedinačno objašnjen i dati su primeri (iz oblasti javnog zdravlja) za svaki od njih. Primeri su obrađeni u programu R.

IZ

Datum prihvatanja teme od strane NN veća: *31.01.2019.*

DP

Datum odbrane:

DO

Članovi komisije:

KO

Predsednik: *dr Ljiljana Gajić, redovni profesor*

Član: *dr Zagorka Lozanov-Crvenković , redovni profesor*

Član: *dr Ivana Štajner-Papuga, redovni profesor*

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
KEY WORDS DOCUMENTATION

Accession number:
ANO
Identification number:
INO
Document type: *monograph type*
DT
Type of record: *printed text*
TR
Contents code: *master thesis*
CC
Author: *Vesna Živanović*
AU
Mentor: *dr Zagorka Lozanov-Crvenković*
MN
Title: *Non-parametric tests in biostatistics with application to public health*
XI
Language of text: *serbian (latin)*
LT
Language of abstract: *s/e*
LA
Country of publication: *Republic of Serbia*
CP
Locality of publication: *Vojvodina*
LP
Publication year: *2019.*
PY
Publisher: *author's reprint*
PU
Publ. place: *Novi Sad, Trg Dositeja Obradovića 4*
PP
Physical description: *3 sections, 52 pages, 10 references, 26 tables, 10 graphs*
PD
Scientific field: *mathematics*
SF
Scientific discipline: *math professor*
SD
Key words: *scales, contingency tables, measure of association, non-parametric tests, R*
UC

Holding data: *Department of Mathematics and Informatics' Library, Faculty of Sciences, Novi Sad*

HD

Note:

N

Abstract: *The need for non-parametric statistics arises from the fact that the data obtained do not always have a normal distribution, so that the methods of parametric statistics cannot be used. The use of non-parametric methods of analysis reduces the risk of inference, since the mentioned methods do not make assumptions about the population, although they may have lower statistical power than parametric methods. In other words, non-parametric methods are "always valid but not always effective", while parametric methods are "always effective but not always valid". From a series of examples in the paper, one can see their usefulness in small samples, where their validity is shown in practice. The use of non-parametric methods, as shown, is very important in medical biostatistics when other data processing methods are inapplicable. Each of the non-parametric tests is explained individually and examples (in the field of public health) are given for each. The examples are covered in R.*

AB

Accepted by the Scientific Board on: *31.01.2019.*

ASB

Defended:

DE

Thesis defend board:

DB

President: *dr Ljiljana Gajić, full professor*

Member: *dr Zagorka Lozanov-Crvenković, full professor*

Member: *dr Ivana Štajner-Papuga, full professor*