CENTAR ZA EDUKACIJU, TOLERANCIJU I MULTIKULTURALIZAM

Martin Rusnak, Viera Rusnakova, Marek Majdan

BIOSTATISTIKA

Maglić, 2013.

CENTAR ZA EDUKACIJU, TOLERANCIJU I MULTIKULTURALIZAM

Edicija naučnih knjiga

Biostatistika

Naslov originala:
Bioštatistika pre študentov verejnéhoo zdravotníctva

Autori:
Martin Rusnak, Viera Rusnakova, Marek Majdan
Katedra javnog zdravstva, Fakultet zdravstva i socijalnog rada, Trnavski univerzitet u Trnavi

Recenzenti:
Prof. MUDr. Peter Krištufek, CSc.
Doc. MUDr. Henrieta Hudečkova, PhD., MPH
Prof. MUDr. Ljudmila Ševčikova, CSc.

Urednik:
Doc. dr Predrag Đurić

Prevod:
Anna Đurić

Maglić, 2013.

CIP – Каталогизација у публикацији
Библиотека Матице Српске , Нови Сад

57 : 311 (035)

РУСНАК, Мартин
Biostatistika [Elektronski izvor] / Rusnak Martin, Rusnakova Viera, Majdan Marek ; [prevod Anna Đurić]. – Maglić : Centar za edukaciju, toleranciju i multikulturalizam, 2013. – 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm. – (Edicija naučnih knjiga)

Prevod dela: Bioštatistika pre študentov verejného zdravotníctva.

ISBN 978-86-88497-03-9
1. Руснакова, Виера 2. Мајдан, Марек
а) Биостатистика - Приручници
COBISS.SR-ID 282466823

SADRŽAJ

PRVO POGLAVLJE

Uvod - osnovni pojmovi i definicije

Sadržaj poglavlja

U uvodu ćemo objasniti svrhu i cilj ove publikacije i navesti osnovne pojmove i definicije. Nakon upoznavanja sa ovim poglavljem, čitalac bi trebalo da stekne dovoljnu količinu pojmova i koncepciju za savladavanje narednih poglavlja. Bavićemo se ciljevima statističkog istraživanja sa akcentom na javno zdravlje. Navešćemo osnovne pojmove i objasnićemo njihovo korišćenje. Objasnićemo odnos između populacije i uzorka populacije, a ujedno ćemo naznačiti i osnovna pitanja koja će biti rešena statističkim testovima. U zaključku ćemo objasniti statističko okruženje projekta R, koje ćemo u celoj publikaciji koristiti za rešavanje primera.

Tabela 1: Ciljevi oglavlja

Ciljevi ovog poglavlja su:
• uvesti i objasniti zadatke statistike u javnom zdravlju i u biomedicinskim naukama • predstaviti osnovne pojmove • dati uvod u statistički program R

Kako i zašto je nastala ova publikacija

Autori reaguju na potrebu da osnaže nastavu statističkih metoda u oblasti nauka koje se bave zdravljem populacije i koje se označavaju kao javno zdravlje. Sa pojmom statistika obično se povezuje i pojam istraživanja, pod kojim ovde podrazumevamo pristup svakodnevnom radu u oblasti javnog zdravlja, a koji primenjuje naučni metod sa ciljem da upozna stvarnost. Problematika istraživanja je prilično široka, pa se ova knjiga posvećuje samo jednom delu istraživanja, a to su odabrane statističke metode koje se koriste u istraživanjima. Ne obrađujemo specijalne statističke pristupe i metode koje nisu rutinske u navedenoj struci.

Šta je, dakle, cilj ove publikacije? U prvom redu da se pripremi pomoćno sredstvo za studente, koje će služiti pri izučavanju statistike, sa krajnjim ciljem njene primene u praksi. Time se nadograđuju već postojeći udžbenici, koji se danas koriste. Povezanost sa klasičnim udžbenicima statistike polazi od činjenice da pominjani autori nisu imali u vidu instrumente koje nam pruža današnje vreme. Ovde se prvenstveno misli na računare i statističke softvere. Razvoj tzv. „*open source*“[1] i drugih slobodno dostupnih programa (gde možemo uvrstiti i Epi-Info autora podržanih od Svetske zdravstvene organizacije) vodio je do otvorenih i do tada nepostojećih mogućnosti primene statističkih metoda. Kao i uvek, osim pozitivnih, ovakav slobodan pristup nosi sa sobom i negativne strane – na primer pogrešno korišćenje statističkih metoda, loš kvalitet ulaznih podataka i neznanje zasnovano na netačnoj interpretaciji rezultata. Ova publikacija može da pomogne pri otklanjanju nekih od navedenih problema, ali i kod racionalne primene potencijala, koje statistika pruža veštom korisniku.

[1] *Open source*, u doslovnom prevodu otvoreni izvor, predstavlja skup računarskih programa, kojima se može pristupiti besplatno i dostupni su svima koji to žele. Ovde spade, na primer, Open office, kao i statistički system R, koji mi koristimo.

Publikacija je u prvom redu pomoćno sredstvo namenjeno studentima javnog zdravlja, ali će je sigurno uzeti u ruke i studenti i apsolventi drugih nezdravstvenih i zdravstvenih smerova. U širem smislu ona pruža osnovnu orijentaciju u oblasti primene statističkih metoda, primere i način njihovog rešavanja, ali ne zamenjuje poznavanje metoda, njihovu snagu i ograničenja. Za primere navedene u ovoj publikaciji nisu navedena uputstva za primenu statističkih metoda bez poznavanja njihove suštine. Time ne mislimo da bi čitalac morao da izučava i poznaje formalnu stranu matematičke statistike sa njenim formaliziranim zapisom i formulama. Ne želimo da se one memorišu bez razumevanja suštine (što se često dešava), nego da vode do razumevanja kako određena metoda funkcioniše, kakav se postupak iza nje skriva. Ne pretpostavljamo da će čitalac sam pisati statističke procedure, ali u određenim situacijama potreban je pogled i na tabele sa rezultatima koje je kreirao statistički program.

Jedna od bitnih odluka koje smo morali da napravimo, ticala se statističkog programa u kojem obrađujemo primere. U slučaju sakupljanja podataka i jednostavnog razvrstavanja, radimo sa programom Excel u verziji koja je stavni deo Open Office-a. Primeri za statističke procedure su predstavljani u programu R4. U pogledu na smisao ove publikacije, nismo uvrstili opis projekta R i opis načina njegovog korišćenja. Pretpostavljamo da će čitalac tragati za mnoštvom izvora koje prate projekat R, ili će koristiti pojednostavljena uputstva na stranici naše katedre.

Potrebno je napisati i kako je publikacija sastavljena i kako preporučujemo da se koristi. Svako poglavlje predstavlja celinu za sebe, usmerenu na rešavanje određene vrste problema. Počinjemo klasično, tj. predlogom postupka u statističkom istraživanju, sakupljanjem i pripremom podataka koji će biti obrađeni. Nastavljamo sa jednostavnom proverom kvaliteta podataka i njihovim osnovnim statističkim opisom. Onda prelazimo na postupke statističke inferencije, koji su podeljeni u više poglavlja. Završavamo logističkom regresijom. Publikaciju nije potrebno čitati od početka do kraja. Radije preporučujemo da se pročitaju uvodna, opšta poglavlja, da se detaljno informiše o testiranju hipoteza, a onda da se orijentiše na dalja poglavljima, prema potrebi. U svakom poglavlju čitalac će naći dosta primera, od najjednostavnijih, pa da kompleksnih statističkih analiza. Kad je primer obrađen programom R, onda je dato uputstvo iz algoritma u jeziku tog programa. Matematičkim opisima je uvek dodato obrazloženje kako ih čitati i interpretirati.

Bitan dodatak publikacije je i veb stranica Katedre za javno zdravlje (http://vz.truni.sk), gde će čitalac pronaći druge detalje, primere kao i kontakte autora. Redovni studenti imaju mogućnost i učenja na daljinu u okruženju MOODLE (pristup preko navedene veb stranice), gde će pronaći ne samo primere, već i elektronske adrese nastavnika.

Šta je statistika, zašto je značajna za javnozdravstvenog radnika

Ništa novo nećemo reći ako konstatujemo da statistika prati problematiku javnog zdravlja od njegovog samog početka. Prisetimo se registrovanja obolelih u Londonu i uloge statistike kod prepoznavanja uzroka nastanka epidemije kolere. Imena kao Vilijem Far i Džon Snou su trajno upisana u istoriju ljudskog znanja. Uprkos tome, nismo odgovrili na pitanje zašto je statistika značajna. Najjednostavniji odgovor je da značaj statistike izvire iz slučajnog karaktera uzroka oboljenja. Slučajnost možemo dokumentovati, na primer, time što neće svaki kontakt sa faktorom koji izaziva oboljenje i dovesti do manifestiranje oboljenja. Primer su oboljenja kao tuberkuloza, HIV infekcija i druge. Mnoga oboljenja su izazvana kombinacijom različitih faktora, a ni jedan od

njih ne možemo označiti kao jedini uzrok. To je, na primer, slučaj kod mnogih hroničnih oboljenja, gde stil života, genetika i životno okruženje zajedno određuju pojavu i razvoj pojedinih oboljenja, kao što su visok krvni pritisak, šećerna bolest, maligni tumor. I tokom studiranja i pri pružanju zdravstvene zaštite koristi se faktor slučajnosti, bilo u formi čovekovih odluka, bilo u reakcijama čoveka. Kod određivanja mera slučajnosti i određenja (determinacija), nećemo moći bez znanja verovatnoće i statistike. Sledeći deo odgovora na pitanje o značaju statističkih metoda je udruženost javljanja određenih fenomena. Znamo da kada treba da napravimo predstavu o nečemu, a šta prevazilazi našu mogućnost da upamtimo sve pojave, koristićemo neku njihovu zajedničku karakteristiku. Na primer, u školi se koristi prosek ocena kao slika o učinku učenika (mada bi bilo pravednije koristiti medijanu). U zdravstvenim istraživanjima često radimo sa populacijama koje se sastoje od mnogih pojedinaca. Nije moguće proučavati ih bez korišćenja određene metode, koja ih zajednički opisuje i pomaže da se saznaju njihove sličnosti i razlike. Ove zadatke rešava statistika.

Definicija 1: Šta je statistika

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

Statistika se prema oblasti primene dalje deli na matematičku i primenjenu. Matematička statistika proučava, razvija statističko znanje i same metode. U oblasti javnog zdravlja govorimo o biomedicinskoj statistici, koja predstavlja primenu statistike u biomedicinskim naukama, a time i u javnom zdravlju. Često se izdvaja i zdravstvena statistika, koja se bavim užom oblašću opisivanja zdravstvenih sistema. Epidemiološka statistika predstavlja skup statističkih metoda koje se često koriste u epidemiološkim studijama. S obzirom na svrhu ove publikacije bavićemo se osnovama biomedicinske statistike, koje se mogu koristiti u svim oblastima bioloških i medicinskih naukama.

Definicija 2: Biostatistika

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

Dva su faktora posebno podstakla razvoj statistike u javnom zdravlju. Prvi od njih je multifaktorijalna priroda nastanka oboljenja, a s tim u vezi i potreba za obradu ne samo jednog para podataka, već cela polja podataka i traženje njihove uzajamne povezanosti. Proračune više nije bilo moguće raditi ručno i tako, paralelno sa razvojem računara, dolazi do njihove sve češće upotrebe u epidemiologiji, a postepeno i u drugim oblastima istraživanja javnog zdravlja. Primenjuju se ne samo veliki skupovi, već i nelinearni i drugi statistički komplikovani postupci. Kompjuteri su, na primer, omogućili sprovođenje populacione studije o povezanosti pušenja i raka pluća[2]. Takođe su vodili ka širenju logističke regresije[3], čija je prva primena u studiji nastanka oboljenja zabeležena 1961. godine i to za određivanje povezanosti koronarne bolesti srca, sistolnog krvnog pritiska i nivoa holesterola[4]. Uprkos tome što su mnogi od ovih postupaka stariji od pola veka, uprkos

[2] Doll, R., Peto, R., Boreham, J., Sutherland, I.: Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ 2004;328:1519 (26 June); http://www.bmj.com/cgi/content/full/328/7455/1519

[3] Cornfield J, Gordon T, Smith WW. Quantal response curves for experimentally uncontrolled variables. Bull.Int.Stat.Inst. 1961;38,pt 3:97-115.

[4] Truett J, Cornfield J, Kannel WB. A multivariate analysis of the risk of coronary heart disease in Framingham.

činjenici da su računari opšte dostupni, uprkos tome da renomirani stručni časopisi ne primaju članak bez odgovarajuće statistike, uprkos svemu tome, stalno se srećemo sa nepravilnom primenom statističkih metoda, kao i sa iznošenjem mišljenja o rezultatima slučajnih (stohastičkih) procesa bez bilo kakvog statističkog obrazloženja. To je i razlog zbog kojeg nastaje ova publikacija, kako se ne bi činila šteta zbog izvorno netačne upotrebe i interpretacije rezultata, te da se ne bi izgubili zanimljivi rezultati istraživačkog rada.

Zašto je statistika nauka? Koristi naučni pristup u upoznavanju sveta oko nas. Zasnovana je na poznavanju sveta koji nas okružuje, a formalizacijom matematički izraženih odnosa među pojavama beleži ih i interpretira. Zajedničke pojave su one koje se pojavljuju u velikoj količini[5], na primer, nivo glukoze u krvi pojedinaca je zajednička pojava, čiji prikaz pruža statistika. Procena obima vrednosti glikemije za zdravo stanovništvo je rezultat statističke uštede.

Jedinica, koja prilikom zajedničke pojave postaje objekat statističkog istraživanja je **pojava**. To je rezultat eksperimenta ili posmatranja. Kod proučavanja javnog zdravlja se češće srećemo sa pojavama koje su posmatrane na skupu zdravih ili bolesnih ljudi. Eksperimenti - ogledi u ovoj oblasti su relativno retki, posebno imajući u vidu etička ograničenja. Zajedničke pojave nastaju kao rezultat neograničenih ponavljanja posmatranja ili kao rezultat posmatranja na obimnim skupovima uzajamno ravnpravnih elemenata. Zajedničke slučajne pojave su one koje nije moguće u potpunosti tačno predvideti pre realizacije eksperimenta ili posmatranja.

Definicija 3: Čime se bavi statistika

Statistika se bavi organizacijom i pregledom podataka i izvođenjem zaključaka o svim podacima, pri poznavanju samo dela njih.

Za razumevanja pojava, a time i poznavanja funkcije statistike, neophodno je poznavanje pojma **populacija**. Ona predstavlja najveći mogući skup entiteta koji predstavljaju predmet našeg interesovanja u datom momentu. Izvesti zaključak o celini predstavlja sastavni je deo i jedan od osnovnih ciljeva statistike. Kad se izračuna statistička vrednost na osnovu podataka iz slučajno odabranog uzorka, a na osnovu poznavanja više pojedinačnih statističkih vrednosti, možemo zaključiti da ona važi za sve jedinice sa istim karakteristikama kakve je imao izabran skup i onda smo primenili **statističku indukciju** za donešenje zaključaka o populaciji.

U praksi to izgleda tako da, kad želimo da zaključimo kakvu visinu imaju deca u uzrastu od dvanaest godina u Slovačkoj, nećemo ići po celoj Slovačkoj i meriti svu decu od dvanaest godina, već ćemo slučajno izabrati (ali pridržavajući se određenih pravila za vršenje slučajnog uzorkovanja) uzorak dece kojoj ćemo izmeriti visinu. Onda izračunamo statističke vrednosti, koje karakterišu ovaj uzorak. Shodno tome, pod određenim okolnostima i uslovima možemo zaključiti da ovaj broj (izračunata vrednost) predstavlja parametar cele populacije dvanaestogodišnje dece u Slovačkoj. Uradili smo statističku indukciju. Drugim rečima, ovaj proces možemo opisati tako da određenu populaciju ili „univerzum“ možemo da prikažemo određenim brojevima - konstantama, koje sumiraju njene osobine. Kod ovog je jako bitno koliko je temeljno istraživač definisao šta je populacija. Primer populacije ili univerzuma mogu biti svi oboleli od šećerne bolesti ili sa visokim

---

J.Chronic.Dis. 1967;20:511-24.

[5] Nije moguće reći koliko pojava čini zajedničku pojavu. Nije dobro tvrditi da je zajednička pojava više nego 30 slučajeva ili slično.

krvnim pritiskom, kod kojih nas interesuje da utvrdimo zavisnost određenih znakova ili efekata lečenja od drugih faktora, kakvi su na primer oblici ponašanja (ishrana, fizička aktivnost) ili davanje insulina, ili lekova koji snižavaju krvni pritisak. Retko kad imamo priliku da upoznamo ili ispitamo sve članove populacije. Ograničenja proizilaze iz nedostatka vremena, materijalnih sredstava ili jednostavno nemogućnosti da se ispitaju svi. Ovaj fenomen je poznat i kod istraživanja popularnosti pre političkih izbora. Firme koje procenjuju šanse kandidata za uspeh (neuspeh) ne kontaktiraju sve birače, već izaberu tzv. reprezentativan uzorak istih, a u njemu su ljudi različitog društvenog položaja, obrazovanja, pola, pa iz rezultata njihovih odgovora izvode mišljenje o rezultatima na državnim izborima (detaljnije o uzorkovanju govorićemo kasnije). Jasno je da rezultat ovog istraživanja ne mora da bude realizovan kad se saberu glasovi birača, već je informacija dobijena na ovakav način i uprkos netačnosti jako bitna, s obzirom da izborni štabovi političkih partija daju veliki novac na ponavljanje istraživanja ovog tipa.

Koncepcija mišljenja o celini na osnovu **uzorka** populacije je jedan od osnovnih postupaka statističkog saznanja. Praktično u svim poglavljima koje slede bavićemo se tehnikom izvođenja statističkih podataka, dakle **statističkom analizom (inferencijom ili indukcijom).**

Shvatanje statističke indukcije je izuzetno važno radi pravilne interpretacije rezultata statističkog saznanja.

Definicija 4: Statistička analiza

Statistička analiza (indukcija,inferencija) je postupak kojim se izvode sudovi o populaciji na osnovu rezultata dobijenih iz uzorka iz ove populacije.

Biostatistika predstavlja jednu podgrupu, sastavni deo cele statistike. Osim nje, postoji i demografska statistika, statistika u socijalnim naukama, kao i druge primenjene statistike. Matematička statistika predstavlja teoretski osnov primenjenih statistika.

Merenje, promenljiva, parametar, statistička vrednost

U biostatistici sakupljamo i vrednujemo informacije o pojavama kod više osoba, uzoraka i eksperimenata. Pojavu izražavamo kao **merenje**, ili kao **posmatranje**. Ono ne mora da bude iskazano samo brojem, može da bude i vrednost iskaza (npr. da-ne, kategorija ili redosled). Detaljnije ćemo se ovim baviti u narednim pasusima. Svako merenje je otkriveno pojavom kod određenog pojedinca, koga obično nazivamo subjektom statističkog saznanja. Dakle, na pojedincu možemo uraditi razna merenja, kao što su uzrast, pol, visina, telesna masa, sistolni i dijastolni krvni pritisak. Budući da statistika predstavlja metodu obrade grupnih pojava, merenje se izvršava na više subjekata. Zato je bitno merenje raditi i izražavati istim načinom. Ista merenja kod različitih subjekata zovemo **obeležja (promenljive, varijable)**. To može da bude na primer visina zabeležena u centimetrima kod svih subjekata. Nije moguće u okviru jedne promenljive izraziti visinu u centimetrima kod nekih subjekata, a u milimetrima ili metrima kod drugih.

Način izražavanja, na primer jedinica u kojoj se merenje sprovelo i zabeležilo[6], zove se

[6] Prirodno je merenje u raznim jedinicama, ali beleženje je nužno samo u jednoj, kao i izračunavanje u jednoj jedinici

jedinica mere promenljive. U odnosu na jedinicu mere promenljive se definiciji promenljivih poklanja velika pažnja. U toj definiciji se navodi ne samo jedinica mere promenljive, već i način merenja. Mnoga, naročito laboratorijska merenja, mogu se dobiti raznim metodama, npr. merenje holesterola može da bude od uzete krvi iz vene ili tzv. suvom hemijom. Vrednosti se mogu razlikovati, u nekim slučajevima suštinski. Takođe način izražavanja vrednosti merenja u promenljivoj mora se tačno definisati, npr. kod beleženja pola nije bitno da li će se muškarac označiti kao 1 ili 0, ali se moramo toga pridržavati kod svih subjekata.

Definicija 5: Promenljiva

Promenljiva je osobina koja poprima razne vrednosti za razne osobe, mesta i stvari.

Promenljive su osnov statistike. Statističke metode zapisuju promenljive, istražavuju njihove zajedničke odnose, ili predviđaju njihov dalji razvoj. Razlikujemo nekoliko tipova promenljivih, koji se mogu uzajamno kombinovati. **Kvantitativne (numeričke) i kvalitativne** promenljive se razlikuju u mnoštvu merljive informacije. Prve izražavaju osobine koje možemo meriti, na primer toplota, masa, koncentracija natrijuma u krvi. Izražavaju se brojčano i imaju jedinicu mere (u kakvim jedinicama su bili mereni).

Kvalitativne (atributivne, **kategorijalne, nebrojčane) promenljive** izražavaju osobine, koje se mogu izraziti u formi **kategorije**, npr. bolestan/zdrav, bez simptoma/sa simptomima. Među kvalitativnim se izdvajaju **nominalne** promenljive, koje možemo meriti samo u smislu pripadanja nekoj grupi, jednoznačno razlikujući se od ostalih grupa. Ne možemo ih kvantifikovati niti poređati. Na primer, sve što možemo reći o dva merenja (pojedincima) jeste da se razlikuju u vrednostnoj promenljivoj A (na primer razlikuju se u polu), ali ne možemo reći da jedna od njih ima manje ili više ovog kvaliteta. Uobičajeni primeri nominalnih promenljivih jesu pol, nacionalnost i slično. **Ordinalne** promenljive omogućavajuređanje u smislu sadržaja manje ili više kvaliteta, međutim ne omogućavaju da se kaže koliko više. Tipičan predstavnik ovog tipa promenljive jeste socio-ekonomsko stanje porodice. Znamo da viša srednja klasa ima bolje socio-ekonomsko stanje nego srednja klasa, ali ne možemo reći da je razlika 20%.

Među **kvantitativnim promenljivim** se izdvajaju **intervalske** promenljive, koje omogućavaju ne samoređanje merenja, već i kvantifikovanje i upoređivanje veličine razlike među njima. Na primer, toplota u Celzijusovim stepenima predstavlja intervalski skalu. Možemo reći da temperatura 40º Celzijusa, jeste veća od temperature 37,8º, a takođe da je razlika 2,2º Celzijusa. **Srazmerne (racio)** promenljive su jako slične intervalskim, ali sadrže apsolutnu nulu. Primer je temperaturna prema Kelvinu. U svakodnevnoj praksi među njima nema razlike.

Kad već spominjemo lorda Kelvina, ovaj poznati fizičar jednom je rekao da predmet naučnog istraživanja treba da budu samo pojave, koje možemo da merimo[7]. Mislio je tada na kvantitativno posmatranje, koje dominira u fizici. U medicini i uopšte u biološkim naukama to nije slučaj, pre bi se reklo da je istina suprotna. Razlika među ova dva tipa podataka nije tako oštra kako bi moglo da nam izgleda na prvi pogled. U praksi se često kvantifikuju i suštinski kvalitativni podaci. Uobičajena je navika da pol označavamo sa 0 i 1, a takođe prisutnost bakterije u mokraći sa jedan ili više krstića. Manje često se pribegava suprotnom postupku kod obrade podataka, dakle konverziji kvantitativnih podataka u kvalitativne. Kad je potrebno kategorizovati neku pojavu, iskoristiće se pomenuti postupak. Ovakve promene promenljivih zovu se **transformacija** promenljivih. Primer je svrstavanje laboratorijskog nalaza u kategoriju referentnih vrednosti ili

---

mere, npr. dužinu deteta nakon rođenja možemo da merimo u centimetrima, ali ćemo je zapisati u npr. metrima.

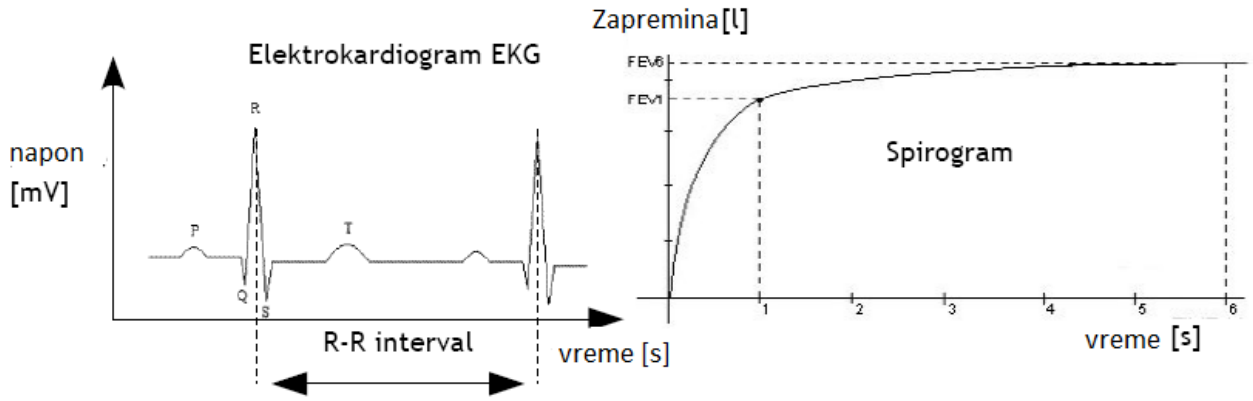
[7] „If you can not measure it, you can not improve it“. Ako nemožete da izmerite, ne možete ni poboljšati. „To measure is to know“. Izmeriti znači znati.

među povišene, odnosno snižene vrednosti. Ovaj postupak se često koristi u klasifikacijskim algoritmima bolesnih stanja.

Slučajne promenljive predstavljaju vrednosti koje su posledica slučajnosti. Samo ove promenjljive treba da budu predmet statističkog saznanja. Suprotno tome je deterministički dobijena promenljiva, gde je slučajnost isključena, a nastaje kao rezultat svrsihodnog rada, gde je verovatnoća postizanja očekivanog rezultata 100%.

Diskretne (prekidne, diskontinuirane) slučajne promenljive su karakterišu prekidima, pauzama, koje indikuju nedostatak vrednosti među određenim vrednostima, npr. vrednost R-R intervala na EKG, vitalni kapacitet u prvoj sekundi izdisaja. **Neprekidne (kontinuirane)** slučajne promenljive nemaju prekide, povezane su, npr. EKG kriva, spirometrička kriva (Slika 1).

Slika 1: Elektrokardiogram sa označenim R-R intervalom i spirogram



Neprekidne slučajne promenljive možemo prevesti u diskretne. Na levoom delu slike vidimo elektrokardiogram. Pojedinim tačkama krive obično dodelimo slovo kao njegovo ime, pa tako udaljenost između dve tačke na dva zapisa predstavlja odraz srčane frekvencije. Obično se navodi udaljenost među vrhuncima R – talasi, tzv R-R interval. Drugi deo slike je beleška promene zapremine izdahnutog vazduha kod ispitivanja funkcije pluća. Za procenu funkcije pluća se obično uzima u obzir zapremina izdaha u prvoj i trećoj sekundi i celokupan izdah vazduha. U oba zabeležena slučaja od zavisne slučajne promenljive dobićemo diskretnu promenljivu. Ponekad (retko kad) diskretne vrednosti prevodimo u neprekidne. U drugom slučaju se često prave greške protivne suštini promenljive. Iako je rast deteta neprekidna radnja, merenje rasta se radi u vremenski određenim intervalima. Najpre merimo svakog meseca posle rođenja, a kasnije jednom u godinu dana.

Nezavisne promenljive su one kojima vrednosti možemo menjati, dok je **zavisne promenljive** moguće samo registrovati ili meriti. Na prvi pogled, ova podela se može činiti obmanjujuće, zato što možemo reći da sve promenljive zavise od nečega. Ova različitost proizilazi iz eksperimentalnog istraživanja, gde su neke promenljive nezavisne od istraživačkog procesa, na primer pol. Na drugoj strani se očekuje da druge promenljive, kao što su telesna visina i masa, zavise od pola, starosti, itd.

Definicija 6: Tipovi promenljivih

Kvantitativne – merljive: • intervalne • srazmerne Kvantitativne: • zavisne • nezavisne	Kvantitativne: • neprekidne • diskretne (prekidne) Kvalitativne (atributivne) – kategorijalne: • nominalne • ordinalne

Sada ćemo preći na rezultate statističkog istraživanja. Razlikujemo rezultate koje se dobijaju na osnovu podataka dobijenih iz celokupne populacije i one koje se dobiju iz izabranog uzorka. **Parametar** predstavlja opisnu meru koja karakteriše promenljivu u populaciji. **Statistička vrednost** je opisna mera izračunata na osnovu uzorka. Razlika između parametra i statističke vrednosti je u načinu izvođenja rezultata statističkog istraživanja: kad se nekom statističkom procedurom izračunava broj koji karakteriše uzorak dobijen slučajnim izborom, onda je to statistička vrednost. Kad se ovaj broj primeni na celu populaciju, onda se pretvara u njen parametar.

Definicija 7: Parametar i statistička vrednost

Parametar predstavlja opisnu meru, koja karakteriše promenljivu kod populacije. Statistička vrednost je opisna mera izračunata na osnovu uzorka.

Populacija

Spomenuli smo pojam **populacija**. Ona predstavlja najveću grupu entiteta koji nas interesuju u određenom trenutku. Predstavlja zato predmet jednog od osnovnih ciljeva statistike, a to je doneti sud o celini. Kad se izračuna statistička vrednost na osnovu podataka iz slučajno odabranog uzorka i na osnovu poznavanja više pojedinačnih statističkih podataka, možemo zaključiti da ona važi za sve jedinice sa istim karakteristikama kakve je imao izabrani uzorak; podrazumeva da smo koristili statističku **analizu** i napravili zaključak o populaciji. U praksi to izgleda tako da kad želimo da zaključimo kakvu visinu imaju deca u uzrastu od 12 godina u Slovačkoj, nasumice odaberemo (ali pridržavamo se određenih pravila za realizaciju nasumimčnog odabira) grupu dece, kojima izmerimo visinu. Onda izračunamo statističku vrednost koja karakteriše ovaj uzorak. Zatim pod određenim okolnostima i uslovima možemo zaključiti, da ovaj broj (izračunata statistička vrednost) predstavlja parametar cele populacije dvanaestogodišnje dece u Slovačkoj. Uradili smo **statističku analizu**. Ovaj postupak je osnovna svrha statističkog proučavanja: na osnovu osobina uzorka donosimo sud o osobini cele populacije. Zato što nikad nećemo znati tačna svojstva, moramo koristiti pretpostavku za njihovo izražavanje.

Uzorak

Iz praktičnih razloga nikad nećemo raditi sa celom populacijom (pogledaj gore). Iz nje odaberemo samo deo, nakon čije obrade izvršimo statističku analizu. Zato je bitno da je uzorak izabran na slučajan način, koji neće dovesti do pristrasnosti rezultata statističke obrade. Ovakav

postupak zovemo **slučajni uzorak iz osnovne grupe**. Slučajnost izbora nam garantuje da svaki element osnovne grupe ima na početku uzorkovanja istu pretpostavku da bude odabran. Jako je bitno održavati pretpostavku jednakih šansi, čime ćemo sprečiti pristrasnost rezultata (bias). Rezultat slučajnog izbora je **uzorak**, koja se sastoji od jasno definisanih elemenata-jedinica statističkog posmatranja, koje u stvari želimo da upoznamo.

Zamislite da želimo da saznamo visinu dece u Slovačkoj. Odabraćemo, međutim, samo decu u gradovima (poznato je da su deca u gradovima viša od dece iz unutrašnjosti), ali rezultat bismo želeli da iskoristimo za celu populaciju Slovačke. Prirodno je da ovaj postupak ne vodi do tačnih rezultata. Ovakvo iskrivljivanje se naziva pristrasnost u izboru ispitanika.

Slučajni uzorak se dobija jednim od sledećih postupaka ili njihovom kombinacijom. Za dobijanje prostog slučajnog uzorka se koriste tabele slučajnih brojeva. One su deo većine statističkih tabela, a njihovo korišćenje je vrlo jednostavno: kod merenja dvanaestogodišnjaka u školi uzmemo spisak sve dvanaestogodišnje dece i pomoću ove tabele odaberemo one, kojima redosled u spisku odgovara broju u tabeli slučajnih brojeva.

Drugi postupak dobijanja slučajnog uzorka je postupak nazvan **mehaničko-sistemsko uzorkovanje**, a radi se prema obeležju koje nije povezano sa obeležje koje se ispituje. Na primer, odaberemo samo onu decu kojima je matični broj deljiv sa tri bez ostatka. Matični broj i njegova deljivost brojem tri sigurno nisu u vezi sa visinom dece.

Sledeća mogućnost je koristiti stratifikovani uzorak u kome se grupa deli na podgrupe, a dalje se radi pomoću tabela slučajnih brojeva. Decu možemo podeliti, na primer, prema socio-ekonomskom stanju porodice iz koje dolaze. To nam garantuje da ćemo imati iz svake grupe dovoljan reprezentativni uzorak. Uopšteno je poznato da socio-ekonomsko stanje porodice utiče na visinu deteta.

Često korišćen je i parni izbor (sparivanje, mečovanje). Ova metoda je zasnovana na izboru jedinica sa zajedničkim oznakama, npr. napravimo parove od pacijenata pre i posle lečenja.

Uvod u statistički program R

Program sa nazivom R je jezik i okruženje za statističke proračune i statističko grafičko prikazivanje i nastao je 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 u mnogome sličan jeziku i okruženju S, odnosno softveru S-plus, koji se prodaje. Ova sličnost proizilazi iz odluke izvornih stvaralaca projekta R, da se u njegovom stvaranju pođe od jezika S. Većina zadataka iz jezika S bez promena u potpunosti funkcioniše i u okruženju R (sa olakšanjem se ponekad za ova dva programa govori kao o dva dijalekta).

R se naziva okruženje i jezik naročito zato što je jako fleksibilan i relativno jednostavno prilagodljiv za bilo koju specifičnu vrstu analize, za razliku od drugih programa za statističku analizu, koje broje skup operacija koje su u suštini nepromenljive. Ovaj fleksibilan sistem je zasnovan na paketima, koji po pravilu sadrže komande za statističke operacije u određenoj specifičnoj oblasti. Tako, na primer, paketi *epibasix* ili *epicalc* sadrže komande specifično namenjene za statističku analizu u epidemiologiji, a paket *survival* sadrži komande namenjene za različite oblike analize proživljavanja u epidemiološkim i zdravstvenim istraživanjima. Prilikom instaliranja okruženja R, automatski se instalira nekoliko paketa koji sadrže standardna statistička izračunavanja i grafički prikaz – tzv paket *base*. Na raspolaganju je onda niz specifičnih paketa komandi, koje je potrebno instalirati u sistem pre njihove upotrebe. Na osnovu licence GNU-GPL svako može da napravi paket komandi i stavi ga na raspolaganje za korišćenje zajedno sa već postojećim paketima.

Okruženje R se od mnogih klasičnih programa za statističku analizu razlikuje i time što je osnova njegovog funkcionisanja komandna linija, odnosno red u koji se zadaje komanda, a R uradi zadato naređenje. Iako će neke od komandi, uprkos svome jednostavnom značenju, uraditi relativno kompleksna izračunavanja, kod tipičnih analiza potrebno je zadati sled nadovezujućih komandi i argumenata, da bismo dobilli traženi rezultat. Na ovakav načinom se napravi *script* ili grupa komandi, na osnovu čega R uradi tražena naređenja, a mi dolazimo do rezultata.

Za pisanje skripta okruženje R uključuje jednostavan tekstualan editor (R editor), koji omogućava pisanje i ređanje skripti u obliku modela, koje je moguće bilo kad ponovo otvoriti i koristiti. Grafički i funkcionalno savršenija alternativa R editoru je softver *Tinn-R 1*. Ovaj program je izvorno nastao kao alternativa *Notepadu*, koji je distribuiran kao deo operativnog sistema *Windows*. U savremenom dobu primarno je namenjen da bude editor skripti za okruženje R. Posle instaliranja moguće ga je jednostavno integrisati sa instaliranom okolinom R, a ima mnogo komandi koje pojednostavljuju i čine efektivnijim pisanje skripti. Isto kao i R, distribuiran je pod licencom GPL.

Okruženje R je za epidemiologiju i istraživanja u javnom zdravlju ili u medicini odlična alternativa za rutinski dostupne komercijalne softvere. Dokaz su desetine dostupnih paketa komandi specijalno napisanih za ovu oblast istraživanja, koje omogućavaju potpunu statističku analizu i grafički prikaz podataka i rezultata od jednostavnih deskriptivnih statističkih radnji, pa sve do kompleksnih prognostičkih modelovanja, ili postupaka odlučivanja u medicini.

Primeri navedeni u nekim poglavljima u ovom tekstu bili su napisani i rešeni u okruženju R. Sve praktične vežbe i zahtevi potražuju instalaciju okruženja R i adekvatnih paketa.

Pre instalacije paketa potrebno je skinuti program za instalaciju kao i upustvo za registrovanje programa u sistemu prema tome kakav sistem koristite. Pojedinosti ćete naći na više izvora, na primer u izvanrednoj knjizi koja navodi osnovne statističke postupke u R[8], kao i u knjizi čiji je uvodi deo dostupan na internetu[9]. Problem navedene literature je u engleskom jeziku, iako ga korisnik neće moći izbeći u programu R. Od slovačkih ili čeških izvora osim knjiga namenjenih za napredne statističare[10], dostupno je više internet stranica[11], [12].

Primeri

U pojedinim poglavljima navodimo primere za provežbavanje postupaka u R. Program R sadrži datoteku podataka. Ona će biti jedan od izvora, dok će drugi biti paket (package) od autora poznate knjige o korišćenju R Dalgarda (2008), koji je pružio javno dostupan paket *IswR*. Dok je u datoteci programa R puno podataka koji se tiču raznih oblasti ljudskog interesovanja, a relativno malo iz oblasti zdravlja i zdravstva, *IswR* sadrži nesrazmerno više primera upravo iz oblasti našeg interesovanja. Koristićemo rešenja iz obe datoteke.

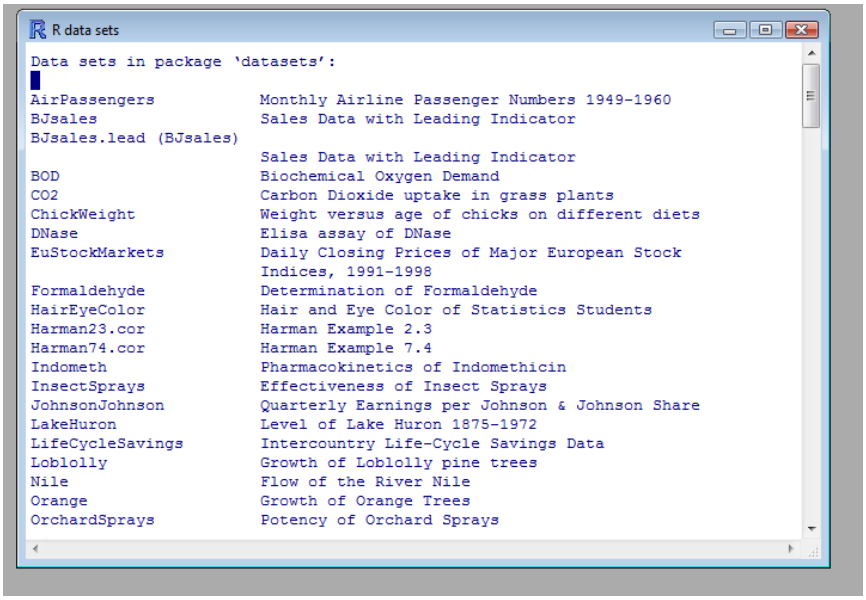
Spisak primera sa podacima o sadržaju direktno u programu R omogućava se dozivanjem komande *data()* bez argumenata. Posle njenog dozivanja otvoriće se prozor sa nazivima datoteka. Ove poslednje pogledajte pozivanjem iste komande, ali kao argument koristite naziv datoteke.

---

[8] Peter Dalgaard. Introductory Statistics with R. Springer, 2nd edition, 2008. ISBN 978-0-387-79053-4.
[9] Rob Kabacoff. R in Action. Manning, 2010. http://www.manning.com/kabacoff
[10] Stano Pekar and Marek Brabec. Moderni analyza biologickych dat. 1. Zobecnene linearni modely v prostredi R. Biologie dnes. Scientia, Praha, 2009. ISBN: 978-80-86960-44-9
[11] http://www.r-project.cz/index.html
[12] www.karlin.mff.cuni.cz/~**komarek**/Rko/Rmanual1.pdf

Sadržaj datoteke ispisaće se pozivanjem komande *list()* sa nazivom datoteke kao argument.

Slika2: Posle skidanja paketa *IswR* on će se instalirati kao datoteka i može da se koristi. Znak # se koristi za komentar, a tekst koji sledi za ovim znakom program R ne obrđuje.



Sažetak

Statistika ima dve osnovne poruke u životu naučnika i istraživača, a samim tim i studenta koji priprema diplomski rad. Zadatak da zabeleži naveden stav, kao i da pomaže pri zaključivanju o populaciji iz koje su se podaci uzeli sa ciljem naučnog istraživanja. Zato je ovo uvodno poglavlje posvećeno i karakteristikama i vrstama podataka i opisuje način kako se njima rukuje. Za obradu podataka predstavljamo odgovarajući statistički program R, koji je slobodno dostupan - bez naplate i samim tim pogodan za korišćenje u akademskom okruženju.

Autori

Grupu autora su činili su prof. dr Martin Rusnak, koji se zajedno sa svojom suprugom prof. dr Vierom Rusnakovom tokom cele svoje stručne karijere posvetio istraživanju u medicini i zdravstvu. Oboje predaju statističke metode za lekare, studente medicine, javnog zdravlja i sličnih struka, a takođe koriste statističke metode za svoja istraživanja. Nastavne tekstove iz ove oblasti objavili su na strani Katedre javnog zdravlja Trnavskog univerziteta.

Treći autor dr Marek Majdan je zaposlen na fakultetu Janog zdravlja i socijalnog rada Trnavskog univerziteta. Od početka rada na fakultetu, već kao doktorant bavio se problematikom statističkog saznanja u oblasti javnog zdravlja. Zbog nedostupnosti drugih kvalitetnih statističkih programa počeo je da koristi statističko okruženje projekta R. Pokazao je njegove izuzetne sposobnosti prilikom analize podataka više istraživačkih projekata na fakultetu, ali i u inostranstvu. Koristi okruženje R u nastavi i preko projekta MOODLE.

Pitanja

1. Na koje od navedenih pitanja ćete tražiti odgovor pomoću statističkih postupaka?
(a) Ko je od vaših školskih drugova najviši?
(b) Postoji li razlika u telesnoj masi školskih drugova, onih koji puše i onih koji ne puše?
(c) Kakav je učinak imalo davanja leka kod jednog pacijenta?

2. Objasnite pojam statističke analize.

3. Da li je moguće merenje telesne temperature kao neprekidne promenljive? Kako je merimo u praksi?

4. Da li je vrednost šećera u krvi neprekidna ili diskretna promenljiva?

5. Imate li program R instaliran u kompjuteru?

6. Grafički predstavite podatke iz uzorka *malaria* u paketu *IswR*.

DRUGO POGLAVLJE

Sakupljanje i obrada podataka

Sadržaj poglavlja

Cilj poglavlja

Postupak obrade podataka koji se koristi u istaraživanju, bez obzira da li se radi o istraživanju u javnom zdravlju, epidemiologiji, ili ispitivanju lekova, bitan je sastavi deo statističkog postupka. Ovaj postupak primenjuje sve metode koje se koriste i u ostalim naukama, ne samo u nauci o zdravlju. Za razliku od tehničkih nauka, bliža mu je deskripcija nego korišćenje dubljih matematičkih analitičkih metoda. Budući da želimo da ova publikacija bude što više praktično uputstvo za rešavanje zadataka, nemamo prostora za pojedinosti teoretskog razmatranja predlaganja i sprovođenja istraživanja. Čitaocu, koji se dublje interesuje za navedenu problematiku, poručujemo odgovarajuću stručnu literturu.

U poglavlju ćemo predstaviti proces obrade podataka, objasnićemo korake koji slede i naznačićemo kakva je njihova primena u praksi. Zatim sledi deo koji govori detaljnije o prikupljanju podataka i njihovom prezentovanju u formi prikladnoj za statističku obradu. Navešćemo takođe način prezentacije sakupljenih podataka za prvobitnu izradu onoga, o čemu ti podaci mogu potencijalno da govore. U zaključku ćemo se dotaći najčešćih grešaka i zabuna koje se prave u ovoj oblasti.

Tabela 1: Ciljevi poglavlja

- Opisati proces obrade podataka
- Predstaviti način prikupljanja podataka
- Prodiskutovati dobre i loše strane različitih načina prikupljanja podataka
- Ukazati na značaj kvaliteta podataka
- Ukazati na probleme bezbednosti podataka

Proces obrade podataka

Proces obrade podataka predstavlja niz od nekoliko koraka. Ovaj proces nije pravolinijski, već u sebi sadrži izgrađen mehanizam kontrole tačnosti urađenog. Slika 1. predstavlja određene korake ovog postupka. Obratite pažnju, koraci su prikazani u pravougaoniku gde treba nešto da se uradi, a tamo gde treba da se opredelimo između dve mogućnosti koraci su prikazani u rombu. U poslednjem navedenom po pravilu postoje dve mogućnosti, ili je nešto urađeno tačno (tako da je odgovor Da) ili netačno (Ne). Dolazi do razgranavanja aktivnosti prema tome kako se ispitivač opredeli.

Predlaganje naučnoistraživačkog projekta

Kako započeti istraživanje (diplomski rad, naučnoistraživački projekat)? Prvi korak, definicija problema - preciziranje zašto se bavimo datom oblašću, kao i postavljanje modela, je osnov bilo kog naučnog postupka zasnovanog na proučavanju podataka. U ovim koracima se formulišu praznine u našem znanju, ciljevi našeg istraživanja, hipoteze i način njihovog potvrđivanja ili odbacivanja. Zainteresovni za naučnu filozofiju naći će odgovore na pitanja koja će se rešavati u ovom koraku kod mnogih filozofa modernog mišljenja, npr. Karla Popera. Ali, šta mora da se uradi u ovom koraku? U prvom redu moramo da upoznamo problem i definišemo **ciljeve –** da damo odgovor na pitanje **šta želimo da dokažemo.** Ali, ne samo jednostavno - da želimo da napišemo diplomski. Ovde istraživač i student moraju dobro da se potrude. Istraživač osim objašnjenja šta u stvari namerava da uradi, po pravilu mora da ubedi nekoga, recimo donatora[13], da u njegovo istraživanje treba investirati. Analogno tome, student mora da ubedi nastavnika da rezultati studentovog istraživanja dokazuju da je student sa studija poneo određeno znanje i veštine.

Ali, kako to da uradimo? Nije na odmet koristi uopšteno primenljivo jednostavno uputstvo za rad, koje navodi nekoliko pravila: u prvom redu kod definisanja važećeg problema i ciljeva neophodno je prostudirati dostupnu literaturu i saznati kako je neko drugi rešio sličan problem. Retkost je da se problem nikada pre toga nije rešavao, iako se, prirodno, to dešavalo u drugim okolnostima. Takođe nije dobro koncentrisati se samo na izvore sa interneta, sa izuzetkom dostupnih renomiranih indeksnih baza i u njima navedenih stručnih, provernih izvora (MEDLINE, COCHRANE, HON sertifikovane stranice, akademski izvori). S obzirom na liberalnost interneta, mnoge pronađene informacije nisu nužno naučno poverene i mogu da budu obmanjujuće. Da ne treba da se kopira rad drugih uz pomoć „copy and paste“, odnosno da se treba kloniti plagijatorstva, ovde ne treba ni posebno naglašavati. Korisna preporuka je i orijentisati se na problem koji je dobro definisan. Ovde u većini slučajeva studente posavetuje profesor koji ima iskustva u ovoj oblasti. Takođe mnogo organizacija, koje pružaju istraživačke grantove, navodi slične savete kako napredovati u predlogu, ali i upravljati projektom[14]. Sledeće pitanje, na koje treba da damo odgovor kod planiranja istraživanja je kako ćemo napredovati dalje, kako ćemo doći do očekivanih rezultata, odnosno koje metode ćemo primeniti. Bitan deo metoda je primena kvantitativnog pristupa, što nam omogućava satistika.

U sledećem koraku potrebno je definisati **plan prikupljanja podataka i statistički model**. U planu će se odrediti koraci pri prikupljanju podataka, specifikovaće se pojedine promenljive. Bitno je takođe odrediti veličinu uzorka[15] i način dobijanja uzorka. U ovom koraku se najčešće pravi greška u tome, da se ne formuliše dovoljno tačno šta i kako će se prikupljati. Primer 1 shematski prikazuje pripremu postupka, kad ispitivač ili student pokušavaju da nađu odgovor na pitanje da li pušenje pre trudnoće utiče na razvoj novorođenčeta, merenjem njegove porođajne mase. Kod definisanja **cilja** izvršiće suštinska opredeljenja, da će se baviti samo dvema kategorijama žena, inače će u svoje istraživanje obuhvatiti žene koje nikad nisu pušile, a kod žena pušača samo one koje su pušile redovno. Na taj način su ograničili broj varijanti, izbacili, na primer, kategoriju povremeni ili bivši pušač. U ovoj etapi bi bilo dobro navesti šta uraditi istraživač kad se sretne sa konkretnom ženom, kakva će biti, da li će je obuhvatiti istraživanjem, uvrstiti je u pušače ili nepušače i slično... Takođe, gore navedeno opredeljenje znači da se neće byviti tačnim podacima o

[13]Donor je pojedinac ili organizacija, koji/koja je finansijski spremna da finansijski podrži predloženo istraživanje. U širem smislu donor je i fakultet na kojem nastaje diplomski rad (postoji više oblika podrške – mentorstvo, pristup biblioteci i slično).

[14]Na stranici http://www.sfcg.org/programmes/ilr/proj.pdf odgovori na ova pitanja.

[15]Određivanje veličine uzorka biće objašnjeno u jednom od narednih poglavlja.

broju popušenih cigareta i vremenskim odnosom prema trudnoći. Neće utvrđivati ni da li je majka pušila i tokom trudnoće. Dakle, u ciljevima istraživač donosi bitne odluke, koje će se onda pokazati u sledećim delovima (kolonama) nacrtane sheme.

Primer 1: Postupak za formulisanje projekta

CILJ/CILJEVI	HIPOTEZE	PLAN PRIKUPLJANJA PODATAKA	STATISTIČKI MODEL
Otkriti razliku u telesnoj masi novorođenčadi čije majke nikad nisu pušile i onih čije majke su redovno pušile.	Majke nepušači rađaju decu sa većom prođajnom masom.	- Opservaciona studija; - 120 porodilja koje su došle da rode na određeno odeljenje, uzorak nije potreban; - Prikupljanje podataka: podaci iz zdravstvenog kartona +razgovor sa majkom; - Promenljive: - telesna masa na rođenju [kg] - telesna dužina na rođenju [m] - pol deteta [M/Ž] - pušenje majke pre trudnoće [D/N] - broj cigareta popušenih u toku života [komada] - uzrast majke [godine] - telesna masa majke [kg] - telesna visina majke [m] - dijabetes majke [D/N]	Upoređivanje proseka telesne mase novorođenadi majki nepušača i majki pušača Pridružujući faktori (confounding): uzrast, telesna masa, telesna visina, broj popušenih cigareta kod majke, pol deteta Dijabetes kod majke – razlog isključivanja iz uzorka

U delu **hipoteza** treba što je moguće tačnije definisati pretpostavke. Kako će se kasnije pokazati, ovaj deo direktno ima veze sa definicijom takozvane nulte hipoteze i statističkim procedurama koje će ovu hipotezu potvrditi ili opovrgnuti. Ako se ovom pitanju ne posveti dovoljno pažnje ili se ono potpuno izostavi, nije moguće postići očekivane rezultate ili to može da košta mnogo truda ili resursa (finansijskih, materijalnih ili ljudskih). U trećem delu, prema shemi, odrediće se postupak istraživanja i utvrdiće se ograničenja, npr. u ovom slučaju se utvrdilo da se neće raditi slučajni uzorak porodilja koje su došle na odeljenje. Ovakva odluka jeste opravdana onda, kada postoji dovoljno snažna pretpostavka da porodilje dolaze u porodolište manje-više

slučajno. Ako bi to trebala da bude skupa privatna klinika, onda je pretpostavka slučajnosti minimalna i možemo opravdano da pretpostavljamo da porodilje iz porodica koje su odabrale tu kliniku neće biti reprezentativni uzorak svih porodilja u zemlji. Ali kad se donosilac odluka opredelio da radi sa porodiljama koje koriste usluge obične klinike, onda je verovatno da sve trudne žene u okolini imaju podjednaku šansu da tamo rode i nije za to potreban dalji slučajni uzorak. Takođe smo se opredlili da će uzorak činiti 120 žena. Odluka je doneta na osnovu ili računanja potrebne veličine uzorka ili na osnovu ograničenja, recimo dužine trajanja istraživanja. Bitne odluke koje se odnose na izbor i oblik promenljivih direktno su u vezi sa onim što je učinjeno u proteklim koracima, a ujedno određuju celokupan postupak prilikom dobijanja i obrade podataka. Obratite pažnju kako su jednosmisleno definisane promenljive. Ovako napisane definicije često uštede mnogo truda pri kraju studije, kad već zaboravimo kako smo kodirali neke promenljive, ili u kojim jedinicama mere su iskazane, a kod velikih studija to može poništiti veliki napor, uložen pri prikupljanju podataka.

U poslednjoj koloni se definiše **statistički postupak**, koje će biti objašnjen u sledećim poglavljima. Na osnovu ovako pripremljenih namera studije možemo da pristupimo realizaciji prikupljanja podataka.

U shemi preporučenog osnovnog postupka (Slika 1) je navedena i **potvrda na pilotnim podacima.** Ovaj korak se često izostavi – naročito u malim studijama – na štetu kvaliteta rezultata i rizika kasnog uočavanja pogrešnog usmerenja. Šta, u stvari, znači ovaj korak? Kao što smo već videli, pre nego što se uradi samo prikupljanje podataka, donosi se niz važnih odluka. One u većini značajno utiču na celo istraživanje. Ako je nastala greška u mišljenju u ovoj fazi, može se desiti da prikupljanje podataka neće biti korektno, a ceo projekat neće uspeti. Zato je zgodno posle prikupljanja određenog dela podataka napraviti proveru i, koliko je to moguće, na osnovu njih proveriti ciljeve projekta. Ako se pokaže da postoje bitna ograničenja, neophodno je vratiti se na početak i ceo postupak korigovati.

O samom načinu prikupljanja podataka govorićemo u sledećem delu poglavlja. Obrada podataka pomoću statističkih metoda i interpretiranje rezultata predmet je narednih poglavlja. Pitanja diseminacije dobijenih rezultata, pripremanje publikacije, prezentacije, jesu značajan deo istraživačkog rada, ali prevazilaze cilj i mogućnosti ove publikacije.

Način prikupljanja podataka

Laboratorijski eksperiment uglavnom ne pruža takvu količinu podataka da bi se pojavili problemi prilikom njihovog sakupljanja. I u ovom slučaju dobro predložen protokol olakšaće njihovu kompjutersku obradu. O uspehu celog tima u obimnim istraživanjima u kojima učestvuje veliki broj ljudi često odlučuje način efikasnog prikupljanja podataka. Takva je situacija u obimnim terapeutskim i profilaktičkim istraživanjima, u epidemiologiji, u socijalnoj medicini, odnosno u svim oblastima javnog zdravlja. Ponovljamo da se pojedinci – pacijenti, ispitanici, koji postanu predmet istraživanja, smatraju statističkim jedinicama i čine **odabran** uzorak. Svaka statistička jedinica je nosilac određenih osobina koje želimo da istražujemo. Ove osobine zovemo **promenljive** (obeležja, varijable) ili statistički **znakovi**. Promenljiva kod svake statističke jedinice dobija **vrednost**. Kod jedne statističke jedinice u jednom trenutku promenljiva dobija jednu vrednost. U našem gore navedenom primeru statistička jedinica je porodilja; promenljiva je pušenje i telesna masa novorođenčeta kod dotične porodilje; vrednosti promenljivih koje beležimo su pušačica A, telesna masa novorođenčeta 2900g. Uslov je da vrednosti promenljivih budu praćene na identičnoj statističkoj jedinici – kod nas na jednoj porodilji.

Takođe, treba misliti na ispravno kodiranje podataka. Neki znakovi imaju posebnu ulogu kod kompjuterske obrade podataka. Može se učiniti trivijalnim upozorenje na razlike među isto prikazanih brojeva i slova, iako se prilikom pisanja pisaćom mašinom često zamenjuju. To je slučaj sa slovom „l“ (el) i brojem „1“ (jedan). Slično je i sa „O“ i „0“ (nula). Većina štampača odštampa precrtanu nulu ili je ova uža nago slovo „O“. Ako se netačno zamene navedeni znakovi, kompjuter će to prepoznati kao različite brojeve i neispravan broj u kojem je nula napisana kao „O“ ili jedinica slovom „l“. U tom slučaju većina programa javlja grešku. Bitna razlika je i korišćenje tačke umesto zareza za označavanje decimalnih brojeva. U slučaju da se koristi zarez umesto tačke, kompjuter neće prihvatiti broj i javljaće grešku.

Rasprostranjena upotreba kompjutera, njihova veličina i korišćenje doveli su do promena beleženja rezultata studija ili eksperimenata. Nije uvek potrebno zabeležiti informaciju na papir pre njenog zadavanja u kompjuter. Laptop, PDA (personalni asistent) ili mobilni telefon omogućuju zapisivanje merenja direktno u elektronski format i njihovo prenošenje u kompjuter ili tamo gde će se uzorak obrađivati. Kompjuter u ruci omogućava zapisivanje podataka direktno prilikom intervjua. Internet omogućava da se na isti način prikupljaju podaci uporedo sa više mesta, kao i njihovo slanje u udaljenu bazu podataka.

Pod bazom podataka podrazumevamo kompjuterski program koji omogućava prikupljanje podataka u istoj strukturi tokom dužeg perioda. Rezultat ovakvog programa je jedan ili više zajednički povezanih uzoraka, gde su podaci poređani na primer prema vremenu i mestu njihovog nastajanja. Jednostavne baze podataka su slika upitnika stavljenog u kompjuter. Komplikovanije su rezultat informacionog sistema u ambulanti lekara opšte prakse, u bolnici, u zdravstvenim ustanovama, u statističkim zavodima, u kancelarijama osiguravajućih kompanija. Deo programa su instrumenti za pretragu, svrstavanje, selekciju podataka prema zadatim kriterijumima. Programi za predlog i pravljanje baza podataka su deo svakog Office paketa i mnogih statističkih programa.

Predložene baze podataka omogućavaju veliki broj podataka za naučno istraživanje i proučavanje. Omogućavaju saradnju timu istraživača, često geografski veoma udaljenih. Takođe, troškovi za vođenje kompleksne baze podataka nisu naročito visoki, budući da se podaci generišu kao deo svakodnevne rutine. Na ovaj način brzo nastaju veliki uzorci podataka, koji omogućavaju proučavanje retkih situacija, oboljenja ili intervencija. Pružaju tačne podatke za kliničku praksu i za javno zdravlje. Jedna od loših strana baza podataka kod statističke obrade je sisitematičnost prikupljanja podataka, što u nekim slučajevima isključuje primenu slučajnog uzorka.

Među poznate baze podataka koje pružaju zanimljive statističke podatke pripada i uzorak iz Framingamske studije o rizičnim stilovima života. U ovoj studiji se podaci sakupljaju već nekoliko desetina godina. Svaka zemlja održava bazu statističkih podataka o zdravlju, o demografiji i slično. Mnoge od njih su dostupne besplatno ili putem plaćanja internetom. Statistički podaci o smrtnosti i nekim karakteristikama zdravlja i zdravstvenoj zaštiti za zemlje udružene u Svetsku zdravstvenu organizaciju za Europu nalaze se u bazi podataka *Health For All*[16].

Otkrivanje uzroka nastanka hroničnih oboljenja traži dugoročno praćenje obolelih. Osim dispanzera, sve više se koristi registar sa određenim tipom oboljenja. Registar je više nego kartoteka, ili dispanzer. Osoba koja je svrstana u registar je, u slučaju hroničnog oboljenja, praćena tokom celog preostalog života. Osim beleženja o toku bolesti, prate se i drugi parametri, koji omogućavaju vrednovanje zdravstvene zaštite pacijenta, lečenje i prognozu obolelog. Registar predstavlja organizacionu i tehničku ustanovu, koja ima cilj da omogući izvršenje intenzivne brige o oboleloj osobi, a ujedno i da obavi kliničko, epidemiološko i operativno istraživanje. Drugim rečima, radi se o bazi podataka o osobama sa određenim tipom zdravstvenog problema. Registri se primenjuju za razne bolesti. U Slovačkoj već duže vreme postoji Nacionalni onkološki registar. Radi se o registru

[16] http://www.euro.who.int/HFADB

incidencije i mortaliteta malignih oboljenja. Uporedo je bio izgrađen i nacionalni registar za decu sa insulin zavisnim dijabetesom (dalje samo IDDM). U registru se sakupljaju informacije o oboleloj deci, koje prikupljaju dečji dijabetolozi. Dete se registruje od prvog dana primanja insulina. Informacije se sakupljaju u centralnoj bazi podataka. Obrađuju se sa ciljem da se utvrdi incidencija i prevalencija ovog oboljenja u populaciji dece u Slovačkoj. Uporedo se podaci dostavljaju Svetskoj zdravstenoj organizaciji u okviru programa DIAMOND. Sledeći registri koji su bili osnovani u Slovačkoj su nacionalni registri pacijenata sa urođenom srčanom manom, sa koronarnom bolešću, sa cerebrovaskularnom bolešću, hroničnim oboljenjem pluća, transplantacioni registar, registar pacijenata sa tubelkulozom, registar pacijenata sa zaraznim bolestima, artroplastični registar i registar pacijenata sa urođenim poremećajem u razvoju.

Kada je reč o aktivnostima koje imaju veze sa dobijanjem informacija od ispitanika, obično se govori o **istraživanju**. Možemo ga realizovati na najmanje na šest načina, koje ćemo sada stručno da opišemo. **Pretraživanje literature** je jako bitna aktivnost, naročito u prvim fazama projekta, kad istraživač defifniše problematiku cele aktivnosti. Ovo je jako važno i u završnoj fazi, kad poredi dobijene rezultate sa stavovima u svetu. Internet predstavlja značajnu pomoć kod traženja izvora za literaturu, ali takođe i značajnu opasnost, budući da u mnogim slučajevima ne pruža informaciju o pouzdanosti izvora. Zato manje iskusni istraživači, a naročito studenti, često koriste ovakve neproverene izvore, a samim tim značajno dovedu u opasnost ispravnost rezultata inače dobrog rada. **Razgovori sa ljudima**, stručnjacima, kolegama, ili sa drugim zainteresovanim, izuzetno su bitni na početku novog projekta, jer mogu pružiti informaciju koja nije bila publikovana, a koja ujedno može da se pokaže kao odlučujuća za uspeh projekta. **Fokusirane (ciljane) grupe** predstavljaju formalizovan pristup razgovoru, kada se grupa ljudi okupi na jednom mestu i naročito pripremljen moderator vodi diskusiju na temu projekta. Ova metodologija je preuzeta iz marketinških tehnika otkrivanja stavova konzumenata prema novom proizvodu. Njihova nepovoljnost je mala reprezentativnost grupe. **Direktni intervju** je jako efikasna metoda saznanja, mada zahteva značajno vreme ispitanika ili troškove, u slučaju kada se koriste usluge specijalizovane agencije. Iz ovih razloga se ovaj pristup koristi prvenstveno u slučajevima kad se pretpostavlja da osobe neće odgovarati na druge načine prikupljanja podataka. **Telefonski intervju** ima prednost u brzini kojom možemo da dobijemo informacije od relativno velike grupe ljudi. Preporučuje se da jedan intervju ne traje duže od 10 minuta. Nepovoljnost je u činjenici da je u nekim državama nivo telefonizacije relativno nizak, a domaćinstva koja imaju telefon su pretežno iz srednjeg ili visokog ekonomskog sloja društva, što snižava reprezentativnost izabranog uzorka. Isto tako bitno, potrebno je biti svestan da u većem domaćistvu žene se češće javljaju na telefon nego muškarci. Ovaj faktor možemo odstraniti, na primer, time da će ispitivač pre početka intervjua pita ko je iz porodice poslednji imao rođendan, ili ko će najskorije imati rođendan. Time će iskoristiti faktor slučaja, a odstraniće jednu od mogućnosti izvora pristrasnosti. Telefonski intervju se radi na osnovu unapred pripremljenog upitnika, gde ispitivač čita pojedina pitanja i zapisuje odgovore. Upitnici poslati poštom su zgodni zbog visoke efikasnosti u odnosu na traškove, a time omogućavaju dosezanje velike grupe ljudi. Jedan od značajnih problema je povratnost, tj. koliko će se od poslatih upitnika vratiti popunjeno, a takođe i tačnost odgovora kod komplikovanijih pitanja može da bude sporna. Poslednjih godina se za ovakve svrhe više koristi internet, odnosno elektronska pošta. Ispitivač mora da bude svestan specifičnosti ovog sredstva komunikacije, a takođe i pristrasnosti koja iz toga proizilazi.

Predlog i primena upitnika za prikupljanje podataka

Jedan od veoma čestih načina prikupljanja podataka od ljudi je korišćenje **upitnika**, iako se neretko primenjuje neprikladno. Ukazaćemo na neka pravila za konstrukciju upitnika, koja važe kod slobodne forme prikupljanja podataka. Cilj istraživača je komunikacija sa potencijalnim ispitanicima, za šta se koristi formular - upitnik. Trudimo se da zagarantujemo potpunu razumljivost i saradnju ispitanika. Zato je bitno poznavati suštinu procesa komunikacije. Definisanje sadržaja upitnika u medicini počinje proračunom stavki koje su zanimljive sa stanovišta dobijanja informacije. Mogu da sadrže karakteristike, anamnestičke podatke, simptome, znakove. Predlog sadržaja treba da obezbedi dobijanje što više informacija o ispitivanom problemu, a da ograniči one koje nisu bitne prilikom rešavanja navedenog pitanja. Kod komplikovanijih upitnika potrebno je pre početka istraživanja na uzorku populacije proceniti kvalitet strukture upitnika. Na taj način se uglavnom otkriju slabosti predloženog upitnika i omogućava njegovo dopunjavanje, odnosno uklanjanje nekih informacija. Kreator upitnika će rešiti dilemu između obimnog sadržaja i bojazni da se izostave suštinska pitanja, naspram kratkog i informativno nepotpunog upitnika. Ispitanik ne mora uvek da bude raspoložen da odgovara na veliki broj pitanja istim kvalitetom i potpuno. Zato je dobro predložen upitnik racionalni kompromis između potpunosti i dužine. Svoj zadatak ovde igra i posledična statistička analiza - što više podataka ima upitnik, time je komplikovanija njihova obrada. Dugačak upitnik ima, dakle, i nepovoljnost komplikovanije obrade.

Proces predloga upitnika je moguće razdeliti u tri koraka (Tabela 2).

Tabela 2: Proces odlučivanja prilikom predlaganja upitnika

● Definisanje cilja upitnika
● Izbor pitanja
● Odluka o obliku i formulaciji pojedinih pitanja
● Raščlanjivanje upitnika i postupnost pitanja

Predlog upitnika počinje time da ispitivač razmisli kako da realizuje povezanost između definisanog cilja istraživanja i instrumenta za njegovo postizanje, tj. upitnika. Zato je potrebno izdvojiti pojedina istraživačka pitanja i razmisliti kako ćemo putem njih dobiti činjenice koje su predmet istraživanja. U ovom koraku ispitivač mora da pođe od dobrog poznavanja problematike, naročito iz literarnih izvora. Trebalo bi da se potrudi da nađe upitnik koji je već bio isproban na sličnom problemu ili da kombinacijom već isprobanih upitnika napravi novi. Povoljnosti ovakvog pristupa su brojne. U prvom redu je u ovakvom slučaju proizvod već postoji, čime se vidljivo povećava pretpostavka uspeha projekta. Upitnici iz nacionalnih ili međunarodnih projekata su uglavnom provereni, validirani i omogućavaju upoređivanje sa prethodnim studijama.

Ispitivač se često trudi da prikupi više informacija nego što je potrebno za postizanje cilja. Pretpostavlja da će ih koristiti ili odmah ili kasnije. Ovakav pristup se obično označava kao „datizam“. On je opasan iz tog razloga da zahteva veće opterećenje pri dobijanju potrebnih informacija i donosi potencijalni rizik da se izgubi prvobitan cilj. Što manje prostora se posvećuje pripremi studije, to je veća opasnost da prikupljanje podataka bude opterećeno prikupljanjem disproporcionalne količine podataka. Ako upitnik sadrži, na primer, 20 pitanja, ispitanik će posvetiti svakome od njih veću pažnju nego u slučaju da upitnik sadrži 50 pitanja. Često se dešava da se u upitnik uvedu promenljive čija upotreba nije neophodna. Onda se povećava broj stavki i

bespotrebno se povećava količina podataka za obradu. Desilo se da je lekar tražio da se uvede praćenje podataka o ehokardiografiji, iako se ovo ispitivanje u datoj grupi ljudi radi samo izuzetno. Posle pola godine praćenja čudio se što je samo manje od 1% pacijenata imalo ovo ispitivanje i da ga nije moguće vrednovati. Posle odstranjivanja stavki koje su se odnosile na ovo ispitivanje, upitnik se značajno skratio. Pojednostavilo se i ubrzalo njegovo unošenje u kompjuter bez gubljenja suštinskih problema. Zato se prilikom većih studija, u kojima učestvuje institucija ili istraživač, počinje sa preliminarnim prikupljanjem na relativno malom, pilotnom uzorku. Ovako je moguće odstraniti mnoge greške, a studiju dovesti do eliminisanja nedostataka dobijenih u preliminarnoj analizi.

Kod predloga oblika i formulacije pitanja neophodno je dobro poznavati više različitih načela. Neka od njih su jednostavnost i preglednost. Korisno je izbegavati nepotrebne naslove i brojčano označavanje pitanja. To ne znači da treba izostaviti naslov. Bitno je takođe obratiti se ispitaniku sa objašnjenjem svrhe istraživanja i zahvaliti se za vreme i trud prilikom odgovaranja. Posebno je važno navesti adresu ispitivača, naročito kada odgovore dobijamo bez asistencije ili poštom. U slučaju da šaljemo upitnike poštom, dobro je priložiti i koverat sa ispisanom adresom i poštanskom markicom. Dužina upitnika često utiče na broj odgovora, a u praksi se pokazalo da, čim je ovaj kraći, to su ljudi pre skloniji odgovaranju. Isto tako bitno je da je upitnik što bolje čitljiv, pa treba koristiti najjednostavniji tip pisanja, npr. *Times*. Kurziv rezervišite za instrucije kako popuniti upitnik. Uzmite u obzir i starije ljude, koji možda imaju problem prilikom čitanja sitnih slova.

Ne postoji opšte uputstvo kakav redosled pitanja treba da bude. Neki autori savetuju postepenost od jednostavnih ka komplikovanijih, drugi stavljaju akcenat da izazovu interesovanje odmah u prvim pitanjima. Neki savetuju da se počne sa identifikacijom ispitanika, dok drugi ovaj deo svrstavaju na kraj. Prilikom koncipiranja upitnika savetuje se da se najpre postave pitanja opšteg karaktera, a posle njih specifična. Grananje upitnika bitno štedi trud prilikom odgovaranja, ali kod komplikovanog grananja može da dovede do nepreglednosti i sniženja kvaliteta odgovora. I u ovom slučaju se vidi da je bitno poznavati uopštene karakteristike uzorka na kome se radi istraživanje i pilotnim sakupljanjem uveriti se da je upitnik funkcionalan prema pretpostavkama, odnosno korigovati ga.

Sledeća bitna odluka jeste o **vrsti pitanja**. Na pitanja utiče vrsta prikupljanih podataka (kvantitativni ili kvalitativni), a takođe i odnos među ispitivačem i ispitanikom. Odnos se kreće od relativno neformalnog, ličnog, pa do skroz bezličnih, nepristrasnih upitnika, koje ispunjava ispitanik samosalno.

U formularima se razlikuju dve osnovne grupe pitanja: otvorena i zatvorena. **Otvorena pitanja** (Primer 2) ne nude nikakvav specifičan odgovor. Ispitanik odgovara slobodno, svojim rečima, a ceo njegov odgovor se beleži. Učinak ove vrste pitanja značajno povećava jasna formulacija pitanja. Precizno postavljeno pitanje redukuje obim odgovora. Široko formulisano pitanje, na drugoj strani, može istraživaču da donese nove impulse.

Primer 2: Otvorena pitanja

- Koje lekove ste uzimali u poslednjih pola godine: .....
- Pri otklanjaju kojih teškoća su vam pomogli ovi lekovi: .....
- Na koji način ste došli do lekova koji se dobijaju samo preko lekarskog recepta? .....

Prilikom obrade odgovora se ne može izbeći značajan trud koji treba uložiti u njihovu

transformaciju za dalju obradu. Ovaj proces iziskuje da ispitivač pogleda svaki odgovor i od njih napravi grupe sa istim ili sličnim odgovorima (Primer 3). Prilikom transformacije ispitivač je koristio kodiranje navedeno u tekstu primera. U prva tri slučaja kodiranje je prilično jednostavno, ali u ostalim slučajevima nije bilo tako jasno, budući da učitelj u osnovnoj školi može da ima ne samo srednjoškolsko, već i i visokoškolsko obrazovanje. U ovakvom slučaju na istraživaču je da se odluči o izbacivanju te stavke ili da pokuša da sazna tačan odgovor. Otvorena pitanja su zato češće izvor netačnosti ili dovode do nedostajućih podataka.

Primer 3: Transformacija slobodnih pitanja na stavke koje se obrađuju statističkim postupkom: 1 – osnovno obrazovanje, 2 – srednjoškolsko, 3 - visokoškolsko

Ogovor na slobodno pitanje: Koje najviše obrazovanje ste postigli?	**Transformacija**
● univerzitetsko	3
● završio sam za stolara	1
● ja sam inžinjer statike	3
● učitelj sam u osnovnoj školi	2

Klasifikciju i interpretaciju podataka opterećuje greška koju unosi administrator upitnika. Kako se ponavljajući pokazalo, razni stručnjaci mogu da interpretiraju to isto otvoreno pitanje različito. Ograničenja ovoga tipa pitanja se pojavljuju prilikom kontrolisanja dobijanja medicinskih činjenica, naročito u situacijama gde je potrebno kombinovati ili uporediti odgovore ispitanika. Značajnu ulogu ova ograničenja imaju kod predloga studije, gde u početnim fazama pomažu da istraživač upozna situaciju sa više strana. Na osnovu ovakvog poznavanja možemo da pripremimo ciljana zatvorena pitanja, bez bojazni od gubljenja informacija.

Zatvorena pitanja mogu da se razlikuju po obliku. Ispitanik u svim slučajevima mora da bira lični odgovor od nekoliko ponuđenih odgovora (Primer 4). Pitanja ovog tipa su često dihotomna, dakle sa odgovorima DA ili NE, ili neki drugi par, na primer VISOK/NIZAK. Ponekad se dodaje treća mogućnost, kao što su NE ZNAM ili NIJEDNA.

Primer 4: Primeri zatvorenih pitanja

Da li uzimate lekove svaki dan?	[DA]	[NE]
(Ako ste odgovorili NE, izostavite sledećih 6 pitanja)		
Da li lekove protiv glavobolje uzimate svaki dan?	[DA]	[NE]
Da li lekove za spavanje uzimate svaki dan?	[DA]	[NE]
Da li svaki dan uzimate lekove za smirenje?	[DA]	[NE]

Za obradu podataka davanje treće mogućnosti nije uvek prednost. Korišćenje treće mogućnosti je delimično obrazloženje u pojedinim upitnicima gde mora da postoji sigurnost da je ispitanik odgovorio na svako pitanje. Kad nedostaje odgovor na dihotomno pitanje, nismo sigurni da li se ova stavka previdela. Uvrštavanjem treće mogućnosti dobijamo tu sigurnost.

Na osnovu nekog odgovora na dihotomno pitanje moguće je razgranati upitnik. Primer

dihotomnog pitanja sa razgranavanjem pokazuje kako možemo da koristimo razgranavanje radi jednostavnijeg rada sa upitnikom. To dovodi do štednje snage ispitanika kod ispunjavanja i do jednostavnije obrade. Međutim, nepovoljna činjenica je da se u slučaju greške u odgovoru dobije potpuno suprotna informacija.

Neke nepovoljnosti prethodnog tipa formulacije pitanja otklanjaju se pitanjima sa širokim spektrom odgovora. Obično se označavaju kao pitanja višestrukog izbora. Jednostavan primer je pitanje sa spiskom odgovora u kome ispitanik označi svoj izbor (Primer 5).

Primer 5: Primer pitanja sa višestrukim odgovorima

Označite krstićem u kvadratiću odgovarajući odgovor, odnosno odgovore.

Da li uzimate svakodnevno lekove

- protiv bolova u glavi [ ]
- protiv visokog krvnog pritiska [ ]
- za spavanje [ ]
- za smirenje [ ]

Ova vrsta pitanja ujedno ukazuje ispitaniku na sve moguće odgovore, kao i na one kojo bi mogli da mu promaknu. To je naročito prednost primene ovg tipa pitanja u medicini, gde se pojavljuju neuobičajeni termini. Ispitaniku može da se neki odgovor učini adekvatnim, iako to ne odgovara stvarnosti. I previđanje neke od mogućnosti ubraja se među nepovoljnosti ovoga tipa pitanja.

Drugi oblik pitanja sa višestrukim izborom je **skala kvantifikacije**. U njoj ispitivač definiše odgovarajuću gradaciju odgovora, koja bi trebalo da predstavlja pravilnu lestvicu na kumulativnoj skali (Primer 6).

Primer 6: Gradacija odgovora skalom

Da li imate bolove u glavi (*označite samo jedan odgovor*)

- izuzetno ili nikad [ ]
- ponekad [ ]
- često [ ]
- skoro stalno [ ]

Drugi način izražavanja skale nudi takozvani Likertov dijagram (Likertova skala), (Primer 7). U upitniku se mogu kombinovati tipovi pitanja.

Primer 7: Vrednovanje skalom i ujedno primer predloga upitnika

Tabela za procenu vodiča

a. Obim i svrha

OBIM i SVRHA	
1. Opšti ciljevi stručnih preporuka su jasno navedeni	U potpunosti se slažem 4 3 2 1 U potpunosti se ne slažem
Komentar:	
2. Klinička pitanja obuhvaćena preporukama su jasno objašnjena.	U potpunosti se slažem 4 3 2 1 U potpunosti se ne slažem
Komentar:	

Izbor pitanja se mora podrediti cilju uštede vremena i interesovanju ispitanika. Ovome mora da odgovara i **tekst pitanja.** Ispitivač mora da bira između dva krajnosti: zdravstveni radnici koji će obrađivati odgovore hteli bi da imaju što tačnije formulacije sa korišćenjem svakodnevne stručne terminologije. Ispitanicima, međutim, treba formulisati pitanja pomoću jednostavnog laičkog jezika. Ujedno se ovde krije i opasnost da neće doći do iste interpretacije termina u navedenim grupama. Međutim, mnogi pacijenti koji boluju od određene bolesti poznaju sadržaj termina koji se tiču njihovog stanja. Rešenje navedene dileme jeste relativno jednostavno: koristiti alternativne izraze. Na primer: „*Da li ste nekada imali žuticu (požutele oči ili požutelost kože)?*". Prilikom predloga upitnika bolje je pretpostaviti manje poznavanje ispitanika iz date oblasti. Sofisticirana terminologija nije jedina barijera u komunikaciji. Korišćenje neprecizno definisanih ili višeznačnih izraza (često, ponekad, malo, puno) povećava netačnost odgovora.

Pozitivne i negativne strane primene upitnika

Uopšteno možemo konstatovati da upitnik ima puno prednosti, ali istraživač mora da bude svestan i ograničenja i prepreka. U sledećem pregledu (Tabela 3) pokazaćemo one najbitnije.

Tabela 3: Upoređivanje povoljnosti i nepovoljnosti upitnika

Povoljnost	Nepovoljnost
● Manja opterećenost u poređenju sa intervjuom; ● Prigodni za veće studije; ● Lako možemo da ih analiziramo; ● Mnogi softverski paketi nude podršku zadavanja i tabulacije podataka; ● Snižavaju pristrasnost time što svako dobija isto pitanje; ● Većina ljudi ima prethodno iskustvo sa popunjavanjem upitnika; ● Istraživač nema mogućnost da utiče na odgovor, bilo verbalnim bilo neverbalnim načinom; ● Ispitanik ne mora da menja svoju satnicu, može da popuni upitnik onda kada njemu odgovara.	● Povratnost (niska), koja može drastično da utiče na kvalitet rezultata; ● Niska fleksibilnost upitnika, a time gubljenje informacija o ličnim stavovima ispitanika; otvorena pitanja smanjuju ovu nepovoljnost nauštrb više diskusije; ● Osoba koja je dobila da popuni upitnik ne mora da bude ista ona osoba koja ga popunjava (roditelj u ime deteta, žena u ime svog supruga); ● Nedovoljno obrazovanje značajno snižava kvalitet odgovora.

Sledeća tabela će nam pokazati upoređivanje tipova pitanja (Tabela 4).

Tabela 4: Upoređivanje povoljnosti otvorenog i zatvorenog tipa pitanja

Otvorena pitanja ● Omogućavaju upoznavanje širokog sadržaja mogućih tema, koje proizilaze iz date problematike. ● Moguće je koristiti ih i tada, kada nije moguće sastaviti iscrpan spisak alternativnih odgovora. **Zatvorena pitanja** ● Odgovara se na njih brzo i lako; ● Smanjuju nepovoljnost onima koji nisu navikli da čitaju i pišu (kad ispitanik sam odgovara) ili onima koji nisu navikli da pričaju (kod intervjua); ● Lako se kodiraju, beleže i statistički obađuju; ● Lako se interpretiraju rezultati.

Ako se upitnik obrađivati kompjuterom (šta je danas gotovo uvek tako), potrebno je upamtiti pravilo za pojednostavljenje procesa zadavanjem, na primer time da se pre odgovaranja pripreme numerisani kvadratići. Ako možemo da pretpostavimo da će se koristiti čitač formulara oblikovanje upitnika je jedna od najbitnijih pretpostavki prikupljanja podataka bez grešaka.

Odlučuje ne samo oblikovanje, nego i kvalitet papira i pisanja. I u ovome je značajno pre umnožavanja formulara posavetovati se sa stručnjakom. Posle unošenja podataka na disk slede procedure prethodne oobrade podataka. Ovde svrstavamo kontrolu zadavanja, odstranjivanje pogrešnih podataka, transformaciju podataka, označavanje nedostajućih podataka.

Kvalitet podataka

Jedna od često diskutovanih tema u medicinskoj statistici je kvalitet podataka. Na osnovu nekvalitetnih podataka ni najkomplikovanije postupke neće moći da objasni prihvatljiv zaključak. Jedna od bitnih pretpostavki kvalitetnih podataka je njihova homogenost. To znači da uslovi pod kojima su dobijeni moraju da budu jednaki za sve slučajeve. Dobro predložene studije definišu okolnosti pod kojima je dozvoljeno da se rade merenja. Primer može da bude standardizacija beleženja i kodiranja EKG kod epidemiološke obrade kardiovaskularnih oboljenja, kako to definiše Minesotski kod. On definiše sredinu u kojoj možemo da beležimo EKG, u kakvom stanju mora da bude ispitanik – koliko dugo pre merenja ne sme da puši, jede, raditi naporan posao, koliko dugo pre merenja mora da se odmara, itd. Slična striktna pravila se predlažu i kod merenja krvnog pritiska. To sve treba da obezbedi odstranjivanje interferencije onih faktora koji bi mogli da iskrive rezultate studije.

Ni u najbolje predloženoj studiji se ne mogu izbeći nedostajajući podaci u nekim merenjima. Često ne možemo da dobijemo neke od podataka. Može da se desi da će se razbiti epruveta sa krvlju ili sa mokraćom ili da pojedinac odbije da odgovori na neko pitanje. U ovakvim slučajevima nije uvek neophodno isključiti ostale podatke tog pojedinca iz vrednovanja. Podatke koji nedostaju obrađuju se prema određenim pravilima koji minimalizuju grešku. Naravno, to je moguće do onog momenta kada podaci nedostaju kod svih ili kod većine posmatranih jedinica. Podatak koji nedostaje ne možemo da zamenimo sa nulom, zato što i nulta vrednost utiče na rezultat statističke analize. Ne možemo ga izostaviti, jer mnogi programi automatski stave u prazno mesto nulu. Podatak koji nedostaje moramo da označimo nekim dogovorenim simbolom. Često se koristi „*“ (zvezdica), ili samo „.“ (tačka), odnosno drugi dogovoreni simbol. Dobar statistički program ne uzima u obzir prilikom proračuna ovaj podatak, i upozori da je u podacima određeni broj nedostajućih podataka.

Ako su svi podaci svrstani u formular, onda je potrebno saznati da se nije pojavila greška prilikom zadavanja. Kod brojčanih podataka se često pojavi greška koja izmeni vrednost, da se na neki način izmakne iz rutinskog sadržaja podataka. Ilustrovaćemo to na jednostavnom primeru, u kojem je istraživač utvrđivao pol, starost, visinu i masu kod grupe ljudi, a rezultate je zapisao u tabelu (Primer 8) programa OpenOffice Calc[17] i hteo je da proveri da li nije napravio grešku. Prirodno je da kod tako malog uzorka nije problem uraditi kontrolu kvaliteta i bez pomoći programa, ali u ovom slučaju želimo da ilustrujemo postupak koji je potreban u slučaju većeg broja podataka. Inicijale osoba uključenih u uzorak nije moguće prekontrolisati na moguću grešku. Pol je ispitivač odlučio da kodira sa slovom M za muškarca, a slovom Z za ženu. Jednostavnim saznanjem koliko slova M i Z imamo u uzorku uverićemo se o tačnost, odnosno pogrešnost kodiranja pola (Primer 9). Koristićemo pritom komandu COUNTIF(obim;uslov). Budući da znamo da imamo zajedno 8 merenja, rezultat 7 ukazuje na činjenicu da jedno merenje nije tačno kodirano. I stvarno, u merenju sa inicijalima R.M. kao pol je navedeno slovo Y umesto M ili Z. Kako se dakle opredeliti, da li je

[17]Koristimo tzv. open source program, koji je moguće skinuti sa interneta besplatno. Štaviše, većina prikaza je ista ili slična sa prikazom Microsoft Office Excel.

reč o muškarci ili o ženi? Pomoću inicijala (ili drugom identifikacijom ako je imamo) nekada imamo mogućnost da se vratimo nazad na datu osobu i ispravimo grešku. Ako takvu mogućnost nemamo, onda radimo sa podatkom kao nedostajućom vrednosti ili izbacimo ceo subjekat iz obrade ili shvatimo da se ova greška često dešava kod korišćenja domaće tastature, kad taster označen sa Z u stvarnosti zapisuje Y i obrnuto. Kad ispitivač odluči da sabere određen rizik greške prilikom poslednjeg pregleda, može grešku da popravi.

Primer 8: Podaci o polu, utrastu, visini i masi sakupljeni i zapisani od strane ispitivača

Inicijali	**Pol [M/Z]**	**Uzrast [godine]**	**Visina [m]**	**Telesna masa [kg]**
J.M.	M	54	1,56	89
F.R.	Z	34	1,89	98
H.K.	Z	30	1,67	6O
E.H.	M	45	1.78	88
F.W.	M	48	1,76	72
R.M.	Y	3	1,82	105
Z.Č.	Z	99	1,66	73
L.B.	M	36	1,73	82

U koloni koja sadrži podatke o uzrastu, navedena je minimalna starost 3 godine, signalizirana komandom MIN (obim). Budući da istraživač nije radio sa decom, odmah mu je bilo jasno da u spisku ima grešku. Kod njenog ispravljanja ima slične mogućnosti kao u prethodno navedenom primeru. U dve poslednje kolone vidimo broj merenja 7, što signalizira grešku. U merenju E.H. u promenljivoj kod visine je za decimale korišćena tačka umesto zareza, što je program vrednovao kao tekst, a ne kao broj. U merenju H.K., pak, bilo je korišćeno slovo O umesto broja 0, sa istim tim efektom. CALC program nudi takođe komandu DATAPILOT, koja mnoge od navedenih operacija uradi jednostavnije, ali njegovo korišćenje zahteva bolje poznavanje postupaka u programu.

Kontrolu zadavanja podataka, a time i izbegavanje mnogih grešaka, možemo da uradimo i direktno u programu, tako da u pojedinim poljima koristimo komandu podataka VALIDATE. Pomoću postavljanja uslova možemo lako da kontrolišemo zadate vrednosti i izbegnemo greške.

Primer 9: Zabeleška o rezultatima merenja u programu OpenOffice Calc zajedno sa izračunavanjem

Inicijali	Pol	Uzrasr	Visina	Telesna masa	
	[M/Z]	[roky]	[m]	[kg]	
J.M.	M	54	1.56	89	
F.R.	Z	34	1.89	98	
H.K.	Z	30	1.67	60	
E.H.	M	45	1,78	88	
F.W.	M	48	1.76	72	
R.M.	Y	3	1.82	105	
Z.Č.	Z	99	1.66	73	
L.B.	M	36	1.73	82	
Broj jedinica	7	8	7	8	"=COUNTIF(B3
Minimum		3	1.56	60	=MIN(C3:C10)
Maksimum		99	1.89	105	=MAX(E3:E10)

U R-u ćemo dobiti isti rezultat koristeći komandu *summary(),* (Primer 10).

Primer 10: Zadavanje, kontrola podataka i uklanjanje grešaka u programu R

```
> uzrast <-c(54,34,30,45,48,3,99,36) #zadavanje podataka, svi su celi brojevi, odvojeni zarezom
> uzrast # ispis sadržaja promenljive
[1] 54 34 30 45 48  3 99 36
> visina <-c(1.56,1.89,1.67,1,78,1.76,1.82,1.66,1.73) # zadali smo vrednosti u promenljivu visina, ali smo pogrešili kod jedne vrednosti, gde smo umesto da decimalni broj napišemo sa tačkom koristili zarez, što je program interpretirao kao posebnu vrednost, pogledaj zabelešku
> visina
[1]  1.56  1.89  1.67  1.00 78.00  1.76  1.82  1.66  1.73
> visina <-c(1.56,1.89,1.67,1.78,1.76,1.82,1.66,1.73)# ispravljena vrednost
> visina
[1] 1.56 1.89 1.67 1.78 1.76 1.82 1.66 1.73
> masa <- c(89,98,6O,88,72,105,73,82)#  umesto znaka 0 bilo je  zadano O. Program to odmah javlja kao neočekivan simbol)
Error: unexpected symbol in "masa <- c(89,98,6O"
> masa  <- c(89,98,60,88,72,105,73,82)# posle ispravke to je u redu
> masa
[1]  89  98  60  88  72 105  73  82
# prekontrolišemo minimalne i maksimalne vrednosti i broj promenljivih
> summary(uzrast)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  3.00   33.00   40.50   43.62   49.50   99.00
> summary(visina)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.560   1.667   1.745   1.734   1.790   1.890
> summary(masa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 60.00   72.75   85.00   83.38   91.25  105.00
> length(uzrast)
[1] 8
> length(visina)
[1] 8
> length(masa)
[1] 8
```

Kodiranje zadatih podatka

Mnoge odgovore je potrebno pre kodirati unošenja u računar. Ovaj postupak je neophodno sprovesti sa velikom preciznošću. Ispitanici odgovaraju na pitanja slobodnije, a onda se njihov odgovor transformiše u brojčani kod prema unapred pripremljenom ključu. Kodiranje bi trebalo da rade radnici naročito uvežbani za ovu namenu. Mnogi upitnici imaju označeno mesto za kod, tako da ispitanik zna da tamo ne treba da piše. Primer kodiranja je označavanje naseljenog mesta poštanskim brojem. Ne možemo da očekujemo da će svi ispitanici tačno odgovoriti poštanskim brojem svog naseljenog mesta, pa se zato naseljena mesta kodiraju pri obradi popunjenih upitnika. Drugi primer je korišćenje međunarodne klasifikacije bolesti MKB (ICD). Bitno je da kodiranje vrši stručno lice, koje zna da proceni i veća odstupanja od svakodnevno korišćene terminologije. Ovakvih sistema je više, a odlučiti se za jedan od njih iziskuje poznavanje suštine procesa i cilja koji želimo da postignemo. Dok se prikupljaju podaci za pet bolesti, dovoljno je da se kodiraju brojevima od jedan do pet. Ako se uzimaju u obzir sve bolesti disajnog sistema, onda je pogodno da se iskoristi MKB. Ako ova ne odgovara kao prilično gruba i netačna, onda se nudi klasifikacija SNOMED ili drugi kodirajući sistem. Vrednosti se mogu kodirati na različite načine. Na primer, u CALC-u je moguće koristiti komandu IF (kriterijum; tačno; netačno). Ova komanda najpre vrednuje kriterijum, a onda stavi u ćeliju broj ili tekst ili drugu vrednost prema tačnosti, odnosno netačnosti rezultata testiranja kriterijuma. Za kodiranje je moguće koristiti i komandu FIND and REPLACE iz menija EDIT. Većina statističkih programa nudi više mogućnosti kodiranja ili transformacije promenljivih.

Osnove bezbednosti i čuvanja podataka

Ne misli se na bezbednost podataka u smislu straha zbog njihove zloupotrebe. Nećemo ovde ni razmatrati etička pitanja i čuvanje ličnih podataka. Misli se na jednostavne principe koji sprečavaju gubitak podataka. Najgore što može da se desi kad utrošite nekoliko dana i noći na

prikupljanje podataka i kad ih prebacite u kompjuter i, odjednom, kad ste na kraju, kompjuter se pokvari ili podatke izbrišete greškom. Zato želimo da naglasimo potrebu pravljenja kopija. Kad su autori ove publikacije radili studiju za SZO, onda su u ugovoru imali obavezu da od svih zadatih podataka naprave tri kopije, jednu da imaju u kompjuteru i to ne onom koji se koristi za rutinski rad, drugi da imaju u mediju za čuvanje podataka, koji se nalazi mimo kompjutera, a treću kopiju na mediju van zgrade radnog mesta (za slučaj požara ili elementarne nepogode). Iako studentski rad obično ne prelazi nekoliko hiljada redova, dobro je naviknuti se na pravljenje „bekapa“ (back-up). Dokazana je korist bekapovanja na eksternom disku, a takođe i redovno kopiranje celih direktorijuma. Bekapovati možemo takođe i na internetu. Postoji više komercijalnih i nekomercijalnih davalaca prostora za bekapovanje podataka[18]. Za bekapovanje je moguće koristiti i CD ROM, DVD, USB disk. To je samo pitanje sistematičnog pristupa radu i uštede puno vremena i nepotrebnog stresa.

Zaključak

U ovom poglavlju smo naveli principe sistematičnog pristupa obradi podataka. Ne bismo želeli da stvorimo utisak da je upitnik jedini i najbolji instrument za dobijanje podataka. Onda smo predstavili primer instrumenata za kontrolu kvaliteta podataka, a na kraju smo pažnju posvetili bezbednosti podataka. Principi na koje smo skrenuli pažnju koristiće se u bilo kojoj obradi podataka. U sledećem poglavlju pažnju ćemo posvetiti prvobitnom predstavljanju prikupljenih podataka.

[18]Ponuda stalno raste. Na primer, idrive http://www.idrive.com/ pruža 2 GB besplatno, a za više potrebno je platiti u suštini simboličnu cenu.

Vežbe

1. Pripremite upitnik o zadovljstvu studenata nastavom predmeta statistika i iskombinujte otvorena i zatvorena pitanja.

2. Pripremite fiktivne odgovore dobijene od 15 ispitanika i prikažite ih u obliku tabele, saznajte proračune i minimalne i maksimalne vrednosti.

TREĆE POGLAVLJE

Pretstavljanje i prvobitna obrada podataka

Sadržaj poglavlja

Cilj poglavlja

Posle uspešno završenog prikupljanja podataka, istraživač može da počne da se bavi pitanjem šta je u stvari uspeo da sazna. U ovom koraku se pažnja usmerava na obradu podataka, pitanja koja se tiču osnovnih karakteristika uzorka, način predstavljana podataka, zaključivanje o ulozi pojedinih promenljivih i slično. Isto kao i u prethodnim poglavljima, primeri su u statističkom programu R.

Tabela 1: Ciljevi poglavlja

●Informisati se o prvim koracima posle prikupljanja podataka ●Predstaviti pravila prikazivanja podataka u formi tabela i grafikona ●Navesti najčešće greške kod predstavljanja podataka

Kakve podatke sam, u stvari, dobio?

Ovo je prvo pitanje koje istraživač sebi postavi kad završi prikupljanje podataka. Šta treba da uradi kako bi se orijentisao u tom pravcu? Podelićemo taj proces na nekoliko koraka, od kojih će nam svaki od njih približiti različite karakteristike naših podataka. I istraživač po pravilu napreduje na isti način, a često mora i da se vrati na početak.

U prvom koraku istraživač bi trebalo da proveri šta je u stvari prikupio. Počnimo sa tim najjednostavnijim, koliko podataka je prikupljeno i koliki deo tih podatak je upotrebljiv. Počnimo pitanjem: koliko podataka imam. U programu EXCEL postoji koamnda *COUNT(od:do*[19]*)*, koja daje broj elemenata u odgovarajućoj koloni, redu ili drugačije definisanom izboru (više kolona ili redova). To je isto moguće i u programu R, a postupak je ilustrovan Primerom 1.

Primer 1: Zadavanje podataka

Zadavanje: Istraživač je ispitivao vezu holesterola u krvi sa sistolnim i dijastolnim krvnim pritiskom kod 11 muškaraca i žena. Izmerene vrednosti je poslagao u tabelu. Hteo je da sazna koliko ima u svakoj grupi merenja.

Postupak: u prvom koraku napravio je matricu (tabelu) izmerenih vrednosti, tako da je najpre stavio izmerene vrednosti u pojedine promenljive, a onda je od njih napravio matricu *pr1.3*.

```
uzrast <- c(23,45,36,29,56,54,33,27,23,47,45)
hol <- c(6.8,4.5,6.4,4.7,5.1,6.2,19,6.2,16.9,4.9,7.3)
skp <-c(135,145,157,145,135,155,146,176,135,129,144)
```

[19] Od „jeste početna adresa“, do „jeste konačna adresa“ ćelije spiska

dkp <-c(75,80,80,75,90,90,75,80,80,75,90)

mas <- c(66,88,72,67,82,102,65,88,55,70,63)

Komanda *c()*, iz osnovne datoteke R, napraviće vektor vrednosti, a pridruživanjem pomoću <- *ili* = napraviće je prema nazivu promenljive. Komanda *cbind()* spojiće pojedine vektore u matricu po kolonama, a ako želimo da spaja po redovima koristićemo analognu komandu *rbind()*.

pr1.3 = cbind(uzrast, hol, skp, dkp, mas)

Slika 1: Relultat korišćenja komande *summary()* za osnovno opisivanje prikupljenih podataka

Komanda *summary()* daće pregled osnovnih karakteristika pojedinih promenljivih. To su najmanja vrednost (Min.), prvi kvartil (1st Qu.), medijana (Median), aritmetička sredina (Mean), treći kvartil (3rd Qu.), najveća vrednost (Max.).

```
> summary(pr1.3)
     uzrast          hol              skp             dkp              mas       
 Min.   :23   Min.   : 4.50   Min.   :129.0   Min.   :75.00   Min.   : 55.00  
 1st Qu.:28   1st Qu.: 5.00   1st Qu.:135.0   1st Qu.:75.00   1st Qu.: 65.50  
 Median :36   Median : 6.20   Median :145.0   Median :80.00   Median :70.00  
 Mean   :38   Mean   : 8.00   Mean   :145.6   Mean   :80.91   Mean   : 74.36  
 3rd Qu.:46   3rd Qu.: 7.05   3rd Qu.:150.5   3rd Qu.:85.00   3rd Qu.: 85.00  
 Max.   :56   Max.   :19.00   Max.   :176.0   Max.   :90.00   Max.   :102.00
```

Šta je istraživač saznao? Saznao je da najmlađi ispitanik ima 23 godine, a najstariji 56 godina. Minimalne i maksimalne vednosti će mu nagovestiti da li je negde pogrešio, odnosno da li u uzorku ima nekoga starijeg od 100 godina ili mlađeg od 20 godina (s obzirom da je radio sa odraslim ljudima).

Nedostajući podaci

Može da se desi, a dešava se često, da i pored najveće volje ili nedostaju podaci ili su neki podaci pogrešni. Kako to može da se desi? Situacije kad se razbije epruveta sa krvlju ili mokraćom, pojedinac istupi iz istraživanja i ne možemo ga pronaći, pacijent odbije da ide na neko istraživanje, često su realnost u istraživanju. U svim navedenim slučajevima nije mogla da se dobije vrednost jedne ili nekoliko promenljivih, iako je većina vrednosti bila ustanovljena.Takođe može da se desi da se zbog greške u zadavanju unela vrednost koja je jasno izvan obima, na primer uzrast umesto 49 godina zadat je kao 149, ili je krvni pritisak umesto 118 mmHg bio zadat kao 1118 mmHg. U ovakvim slučajevima možemo da logično izvedemo tačnu vrednost, ali šta u slučajevima kad to nije moguće? Tada označimo nedostajući podatak, ili na engleskom „missing value". Ovakvo merenje u

R označimo pridruživanjem specijalne vrednosti NA[20].

Primer 2: Istraživač prilikom merenja krvnog pritiska nije saznao jednu vrednost u četvrtom merenju i naveo ju je kao nedostajuću

U programu R zadaćemo nedostajuću vrednost slovima NA.

> skp <-c(135,145,157,NA,135,155,146,176,135,129,144)

Ispis vrednosti promenljive skp biće

```
< skp
 [1] 135 145 157  NA 135 155 146 176 135 129 144
```

Komanda sumary() onda daje sledeću tabelu, koja karakteriše zadate podatke

```
> summary(stk)
 Min. 1st Qu.  Median Mean   3rd Qu.   Max.   NA's
129.0            135.0   144.5   145.7  152.8    176.0  1.0
```

Tabela ili grafikon?

Najjednostavnije je prikupljene podatke svrstati u tabelu. Da li je to i najbolji način prikazivanja? Počnimo time što ćemo da objasnimo čemu takva tabela služi i kako se pravi. Intuitivno svi ih koristimo prilikom prikupljanja podataka, kad svakom pojedincu priključimo dobijene, odnosno izmerene vrednosti. Ili se to radi u obrascu u formi protokola, ili se koristi „spreadsheet“, npr. EXCELL ili drugi program. U svakom slučaju, cilj ove tabele je da zabeležimo vrednosti tokom istraživanja. Primer ovakve tabele, konstruisane u EXCELL, je Primer 3. Programsko okruženje R nam nudi više instrumenata za izradu tabela, ali to je komplikovanije nego pri korišćenju tekstualnog editora ili *spreadsheet*-a.

Nezavisno od namene, tabele je potrebno je označiti nazivom. Prilikom beleženja rezultata eksperimenta ili merenja takođe je potrebno označiti tabelu pogodnim nazivom, koji može da se pokaže kao koristan, naročito kasnije kada bude potrebno da se vratimo protokolu i podsetimo se šta ta tabela u stvari prikazuje. Tamo gde se radi sa većim brojem promenljivih, često je potrebno koristiti i podnaslove. Budući da ovaj tip tabele služi ujedno i kao protokol, dobro je da se uvede i datum beleženja i ime osobe koja je unela podatke. Kolone moraju da imaju naziv promenljive i jedinice mere. Kod svakog merenja trebalo bi da postoji prostor za komentar. Broj merenja odrediće i broj redova.

Kako se opredeliti kad hoćemo da podatke koje smo prikupili prezentujemo kolegama ili stručnoj javnosti? U prvom redu moramo da shvatimo da su tabela i grafikon različiti načini predstavljanja iste informacije. Tabela može da obuvata više detalja, same podatke podeljene u grupe. Grafikon daje širi pogled na situaciju, a time omogućava i uočavanje trendova i povezanosti. Dakle, kad god je moguće, za predstavljanje detalja prikupljenih podataka bolje je da koristimo tabelu, iako grafikon pruža još i mogućnost povezivanja pojedinih tačaka na vrednosnim krivama. Tabela je preglednija, ali je često teže saznati uzajamne veze između promenljivih ili ustanoviti

[20] NA znači “not available”

trendove. Dobra tabela treba da bude pregledna, sa jednostavnim i jasno razumljivim sadržajem svakog polja. Nije potrebno iscrtavati sve linije, bilo vertikalne ili horizontalne u tabeli; previše linija će smanjiti preglednost tabele. Ispravno odabrane linije pomoći će da se istaknu bitni delovi tabele. Ponekad je dobro istaknuti značajan rezultat ili ceo red podvući koristeći liniju ili masnim slovima (boldom), ili bojenjem ćelije ili linije. Uvek je bitno imati na umu čitaoca i olakšati njegovu orijentaciju. Zato se akcenat stavlja na naziv tabele, koji mora nedvosmisleno i u potpunosti objasniti njen sadržaj. I u slučaju da čitalac nije video tabelu, u kontekstu sa tekstom u članku ili pradavanjem mora da bude jasno iz samog naziva tabele o čemu nam ona govori.

Primer 3: Tabela rezultata sa nazivom

Table 7. Estimates of Odds Ratios for scores of adherence to recommendations from TBI guidelines and favourable outcome, adjusted for age, Injury Severity Score (ISS) and Glasgow Coma Scale (GCS)

Recommendation	Coefficient Estimate	Standard error	OR	OR_low (95%)	OR_up (95%)
Score #2 Prehospital resuscitation	-0.02	0.16	0.98	0.72	1.33
Score #3 Resuscitation of PB and O_2	0.17	0.06	1.18**	1.04	1.34
Score #4 Indication for ICP	0	0.04	1	0.92	1.09
Score #5 ICP treatment threshold	0.02	0.07	1.02	0.88	1.17
Score #6 Type of monitoring	-0.08	0.17	0.92	0.67	1.28
Score #7 CPP	0.02	0.05	1.02	0.93	1.11
Score #8 Hyperventilation Standard	0.02	0.04	1.02	0.94	1.1
Score #8 Hyperventilation Guideline	0.03	0.05	1.03	0.95	1.13
Score #9 Mannitol	-0.03	0.04	0.97	0.89	1.05
Score #10 Barbiturate use	-0.08	0.06	0.93	0.83	1.04
Score #11 Steroid use	0.04	0.04	1.04	0.96	1.12
Score #13 Prophylactic use of anti-seizure drugs	0.02	0.04	1.02	0.95	1.1
Total Scores	0.01	0.01	1.02	0.99	1.04

** p<0,01

Primer 4: Tabela prikupljenih podataka. Obratite pažnju da je svaki ispitanik označen na dva načina, prvi ID je uzet sa mesta uzorkovanja, a drugi ID (INT PID) je interna oznaka u bazi podataka. Ovaj način obeležavanja će omoguti orijentaciju u ličnoj bazi podataka, ali i sačuvati mogućnost povezivanja sa izvorom podataka.

ID	INT PID	Count ICP	AGE	GENDER: Male 1 Female 0	ISS	TRISS	GCS	PUP_ASY Yes=1;No =0	PUP_BIL Yes=1;No= 0	Basal Cisterns Opened 1 Closed 0
10121	1	0	12	1	12	96	8	0	0	0
10122	2	0	46	1	21	73	3	1	0	0
10123	3	1	33	1	17	0	0	0	0	0
10124	4	0	68	1	17	35	3	0	0	0
10125	5	1	1	1	45	25	3	0	0	0
10126	6	0	41	1	0	0	0	0	0	0
10141	1	1	72	1	33	71	12	0	0	1
10142	2	1	17	1	18	90	4	0	0	1
10143	3	1	29	0	21	98	12	0	0	1
10144	4	0	89	0	16	90	11	0	0	1
10145	5	0	16	1	29	90	7	0	0	1

Često se zaboravlja na izražavanje decimala. Danas, kada kompjuteri izračunavaju rezultate na mnoštvo decimala, veoma je bitno shvatiti koliko brojeva iza decimalnog zareza (ili tačke) ima smisla. Uopšteno važi da što manje cifara iza decimalnog zareza, to bolje. Pogrešnom tačnošću se zove postupak kad se brojčani podaci predstavljaju sa većom tačnošću od realne. Budući da je tačnost[21] prebačena u preciznost, reproduktivnost ili ponavljanost, ovakav postupak onda vodi ka neobrazloženom višku sigurnosti. U nauci se zato koristi dogovor koji kaže da ako nije unapred određena granica grešaka, broj cifara korišćenih prilikom prezentacije podataka mora da bude ograničen tačnošću samih podataka. Ako, dakle, merimo uzrast brojem proživljenih godina, nije ispravno prikazivati prosečan uzrast sa jednim ili više decimalnih mesta. Ako za merenje određenog parametra koristimo uređaj ili metodu koja radi sa tačnošću na dva decimalna mesta, ne možemo rezultate statističke obrade prezentovati na tri li više decimalnih mesta. Ako izračunavamo procente, trebali bismo ih zaokruživati na cele brojeve, zato što sam procenat predstavlja stotinu, dok kad koristimo decimalno mesto prikazaćemo hiljade, znači promile.

Grafičko prikazivanje

O prikupljenim podacima puno govori njihov grafički prikaz. Jednostavan dijagram izmerenih vrednosti puno će reći o njihovom karakteru. Komanda *plot(stk, main="promenljiva:STK", xlab="merenje", ylab="vrednost")* će nacrtati dijagram razlaganja podataka (Primer 5).

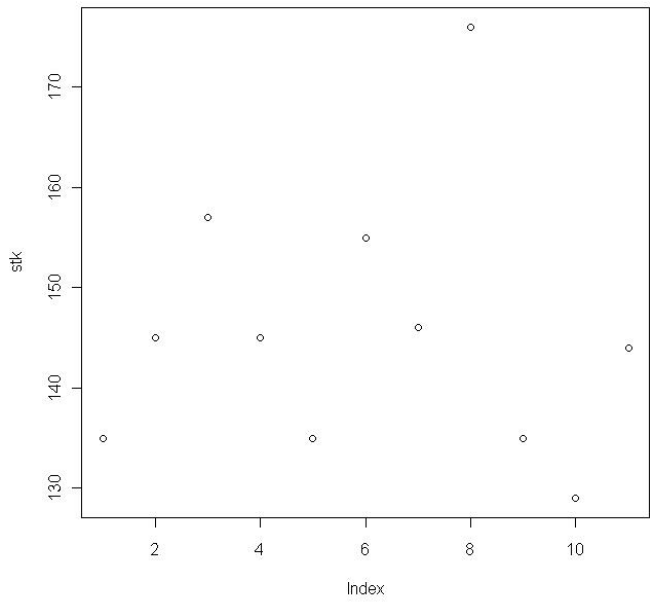
U komandu *plot* smo, osim nazivu promenljive *stk*, poslali tekst, kojim smo dali naziv grafikonu (*main="promenljiva:STK")*, takođe smo odredili nazive za osu x i y (*xlab="merenje"*,

[21] **Tačnost** u nauci, inžinjerstvu i u statistici predstavlja stepen odgovornosti merenja ili izračunate vrednosti prema njenoj stvarnoj (istinitoj) vrednosti. **Preciznost** (takođe nazvana reproducibilnost ili ponavljivost) izražava stepen u kojem sledeća merenja ili izračunavanja daju iste ili slične rezultate.

ylab="vrednost").

Dobra osobina R je činjenica, da će se rezultujuća slika preneti u drugi dokumenat time, da na pritku slike nacrtane programom, posle pritiska na rečenicu *File-Copy to the clipboard-as metafile* se u memoriju stavi određena slika, a pomoću naredbe *zalepi,* odnosno *paste* staviće se u dokument sa tekstom (npr. u ovaj dokument). Moguće je to uraditi i tako da se rečenicom *File-Save as-JPG* upamti slika u uzorku .jpg, a taj se onda stavi u tekst.

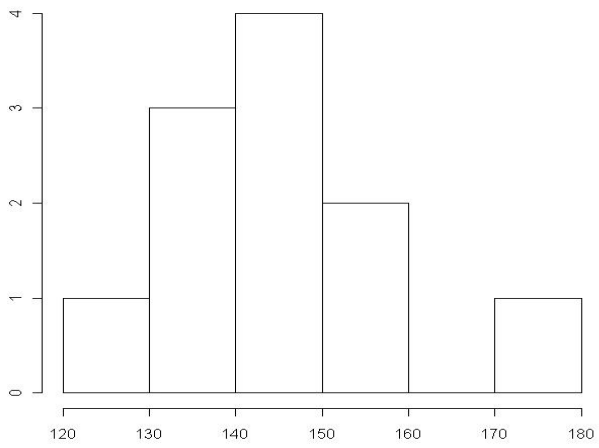
Primer 5: Dijagram razgranatih podataka, takođe označavan kao tačkasti, u engleskom kao scattergram ili dot-plot chart



Sledeća slika dokumentuje rezultat prenosa slike iz R u editor za tekst.

Primer 6: Histogram frekvencija, takođe nazivan frekventni poligon izmerenih vrednosti svrstanih u grupe sa množenjem 10, iscrtaće se pozivanjem komande hist(stk)

promenljiva: STK



izmerena veličina

Istraživač iz histograma ujedno može da isčita koliko vrednosti ima u intervalima po deset, npr. u intervalu 130 do 140 ima 3 vrednosti.

Stablo i list (stem and leaf) je način prikazivanja frekvencije, izražen bez korišćenja grafike.

Primer 7: Stem and leaf prikazivanja frekvencije podataka

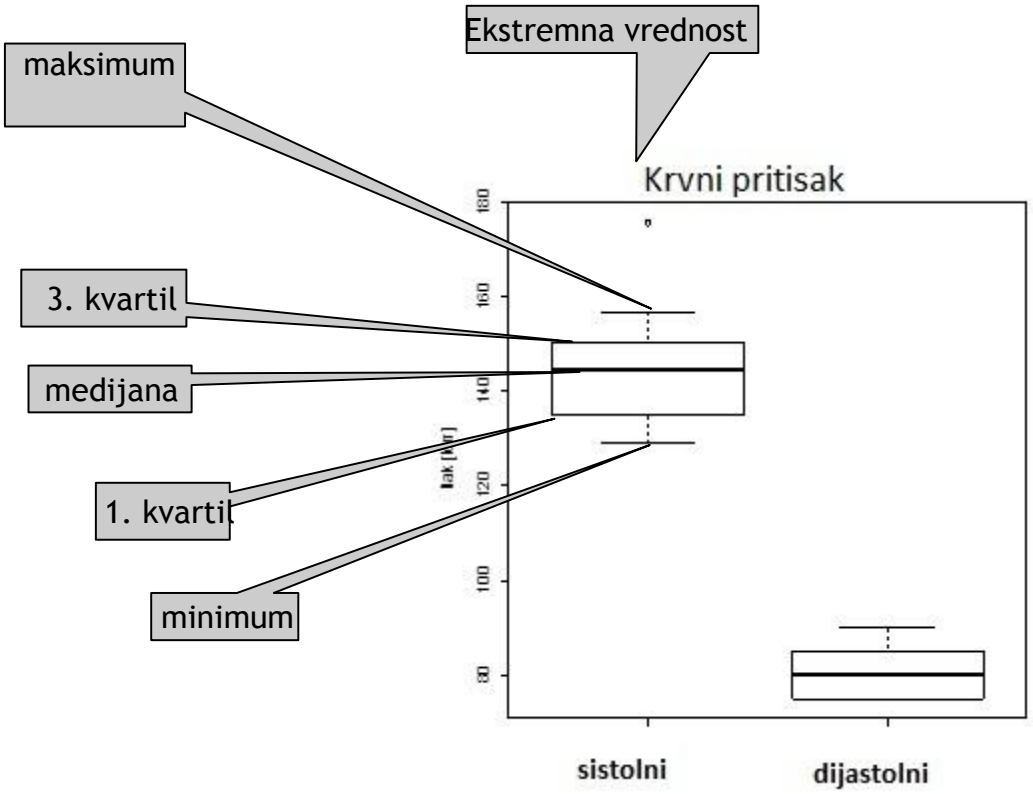
```
> stk
 [1] 135 145 157 145 135 155 146 176 135 129
144

> stem(stk, scale=1)
  The decimal point is 1 digit(s) to the right of
the |
  12 | 9
  13 | 555
  14 | 4556
  15 | 57
  16 |
  17 | 6
```

Program razdeli sva posmatranja na vrednosti „stabla“ i „listova“ tako da vrednosti „stabla“ označavaju razred (Class) a vrednosti „listova“ broj (Count). Svaki red se počinje razredom, u našem slučaju je reč o stotinama i deseticama, npr. broj 12 izražava 120, a 13 izražava 130. Iza

vertikalne linije se nalaze „listovi“, koje govore kakvi se podaci pojavljuju u određenom razredu, u našem primeru se iza broja 12 nalazi broj 9, a time se prikazuje činjenica da se broj 129 pojavljuje jednom. U slučaju ponovnog pojavljivanja vrednosti, broj listova će se povećati, npr. u razredu 130 je broj listova sa vrednošću 5 tri puta, dakle to se prikaže trima peticama. U razredu 150 je jedna vrenost 155 i jedna 157 i zato iza linije piše 57. Prednost ovog prikazivanja raspodele vrednosti je i u tome da veoma dobro prikazuje ekstremne vrednosti.

Za zapisivanje podataka i njihovo kompleksno prikazivanje jako je koristan „box plot“, što bismo mogli da prevedemo kao „prikaz u kutiji“ ili kvartilni grafikon. Držaćemo se ipak izvornog engleskog naziva. Boksplot grafički prikazuje glavne zabeležene mere centralne tendencije, kao što su aritmetička sredina i medijana, ali isto tako i mere disperzije: raspon, prvi i treći kvartil. Boksplot indikuje i oblik rastavljanja vrednosti.

Primer 8: Boskplot grafikon sa podacima o sistolnom i distolnom krvnom pritisku (njihovo zadavanje pogledaj u primeru 1)



Ovaj grafikon (Primer 8) smo dobili komandom:

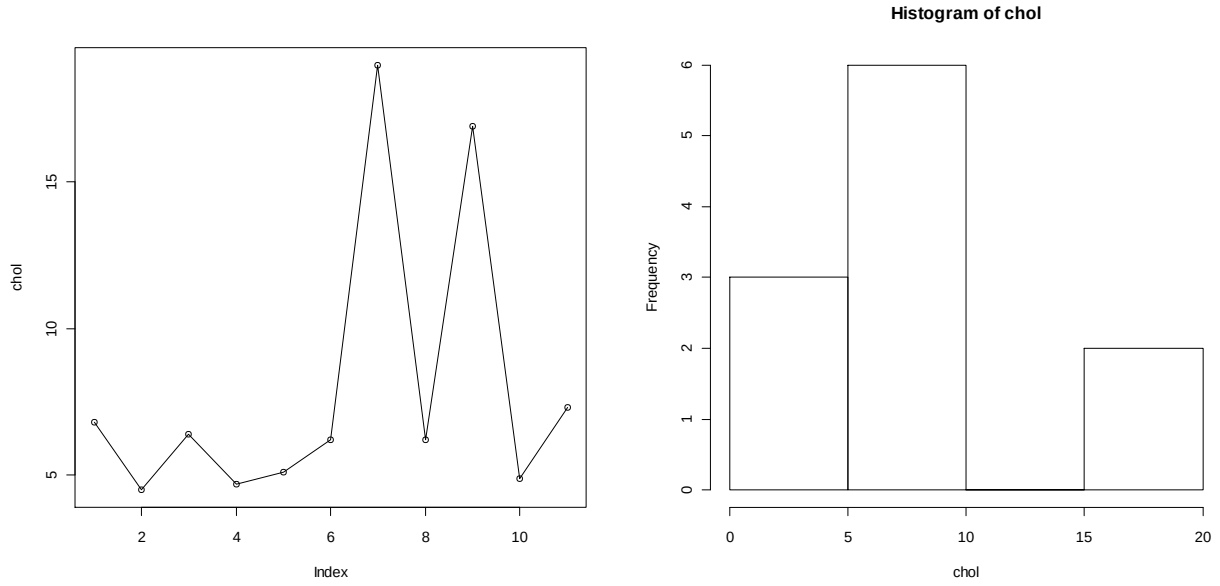
> boxplot(stk, dtk, main = "Krvni pritisak", ylab = "pritisak [torr]", names = c("Sistolni", "dijastolni"))

Uočite da smo dodali naziv grafikona *main = "Krvni pritisak"*, a takođe smo dodali i naziv ose y *ylab = "pritisak [torr]"*, a na kraju smo nazvali pojedine promenljive *names = c("Sistolni", "Dijastolni").*

Podelom raspona podataka (minimum - maximum) na četiri jednaka dela dobijamo takozvane kvartile. Ove ćemo objašnjavati u sledećim poglavljima.

Navešćemo i najčešće korišćene grafikone, od kojih je svaki odgovarajući za prikazivanje određenih podataka. **Stubičasti** prikazuje podatke dobijene ili važeće u istom periodu (npr. upoređivanje broja kreveta među bolnicama, broj pacijenata sa razlilitim komplikacijama povreda, broj studenata prema popušenim cigaretama i sl.). Ovakav grafikon koristimo onda kad ne pretpostavljamo da postoji veza među pojedinim kategorijama podataka. Kad hoćemo da prikažemo broj ljudi prema starosnim kategorijama ili prema broju popušenih cigareta ispravno je koristiti stubičasti grafikon. Kad bi koristili linijski grafikon, mogli bi da steknemo utisak da između kolona sa brojem ljudi u starosnoj grupi 21-30 godina i starosti 31 – 40 godina postoji između i neka prelazna grupa, čiju vrednost znamo na osnovu linije koja povezuje obe vrednosti koje prikazuje.

Slika 1: Linijski i stubičasti grafikon promenljive hol



Linijski grafikon predstavlja podatke zabeležene u vremenu, pokazujući nekakvo vremensko kretanje ili trend (npr. kretanje incidencije žutice tokom dve nedelje, kretanje broja povreda u okrugu za godinu prema mesecima i slično). Ovakav grafikon onda omogućava pretpostavku da su između dve izmerene tačke prisutne vrednosti koje se mogu pročitati iz grafikona (iako to ne mora da bude u potpunosti tačno). Na osnovu grafikona možemo jednostavno da kažemo koji pokazatelj i kada je dobio najveću ili najmanju vrednost. Takođe možemo da zaključimo o prisustvu ili odsustvu određenog trenda.

Grafikon u oblik kolača predstavlja podatke koji zajedno čine celinu od 100%. Svaka vrednost je prikazana kao isečak iz kruga (npr. pojava pojedinih uzroka smrtnosti u odnosu na sve uzroke, podela uzorka prema broju popušenih cigareta i sl.). Omogućava posmatranje proporcije delova iz celine podataka. Njime je moguće prikazati kvantitativne i kvalitativne podatke, od kojih su poslednji veoma pogodni za ovu vrsta grafikona. U celini, pak, moramo da upozorimo da iako je grafički prikaz u obliku kolača zgodan za upoređivanje, naše oko nije naviklo da upoređuje površine. Bolje je koristiti stubičasti grafikon za upoređivanje vrednosti.

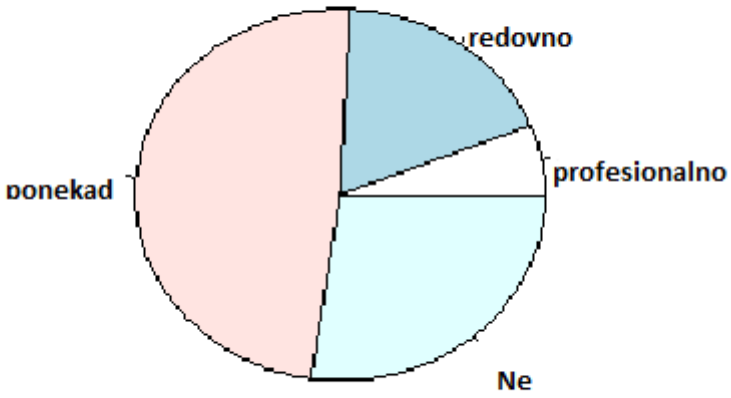
Primer 9: Grafikon u obliku kolača

Student je hteo da sazna koliko se njegove kolege bave sportom, a odgovore je svrstao u četiri kategorije, kojima je priključio frekvencije odgovora. Dobio je sledeću informaciju: četiri studenta su se bavila sportom profesionalno, 12 redovno, 32 ponekad, a 18 se nije uopšte bavilo sportom. Rezultate je prikazao grafikonom u obliku kolača.

Rešenje:

> sport<- c(4, 12, 32, 18)# zadavanje ulaznih podataka

> pie(sport, labels = c("Profesionalno", "Redovno", "Povremeno", "Ne")) #pozivanje komande za oznake prikazane pod nazivom merenja



Uopšteno je potrebno kod pravljenja grafikona i tabela pridržavati se principa jednostavnosti i razumljivosti. Rad neće dobiti na kvalitetu ako grafikon ili tabela budu vrištali od boja. Jednostavnost donosi sa sobom i potrebu dubljeg razmišljanja nad ciljem korišćenog grafikona. Isprazno iskorišćavanje mogućnosti kompjuterskog generisanja grafikona često prikriva dobru nameru autora. Razumljivost opet podrazumeva da će svaka stavka u grafikonu biti nedvosmisleno opisana, dakle na osama će biti prikazani ne samo naziv, već i jedinice mere promenljve, pojedine tačke ili kolone će se jasno razlikovati, u nazivu grafikona biće jasno napisano šta taj grafikon prikazuje, koje podatke, kako su bili obrađivani.

Sažetak

Tačno prikazivanje podataka je jako bitno zbog narednih statističkih saznanja. Grafičko prikazivanje podataka i ispravno korišćenje prikaza naročito može da dovede do odlučivanja u narednom postupku na izražajan način, kao i do orijentisanja u smislu potvrde hipoteza, a takođe može da prikaže i rezultate naučnoj javnosti. U vežbama ćemo predstaviti nekoliko studija sa realnim podacima. Pokušaćemo da podatke ubacimo u promenljive u R i da ih sačuvamo za dalje korišćenje u sledećim poglavljima.

Vežbe

1. Istraživač[22] je ispitivao uticaj više fizioloških faktora na stanje pacijenata sa cističnom fibrozom. Podaci koje je prikupio nalaze se u uzorku *cystfibr* iz datoteke *IswR* (otvoriti korišćenjem komande *library(ISwR)*). Odredite veličinu pojedinih promenljivih, minimum i maksimum i nacrtajte sliku sa grafičkim prikazom svih podataka; nacrtajte pojedine slike za svaki podatak pojedinačno.
2. J.Klemensen[23] je posmatrao rasprostranjenost karcinoma u četiri grada u Danskoj. Podaci koje je prikupio nalaze se u uzorku *eba1977* u datoteci *ISwR*. Odredite dužinu promenljivih, minimum i maksimum i nacrtajte sliku sa grafičkim prikazom podataka, gde to ima smisla i nacrtajte pojedine slike za svaki podatak samostalno.

[22] O'Neill et al. (1983), The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis, Am. Rev. Respir. Dis.,128:1051–1054.

[23] J. Clemmensen et al. (1974), Ugeskrift for Læger, pp. 2260–2268.

ČETVRTO POGLAVLJE

Mere centralne tendencije i disperzije

Sadržaj poglavlja

Cilj poglavlja

Kad smo uspeli da prikupimo podatke i stekli osnovnu predstavu o njihovim celokupnim karakteristikama, moramo da probamo da nađeme takve brojčane karakteristike koje će sakupljene podatke prezentovati sa najvećom mogućom preciznošću i jednoznačnošću. U ovom poglavlju objasnićemo osnovne mere koje se najčešće koriste. Ukazaćemo kako ih razumno i tačno interpretirati. Nasuprot čestom korišćenju ovih mera, mnogo puta intuitivno, njihovo korišćenje i interpretacija često su pogrešni. U ovom poglavlju čitalac će se informisati o osnovnim postupcima kod izračunavanja mera centralne tendencije i disperzije, razjasniće njihove osobine i saznaće kada i kako da ih ispravno koristi i interpretira. Kao i u prethodnom poglavlju, primeri su dati u statističkom programu R.

Tabela 1: Ciljevi poglavlja

3. Razjasniti mere centralne tendencije i disperzije
4. Biti svestan osobina pojedinih mera
5. Upoznati se sa postupkom izračunavanja i interpretiranje rezultata

Mere centralne tendencije

Pod merom podrazumemo broj koji karakteriše dati skup podataka. Pod centralnom tendencijom podrazumevamo usmerenost podataka ka centru, sredini. Samo jedna mera, korišćena samostalno, nije dovoljna da opiše skupa podataka. Neophodno je kombinovati je sa drugim merama. Zato ovo poglavlje u nazivu ima i disperziju, koja je mera opisivanja difuzije skupa podataka.

Definicija 1: Mere centralne tendencije

Mera centralne tendencije uzorka je broj koji karakteriše centralno usmeravanje skupa podataka. Najčešće korišćena mera centralne tendencije odabranog uzorka su aritmetička sredina, medijana i mod.

Aritmetička sredina

Uobičajena mera centralne tendencije je aritmetička sredina. Dobija se sabiranjem svih vrednosti u uzorku i deljenjem tog zbira sa brojem vrednosti. Razlikujemo aritmetičku sredinu populacije i aritmetičku sredinu uzorka. Računanje obe aritmetičke sredine je jednaka. Ako imamo podatke za celu populaciju (izuzetno, na primer kao razultat popisa stanovnika), onda govorimo o aritmetičkoj sredini populacije i označavamo je grčkim slovom μ (mi). Najčešće radimo sa podacima dobijenim od uzorka populacije i u tom slučaju govorimo o aritmetičkoj sredini uzorka, a

označavamo ga sa $\overline{X}$ (x bar). Aritmetička sredina ima osobine koje određuju njenu široku primenu. Praktično sva statistička saznanja urađena na kvantitativnim promenljivim koriste aritmetičku sredinu za dobijanje prve informacije o promenljivoj. Osobine aritmetičke sredine je preporučuju za često korišćenje, mada njena interpretacija nije bez problema.

Glavna prednost aritmetičke sredine je njena jednostavnost, laka razumljivost i lako izračunavanje. Razumljvost dovodi do toga da i deca u osnovnoj školi brzo nauče da koriste aritmetičnu sredinu (prosek) za računanje pretpostavljene zaključne ocene. Ova jednostavnost i prirodna razumljivost dovodi do njene izuzutno česte primene, ali i do pogrešne interpretacije rezultata. Iz škole znamo da je zaključena ocena bila prosek ocena. Uticale su na to, između ostalog, i ekstremne vrednosti (prisetite se očekivanja da će jedna slaba ocena biti ignorisana).

Osetljivost na ekstremne vrednosti utiče na vrednost aritmetičke sredine, tako da može da se desi da će ona u određenim slučajevima postati neupotrebljva kao mera centralne tendencije. Izračunavanje prosečne ocene učenika koji je imao iz određenog predmeta određene vrednosti, kad uzmemo u obzir i izostavljanja ekstremnih, je ilustracija osobina aritmetičke sredine (Primer 1).

Primer 10: Izračunavanje zaključene ocene kao primer proseka za ekstremne vrednosti (U slovačkom sistemu školovanja ocena 1 znači odličan, a ocena 5 nedovoljan.)

Ocene: 1, 3, 2, 1, 1, 1, 5

onda je prosek ocena bio 2; međutim, ako učitelijca ne bi uzimala u obzir jedinu peticu koju je učenik dobio, onda bi situacija bila drugačija

Ocene: 1, 3, 2, 1, 1, 1

sa prosekom 1,5. Ovde bi učiteljica mogla da se opredeli i za jedinicu kao zaključenu ocenu.

Sledeća osobina aritmetičke sredine je njena unikatnost, koja određuje da za dati skup podataka postoji samo jedna jedina aritmetička sredina. Za ilustraciju kako aritmetička sredina izražava centralnu tendenciju, navešćemo primer.

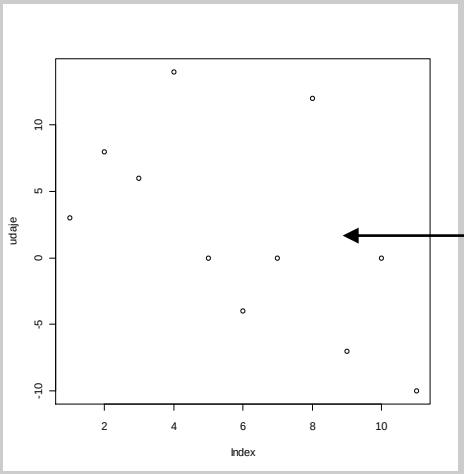
Primer 11: Izračunavanje aritmetičke sredine i odstupanja podataka od nje

Zadajemo: Imamo skup rezultata merenja: 3 8 6 14 0 -4 0 12 -7 0 -10 i želimo da izračunamo njihovu aritmetičku sredinu.

Rešenje:

```
> podaci <- c(3, 8, 6, 14, 0, -4, 0, 12, -7, 0, -10)
> podaci
 [1]  3  8  6 14  0 -4  0 12 -7  0 -10
> aritmetickasredina <- sum(podaci)/length(podaci)
> aritmetickasredina
[1] 2
```

> plot(podaci)



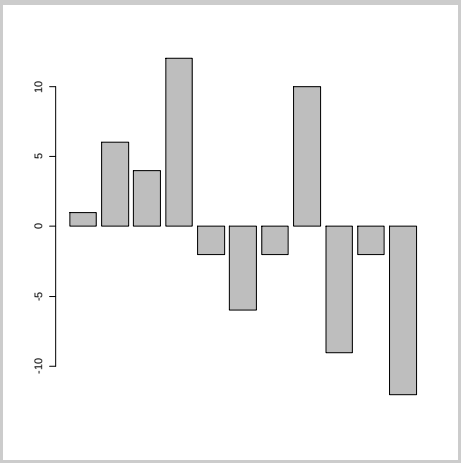
Pogledaćemo kako se pojedini podaci razlikuju od aritmetičke sredine:

> dif <- podaci - aritmetickasredina

> dif

[1] 1 6 4 12 -2 -6 -2 10 -9 -2 -12

> bargraph(dif)



Zbir odstupanja od aritmetičke sredine pozitivnih i negativnih:

> dif

[1] 1 6 4 12 -2 -6 -2 10 -9 -2 -12

> pos <- sum(1, 6, 4, 12, 10)

> pos

[1] 33

> neg <- sum(-2, -6, -2, -9, -2, -12)

> neg

[1] -33

Vidimo, da je zbir odstupanja pozitivnih i negativnih jednak. Dakle, stvarna aritmetička sredina određuje centralnu tendenciju podataka.

Medijana

Medijana konačnog zbira je vrednost koja deli skupove na dva jednaka dela, takva da je broj vrednosti većih ili jednakih medijani jednak broju vrednosti manjih ili jednakih medijani. 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.

Ako bi učiteljica koristila medijanu za vrednovanje, onda bi računanje izgledalo kao u sledećem primeru (Primer 3).

Primer 12: Izračunavanje medijane za grupu brojeva

Ocene: 1, 3, 2, 1, 1, 1, 5

onda je prosek (aritmetička sredina) ocena 2.

Posle poređanih ocena prema veličini: 1,1,1,1,2,3,5

ostaje u sredini broj 1, što je medijana za date brojeve.

Ako bi učiteljica koristila medijanu za izračunavanje krajnjeg vrednovanja, ne bi morala da se bavi ekstremnim vrednostima (na primer peticom).

Osobine medijane predodređuju je za širu upotrebu nego što je to danas slučaj. Ona je unikatna kao i aritmetička sredina, postoji samo jedna medijana za dati skup podataka. Izračunava se jednostavno. Senzitivnost na ekstremne vrednosti je manja nego kod aritmetičke sredine, odnosno one ne utiču na medijanu tako snažno kao na aritmetičku sredinu. Medijana i aritmetička sredina se pri normalnoj distribuciji promenljivih izjednačavaju (o podeli promenljivih u sledećim poglavljima). Izračunavanje medijane u našem statističkom okruženju R nalazi se u sledećem primeru.

Primer 13: Izračunavanje medijane u okruženju R

Koristićemo podatke iz primera o aritmetičkoj sredini (Primer 2).

```
> podaci
 [1]  3  8  6 14  0 -4  0 12 -7  0 -10
> sort (podaci)
 [1] -10 -7 -4  0  0  0  3  6  8 12 14      # samo za ilustraciju postupka
```

> median (podaci)

[1] 0

Rezultat je 0.

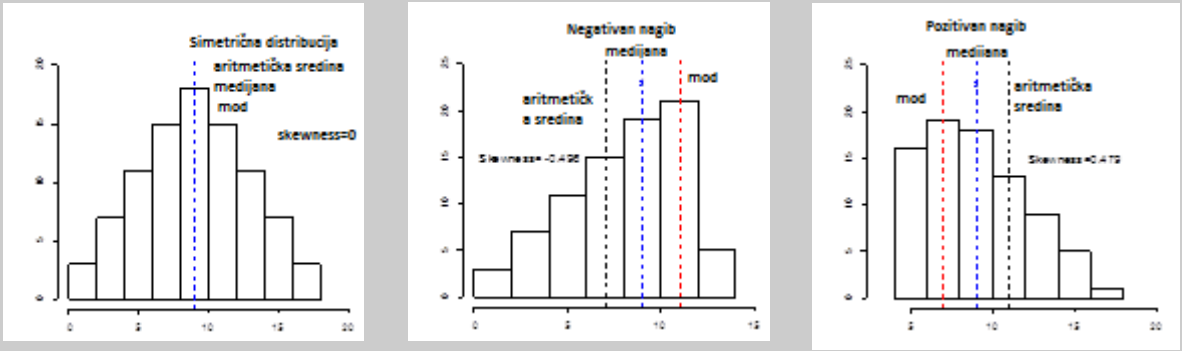
Mod

U datom skupu vrednosti mod je ona vednost koja se najčešće pojavljuje. Možemo da ga koristimo za prikaz kvalitativnih podataka. U primeru sa ocenama mod će ponovo biti jedinica. Najčešće se koristi za vrednovanje skora. Ako su svi brojevi skupa jednaki, onda nema moda. Ako se u skupu pojavljuje samo jedan jedini broj, onda govorimo da je skup unimodalan. Ako se dva broja pojavljuju najčešće, onda govorimo o bimodalnom skupu podataka.

Oblici distribucije podataka oko srednje vrednosti

Za zaključivanje o mogućnostima statističke obrade podataka bitno je poznavanje oblika distribucije podataka. To ne mora da znači da ih moramo prikazati vizuelno. Statistička vrednost, koja karakteriše simetriju podataka, naziva se **koeficijent asimetrije** ili **nagib**. Ako je njegova vrednost bliska nuli, distribucija podatako oko aritmetičke sredine je približno simetrična. Ako dobije vrednosti veće od nule, podaci su više pomereni na levo od aritmetičke sredine, a ako je vrednost manja od nule, onda su pomereni na desnu stranu od aritmetičke sredine. U slučaju vrednosti većih od nule, govorimo o pozitivnom nagibu i u ovom slučaju aritmetička sredina je udesno od medijane, a mod je ulevo od medijane. U slučaju vrednosti manjih od nule, govorimo o negativnom nagibu i postavljanje aritmetičke sredine prema medijani i modu je suprotno. Sledeća slika ilustruje ovakve slučajeve.

Primer 14: Prikazivanje tri tipa distribucije podataka oko aritmetičke sredine



(a) (b) (c)

Za izračunavanje nagiba koristićemo komandu *skewness iz datoteke e1071.*

Za prikazivanje situacije sa slike (a) napravićemo objekat sa nazivom *norm*

> norm=c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,8,8,8,8,9,9,9,10,10,11)

> library(e1071) #učitavamo datoteku 071

> skewness(norm) #Komandom *skewness* zatražićemo izračunavanje nagiba.

Rezultat:

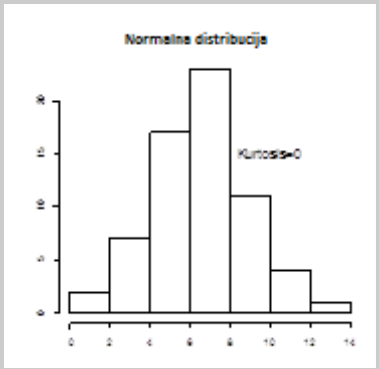
> skewness(norm)

[1] 0

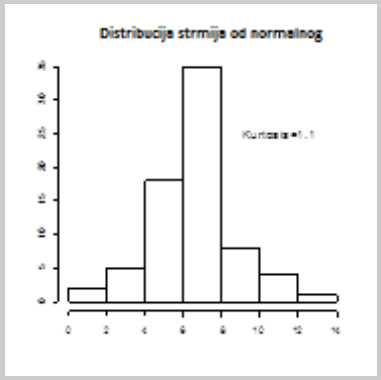
Nagib je jednak nuli, pa zaključujemo da su podaci jednako nagnuti kao pri normalnoj raspodeli i simetrični su u sredini. Uverićemo se u to i izračunavanjem aritmetičke sredine i medijane, koje obe iznose šest. Grafikon (b) prikazuje primer kada je distribucija nagibljenija sa nagibom manjim od nule (-0,5) i podacima pomerenim udesno, a grafikon (c) je suprotan primer, kad je nagib veći od nule, dakle distribucija podataka je pomerena u levo. Na podacima iz primera u trećem poglavlju ilustrovaćemo kako u stvarnosti izgledaju rezultati izračunavanja nagiba.

Zašiljenost, strmost (ili spljoštenost) karakteriše koncenraciju vrednosti obeležja oko srednje vrednosti. Koeficijent spljotenosti informiše da li je koncentracija vrednosti istraživanog obeležja veća ili manja nego u uzorku sa tzv. normiranom normalnom raspodelom. Ako je koeficijent spljoštenosti veći od nule, podela vrednosti obeležja je strmija (šiljatija) nego normirana normalna podela, a ako je manja od nule, raspodela vrednosti obeležja je zaravnjenija, šira nego normirana normalna podela (o normalnoj podeli govoricemo kasnije).

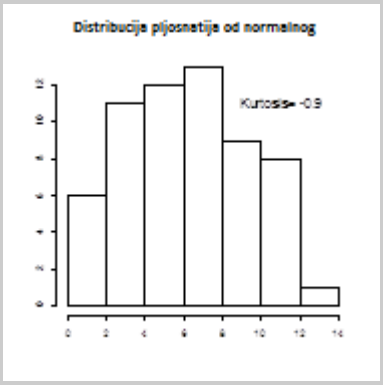
Primer 15: Tri tipa distribucije i izračunavanje spljoštenosti



(a)



(b)



(c)

Za izračunavanje zašiljenosti koristićemo komandu *kurtosis* iz datoteke *e1071*.

Napravimo objekat sa nazivom norm:

>norm=c(1,2,3,3,3,4,4,4,4,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7,7,7,7,7,7,8,

8,8,8,8,8,8,8,9,9,9,9,9,9,9,9,10,10,10,11,11,11,12,13)

> library(e1071) # učitajmo datoteku *e1071*

>kurtosis(norm) # zahtevajmo izračunavanje pljosnatosti.

Rezultat:

> kurtosis(norm)

[1] -0.005227106

Pljosnatost je ispala praktično jednaka nuli, pa zaključujemo da su podaci jednako zašiljeni kao normirana normalna rapodela. Grafikon (b) prikazuje primer kada je podela zašiljenija sa zašiljenošću većom od nule (1.1), a grafikon (c) suprotan slučaj, kad kad je zašiljenost manja od nule, dakle distribucija podataka je šira.

Mere disperzije

Posle određivanja centralne tendencije, potrebno je odrediti koliko blizu, odnosno koliko daleko su rezultati disperzirani oko centra. Ova pojava se može nazvati raipanje, disperzija ili varijabilnost podataka. Disperzija skupa posmatranja opisuje različitost ustanovljenu posmatranjem. Ako bi svi podaci bili isti, onda ne bi bila prisutna nikakva disperzija, a ako su podaci blizu jedni drugih, onda je disperzija mala. Postoji više mera disperzije i svaka ima različite osobine. Objasnićemo raspon populacije, interkvartilni razmak, varijansu, standardnu devijaciju i koeficijent varijacije.

Raspon populacije

Raspon populacije je najjednostavnija mera disperzije podataka. Izračunava se kao razlika između najviše i najniže vrednosti skupa posmatranja. Njegova primena je ograničena, kada u obzir uzimamo dve vrednosti, i zato je značajno zavisna od ekstremnih vrednosti. Prednost raspona populacije je lako izračunavanje.

Primer 16: Izračunavanje raspona populacije

Za izračunavanje koristimo komandu *range*, pomoću koje ćemo saznati maksimalnu i minimalnu vrednost promenljive.

>x=c(10,23,45,34,23,12,1,3,56,43,23,41,21,12,15)

>range(x)

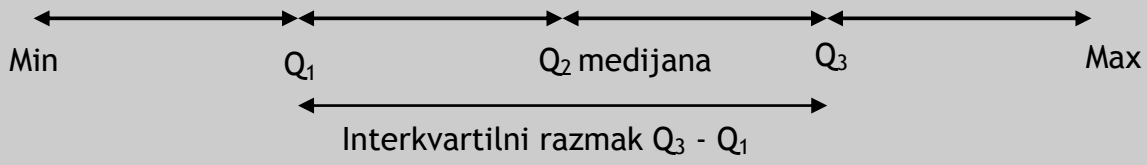
[1] 1 56

Komanda nam je pružila najmanju i najveću vrednost u uzorku. Sam raspon populacije saznaćemo jednostavnim oduzimanjem minimalne od maksimalne vrednosti. Njihova razlika je raspon populacije u uzorku, znači 55.

Interkvartilni razmak

U prošlom poglavlju smo naveli da podelom uzorka na četiri jednaka dela dobijamo kvartile. To su tri vrednosti promenljive, koje nju dele na četiri jednaka dela. Dobijaju se tako da se podaci poređaju prema veličini i podele u četiri iste grupe. Tako svaka od njih sadrži 25% (dakle ¼) svih podataka. Zato se to zove kvartil. Moguće su i finije podele: kada se uzorak podeli na deset delova dobićemo decile ili ako uzorak delimo na sto delova govorimo o percentilima. Jedan od kvartila, drugi po redu, jeste medijana Q_2, koj smo već naveli.

Slika 2: Kvartili i interkvartilni razmak



Interkvartilni razmak ne pada pod uticaj ekstremnih vrednosti i zato pruža bolju i neiskrivljenu sliku o disperziji podataka. S obzirom na njegovo jednostavno izračunavanje i lako prikazivanje (box-plot) mogao bi da bude i češće korišćen. Izračunavanje kvartila omogućavaju komande *quantile() i summary().*

Primer 17: Izračunavanje kvartila i interkvartilnog razmaka. Komande summary() izračunaće i aritmetičku sredinu. Vodite računa o velikim slovima kod pozivanja komande IQR().

Koristićemo promenljivu *chol i*z prethodnih primera.

```
> quantile(chol)
  0%  25%  50%  75% 100%
4.50 5.00 6.20 7.05 19.00
> summary(chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.50    5.00    6.20    8.00    7.05   19.00
> IQR(chol)
[1] 2.05
```

Disperzija – varijansa, srednje apsolutno odstupanje, srednja kvadratna fluktuacija

Tačnija mera rasipanja vrednosti je disperzija, odnosno varijansa. Izračunavanje se zasniva na sabiranju razlika vrednosti pojedinih merenja u odnosu na aritmetičku sredinu. Pre sabiranja su ove vrednosti dignute na kvadrat, da se eliminišu znakovi. Zbir se onda podeli brojem merenja smanjenim za jedan, što se označava kao *n–1*. Ako su vrednosti posmatranog skupa blizu aritmetičke sredine, onda je njihova disperzija mala i obrnuto. Razlika n–1 u izračunavanju se zove broj stepena slobode. Za pojašnjenje navešćemo primer, koristeći podatke o holesterolu iz prethodnog poglavlja. Izračunata varijansa 25,2 je prema aritmetičkoj sredini osam jako velika, što pokazuje da su podaci značajno rasuti od aritmetičke sredine, a to možemo da vidimo i iz nacrtanog tačkastog dijagrama (slika 2).

Primer 18: Izračunavanje vrednosti varijanse

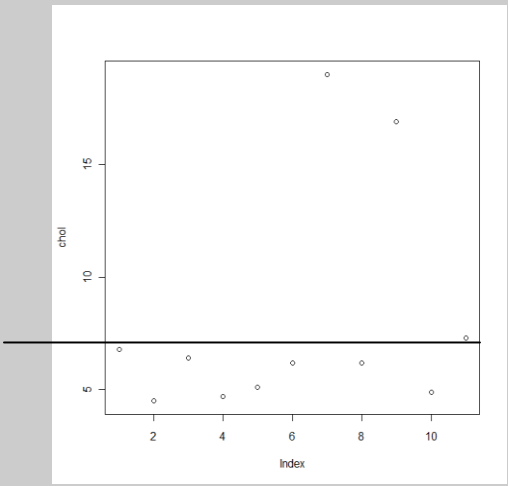
Slučaj	chol	Odstupanja od aritmetičke sredine	Kvadrat
1	6,8	1,2	1,44
2	4,5	3,5	12,25
3	6,4	1,6	2,56
4	4,7	3,3	10,89
5	5,1	2,9	8,41
6	6,2	1,8	3,24
7	19	-11	121
8	6,2	1,8	3,24
9	16,9	-8,9	79,21
10	4,9	3,1	9,61
11	7,3	0,7	0,49
Aritmetička sredina	8		
Kvadrati odstupanja od aritmetičke sredine zajedno			252,34
Broj slučajeva – 1			10
Varijansa			25,23

U okruženju R izračunavanje je jednostavno i sa istim rezultatom.

Koristićemo komandu var.

```
> chol
 [1]  6.8  4.5  6.4  4.7  5.1  6.2 19.0  6.2 16.9  4.9  7.3
> var(chol)
[1] 25.234
```

Slika 3: Tačkasti dijagram predstavljenih podataka, aritmetička sredina je označena punom linijom



Standrardna devijacija

Izračunavanjem korena vrednosti varijanse dobijamo standardnu devijaciju. Time se, u stvari, vraćamo do izvornih vrednosti odstupanja od aritmetičke sredine, budući da smo za izračunavanje varijanse vrednosti dizali na kvadrat. Rezultat je broj 5, koji još uvek predstavlja veliko odstupanje naspram aritmetičke sredine osam, tako da je jasno da su podaci rasuti široko oko aritmetičke sredine.

Primer 19: Izračunavanje standardne devijacije

```
> sd(chol)
[1] 5.023345
```

U svakodnevnoj praksi prednost se daje primeni standardne devijacije u odnosu na varijansu, zato što je lakše shvatljiva. Grubo možemo da kažemo da ona predstavlja presek

odstupanja od aritmetičke sredine. Kad bi sva merenja imala istu vrednost, onda bi vrednost standardne devijacije bila jednaka nuli. Standardna devijacija značajno zavisi od udaljenih izmerenih vrednosti. U primeru 8 izračunali smo interkvartilni razmak sa vrednošću dva. To znači da, zapravo, u uzorku imamo ekstremne vrednosti koje značajno utiču na njegov oblik, što možemo da vidimo i iz prikazanih vrednosti.

Koeficijent varijacije

Standardna devijacija predstavlja prosek odstupanja merenja od njihovih aritmetičkih sredina. Odnos standardne devijacije i aritmetičke sredine moguće je izraziti u procentima i naziva se koeficijent varijacije. Njegovo izračunavanje je vrlo jenostavno (primer 11).

Primer 20: Izračunavanje koeficijenta disperzija

```
> (sd(chol)/mean(chol))*100
[1] 62.79182
```

Iz rezultata možemo da zaključimo da standardna devijacija predstavlja čak 63% vrednosti aritmetičke sredine, što potvrđuje značajnu disperziju podataka.

Sažetak

Mere centralne tendencije i disperzije su verovatno najčešće korišćena sredstva za opisivanje podataka. Uprkos tome, pojavljuju se česte greške prilikom njihove interpretacije ili pri njihovoj primeni. Često se izračunavaju i tamo gde nije moguće očuvati osnovne pretpostavke njihove ispravne interpretacije. Obično se aritmetička sredina dobija i tada, kada nisu zadovoljeni uslovi njene ispravne primene – ne radi se o uzorku iz normalne distribucije. Takođe, korišćenje mere centralne tendencije samostalno, bez mere disperzije – navođenje samo aritmetičke sredine, je čest nedostatak. Zaboravlja se i na medijanu, a u celosti malo se koriste kvartili i interkvartilni razmak, sa boksplotom za opisivanje uzorka. Ovo poglavlje sažima pravila prilikom odlučivanja, koju meru u datom slučaju koristiti i kako je interpretirati.

Vežbe

1. Za sve kontinuirane promenljive uzorka *cystfibr* iz datoteke *IswR* (primer 1, poglavlje 3) izračunajte vrenosti mera centralne tendencije. Koristite komande navedene u ovom poglavlju.

2. Razmotrite spljoštenost i nagib distribuiranih vrednosti promenljivih *age, height, weight, bmp, fev1, rv, frc, tlc* i *pemax* u okviru uzorka *cystfibr*. Nacrtajte histogram za ove promenljive i saznajte da li je na osnovu mera centralne tendencije kod ovih promenljivih moguće proceniti da su podaci normalno distribuirani.

3. Korišćenjem podataka iz uzorka *eba1977 i*z datoteke *IswR* (primer 2, poglavlje 3) izračunajte kvartile pojedinih promenljivih i njihov međukvartilni razmak i probajte da interpretirate rezultate.

PETO POGLAVLJE

Ocene uzorka

Sadržaj poglavlja

Cilj poglavlja

U uvodnom delu ove knjige smo govorili o smislu statistike i naveli smo da se ona bavi prikupljanjem podataka, njihovim beleženjem I, na osnovu njih, donošenjem zaključaka. Objasnili smo pojam statističke analize kao postupka zaključivanja o populaciji na osnovu rezultata dobijenih iz uzorka te populacije. Dok smo se u prethodnim poglavljima bavili radom sa podacima i naveli osnovne statističke mere, u ovom poglavlju ćemo na osnovu prikupljenih podataka iz uzorka izvoditi zaključke oceloj celoj populaciji. Dakle, počećemo sa delom statistike koji se bavi izvođenjem osobina populacije na osnovu uzorka, što se često naziva i induktivna statistika.

Ovaj deo statistike zasnovan je na koncepciji slučaja, a time prirodno nameće primenu verovatnoće. Budući da u ovom udžbeniku nije moguće objasniti celu koncepciju verovatnoće, trudićemo se da koristimo intuitivan pristup, bez formula i matematički zasnovanih koncepcija. Posle stručnog objašnjenja osnova verovatnoće, preći ćemo na postupak koji omogućava zaključivanje o stvarnim vrednostima ili disperziji kod populacije na osnovu izabranog uzorka.

Tabela 2: Ciljevi poglavlja

●Objasniti osnove verovatnoće ●Navesti neke podele uzoraka ●Objasniti pojam tačkaste i intervalske ocene ●Intervalska ocena aritmetičke sredine populacije

Neophodne osnove verovatnoće

Verovatnoća je svakodnevno korišćen pojam, koji izražava meru nesigurnosti da će se nešto desiti ili se neće desiti. Meteorolozi govore o verovatnoći kišnih padavina, lekari o verovatnoći određene dijagnoze, studenti o verovatnoći da se položi ili ne položi ispit. Verovatnoća je tesno povezana sa slučajnošću. Svi smo u školi učili kako brojem izraziti mogućnost da pri bacanju kocke dobijemo određeni broj, npr. četiri. Znamo da je ovaj rezultat upravo jedna od mogućnosti, kojih ima šest. Rezultat, jedna šestina, zovemo **verovatnoćom** da će pasti ne samo četvorka, već i bilo koji broj na kocki. S obzirom na to da je verovatnoća padanja bilo kojeg broja podjednaka, govorimo da je određeni broj pao **slučajno.** Nismo sigurni koji će broj zaista pasti.

Ovo je vrlo pojednostavljen pogled na verovatnoću. Češće postavljamo pitanja o dešavanjima iz svakodnevnog života i očekujemo odgovor koji će moći da se upotrebi u konkretnom slučaju. Ni statistika ni verovatnoća, međutim, neće dati ovakve odgovore, ne znaju da kažu ko će kad umreti, kakve bolesti ga očekuju ili kada će završiti školu i sa kakvim uspehom. Znaju da kažu da muškarac, koji se rodio 2005. godine u Slovačkoj, može da doživi 70,3 godine. To ne znači da nema muškaraca koji će umreti pre ili kasnije. Na osnovu podataka koliko muškaraca je umrlo u određenoj godini i u određenom uzrastu znamo da izračunamo verovatnoću umiranja ili preživljavanja do sledeće godine, tako što broj muškaraca umrlih u određenom uzrastu podelimo brojem svih koji su umrli. Kad to uradimo za sve uzrasne grupe (npr. petogodišnje) dobićemo verovatnoću umiranja svake od njih. Jasno je da, što je čovek stariji, to će biti veća verovatnoća umiranja.

Definicija 2: Klasična definicija (Pjer Simon de Laplas)

Verovatnoća = Broj relevantnih događaja / Broj svih mogućih događaja.

Drugim rečima: Količnik broja situacija u kojima će se desiti ono što nas zanima i zbir broja situacija u kojima će se dogoditi ono što nas zanima i broja situacija u kojima se neće dogoditi ono što nas zanima.

Normalna raspodela

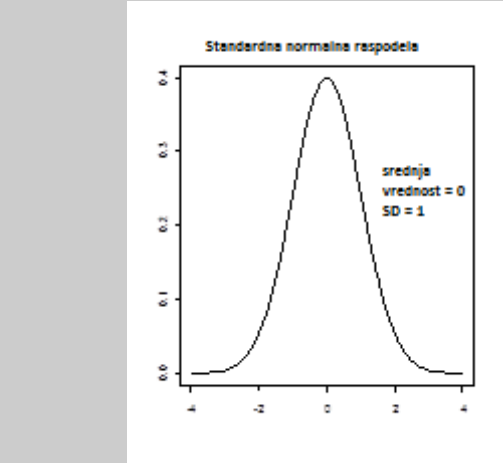
Pretpostavite da ćete izmeriti visinu svih devojaka na univerzitetu. Kakve vrednosti možemo da očekujemo? Velika većina može da ima visinu blizu proseka za devojke, što je utvrđeno istraživanjem sprovedenim u celoj Slovačkoj Republici (VI nacionalno istraživanje telesnog razvoja dece i omladine, 2001.), a to je bilo 165 cm. Prirodno je, međutim, da su neke devojke niže, a neke više. Ako posmatramo njihov broj saznaćemo da, što su devojke niže, to je i njihov broj manji i, isto tako, što su više, to je njihov broj manji. Slika 1 prikazuje rezultat.

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 vresnost 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. Celokupna površina ispod krive iznad ose x je jedinstvena. Zato posle podele površine ispod krive u mestu srednje vrednosti je 50% površine u desno i 50% u levo od srednje vrednosti.

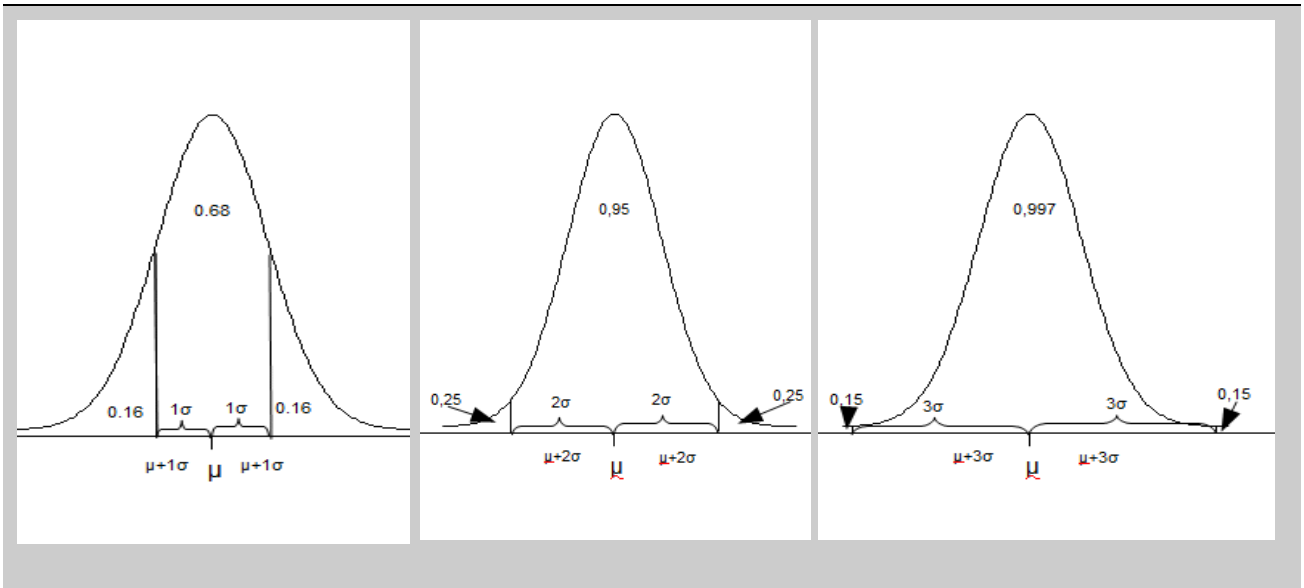
Ako površinu podelimo na mestu koje odgovara jednoj standardnoj devijaciji u desno i jednoj u levo, dobićemo površinu koja približno odgovara 68% ukupne površine, a ako koristimo dve standardne devijacije, onda će površina biti približno 95%, a kod tri podele čak 99,7% površine ispod krive (Slika 2).

Normalna raspodela je tačno određena sa dva parametra: μ, odnosno srednja vednost populacije i σ, odnosno varijansa populacije. Razne vrednosti μ pomeraju grafikon podele duž ose x i razne vrednosti σ menjaju njenu zašiljenost (slika 3).

Slika 1: Kriva standardne normalne raspodele sa srednjom vrednosti 0 i standardnom devijacijom 1

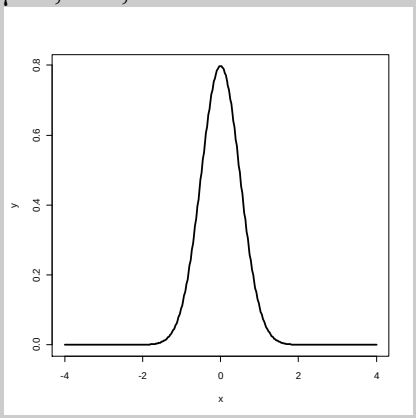


Slika 2: Površina ispod krive standarne normalne raspodele

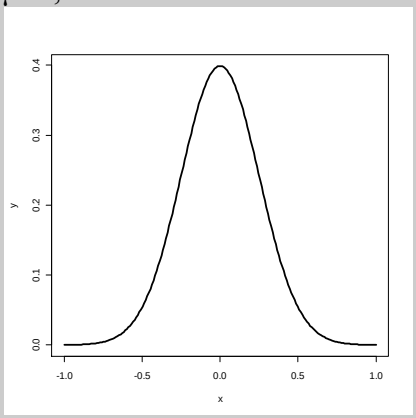


Slika 3: Razni oblici krive standardne normalne raspodele pri različitm vrednostima disperzije i pri konstantnoj aritmetičkoj sredini

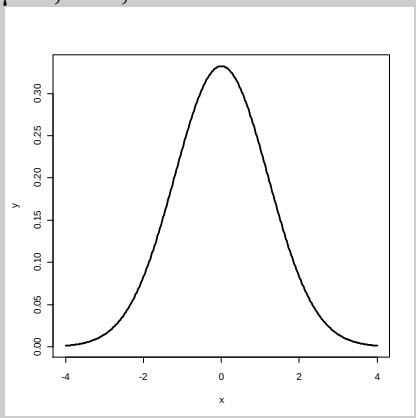
μ=0, σ=0,5



μ=0, σ=1



μ=0, σ=1,2

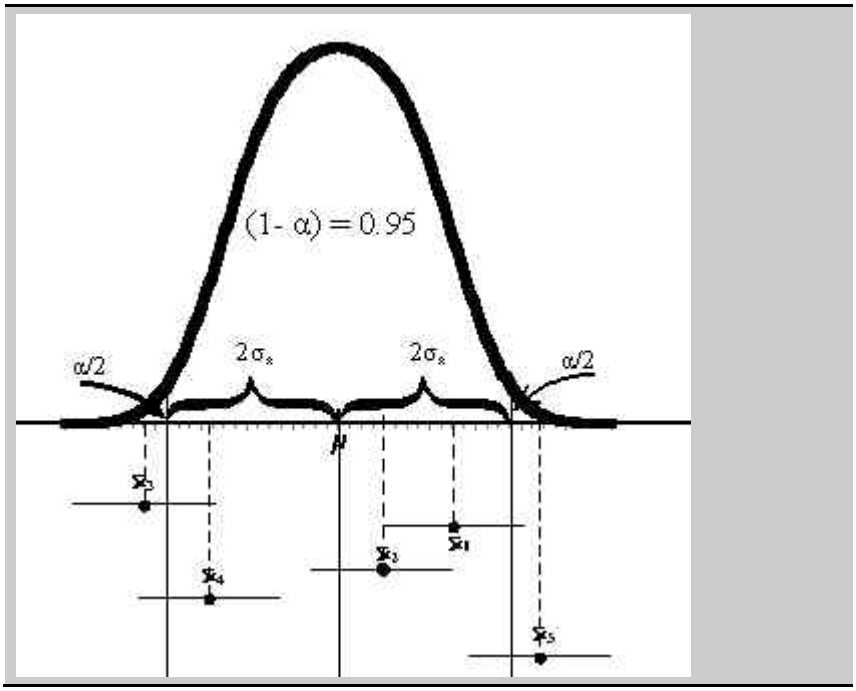


Slučajni uzorak

Princip odabira članova ovog uzorka podrazumeva da svako iz populacije mora da ima jednaku šansu da postane pripadnik uzorka. Sledeći uslov je nezavisnost uzorkovanja, dakle da izbor jednog pripadnika ne sme da ometa izbor bilo kojeg sledećeg. Svakodnevno se ovi uslovi obezbeđuju korišćenjem tabela slučanih brojeva.

I kada obezbedimo da svaki član populacije ima istu šansu da dospe u uzorak, to ne znači da će svi odabrani imati iste osobine. Razlike u srednjoj vrednosti i disperziji mogu da budu različite. Sledeća slika nam ilustruje situaciju ponavljanih uzorkovanja iz normalno distribuirane populacije. Uzorci su označeni njihovom srednjom vrednošću (isprekidana linija) i disperzijom (puna linija).

Slika 4: Ponavljani uzorci iz normalno raspodeljene populacije



Svaki od uzoraka ima različitu srednju vrednost (pogledajte srednje vrednosti uzoraka označenih sa $\bar{x}$) i različitu disperziju. Tri prosečne vrednosti pale su u oblast ograničene sa dve standardne devijacije, dakle 2σ. Dve su pale van ove oblasti - prisetimo se da se u ovoj oblasti nalazi 95% svih podataka, a time i izvora mogućih uzoraka. Ovako uobičajeno izgleda ono šta radi istraživač prilikom korišćenja statističke analize: napravi slučajni uzorak, izračuna njegove karakteristike i proba da sazna gde se nalazi stvarna srednja vrednost populacije i kakva je njena stvarna disperzija. Problem je, u stvari, da ne zna koji uzorak je uzeo.

Utvrđivanje srednje vrednosti primenom slučajne raspodele

Vratimo se na osnovno pitanje statistike, a to je statistička analiza i kako će nam poznavanje normalne raspodele pomoći u ovoj temi. Razjasnićemo da se visina devojaka u Slovačkoj kreće oko prosečne vrednosti 165 cm i neka je standardna devijacija 10 cm. Ali pri svakom istraživanju ipak ne možemo da ispitujemo sve devojke starosti 18 godina i više u Slovačkoj. U ovakvom slučaju koristićemo **slučajni uzorak**. Iz primera sa slike 4 proizilazi da što više se ovakvih uzoraka iz populacije napravi, to je srednja vrednost njihovih srednjih vrednosti bliža realnoj srednjoj vrednosti populacije ili, što je dobijeni slučajni uzorak veći, time je njegova srednja vrednost bliža realnoj. Zato se trudimo da saznamo adekvetnu veličinu uzorka, tako da on najbolje reprezentuje celu populaciju. Takođe moramo da shvatimo da je disperzija između uzoraka zavisna delom i od izbora uzorka. Disperziju populacije ne poznajemo, zato moramo da se orijentišemo ka njenom otkrivanju putem uzorka, čija mera je standardna devijacija. Kada smo razjasnili da varijansa uzorka zavisi od veličine uzorka, moramo je jednostavno dovesti u red tako što ćemo je podeliti brojem merenja u uzorku. Kako bi se lakše izračunavalo, umesto verijanse se koristi standardna devijacija i podeli se korenom broja. Ovaj odnos se označava kao **standardna greška aritmetičke**

sredine (standard error of the mean SEM).

Sada možemo da se vratimo na osnovnu temu ovog poglavlja, posvećenu analitičkoj statistici, odnosno induktivnoj statistici. Najpre ćemo prikupiti činjenice: znamo da imamo populaciju iz koje slučajnim izborom dobijamo slučajni uzorak. Znamo da izračunamo aritmetičku sredinu i standardnu devijaciju ovog uzorka, a time i da odredimo mere centralne tendencije. Poznate su nam osnovne karakteristike normalne raspodele i očekujemo da populacija iz koje je uzorak napravljen je normalno raspodeljena sa aritmetičkom sredinom μ i varijansom σ. Pitanje je kakva je realna vrednost aritmetičke sredine u populaciji?

Računica proizilazi iz činjenice da 95% svih merenja je približno u obimu $\pm 2\sigma$, dakle u rasponu dve varijanse sa obe strane aritmetičke sredine. Pogledajte da u ovom slučaju znamo realnu disperziju (varijansu) parametra u populaciji. Situacija u kojoj poznajemo disperziju populacije, ali ne i njenu aritmetičku sredinu je veštačka (modelska), za svrhe objašnjavanja principa postupka. Postupak je ilustrovan primerom 1.

Primer 21: Približno utvrđivanje stvarne prosečne vrednosti glikemije populacije

Istraživač je merio vrednosti šećera u krvi. Merenje je napravio na 15 ljudi sa sledećim vrednostima u mmol/l:
5.3, 6.1, 5.8, 6.2, 6.4, 5.9, 6.7, 7.1, 6.3, 6.6, 7.4, 6.6, 5.7, 6.4, 5.5

```
> gly<-c(5.3, 6.1, 5.8, 6.2, 6.4, 5.9, 6.7, 7.1, 6.3, 6.6, 7.4, 6.6, 5.7, 6.4, 5.5)
> gly
 [1] 5.3 6.1 5.8 6.2 6.4 5.9 6.7 7.1 6.3 6.6 7.4 6.6 5.7 6.4 5.5
> length(gly)
[1] 15
> mean(gly)
[1] 6.266667
> var(gly)
[1] 0.3323810
> sd(gly)
[1] 0.5765249
```

Aritmetička sredina uzorka je 6,3 mmol/l (zaokruživaćemo zbog preglednosti), varijansa 0,33 mmol/l i standardna devijacija 0,58 mmol/l.
Sada nam je potrebno da izračunamo koji deo disperzije odgovara veličini izabranog uzorka

```
> sqrt(var(gly)/length(gly))
[1] 0.1488581
> sd(gly)/sqrt(length(gly))
[1] 0.1488581
```

Obe metode izračunavanja dale su isti rezultat. Probajte i komandu *std.error* iz datoteke plotrix. Daje nam istu vrednost.

```
> std.error(gly)
[1] 0.1488581
```

Rezultat: 6,3 ± 0,15= 6,45 i 6,15 interpretiramo tako da je stvarna standardna devijacija glikemije u populaciji između 6,15 mmol/l i 6,45 mmol/l.

Da bi približni rezultat doveli do tačnog, moramo da koristimo još jedan parametar, a on se zove **koeficijent poverenja** i označava se kao malo z. On se uvodi i iz razloga da nas nekad ne zanima 95% verovatnoće. Možemo da radimo i sa drugim vrednostima, npr. 99%. Aritmetička sredina uzorka se obično zove i **tačkasti rezultat aritmetičke sredine** populacije i njegova vrednost se navodi kao **procenitelj** (estimator). Ishod ovakvog rezultata, nazvan kao intervalski rezultat aritmetičke sredine se izračunava:

procenitelj ± (koeficijent poverenja) x (standardna greška uzorka).

Vrednosti koeficijenta poverenja naći ćete u tabeli normalne raspodele. Tu se, u stvari, navodi vrednost za nivo poverenja *(1 – α)/2*. Alfa (α) je vrednost verovatnoće za koju srednja vrednost uzorka ne odgovara aritmetičkoj sredini populacije. Takođe treba da shvatimo da se verovatnoća iskazuje ne u procentima, već u decimalnim brojevima između 1 i 0. Tako da alfa sa 99% verovatnoće, što je 0,99, iznosi 0,01, zato što je 1 – 0,99 = 0,01. Vrednost z treba tražiti u tabelama, ali nju program izračunava kao u sledećem primeru, koji pokazuje izračunavanje iz prethodnog primera 1.

Primer 2: Tačna vrednost intervala stvarne srednje vrednosti populacije

```
> gly
 [1] 5.3 6.1 5.8 6.2 6.4 5.9 6.7 7.1 6.3 6.6 7.4 6.6 5.7 6.4 5.5
> alfa<-(1-0.95)/2
> alfa
[1] 0.025
> z<-qnorm(alfa)
> z
[1] -1.959964
> sem=sd(gly)/sqrt(length(gly))
> sem
[1] 0.1488581
> donji<-mean(gly)-abs(z*sem)
> donji
[1] 5.97491
> gornji<-mean(gly)+abs(z*sem)
> gornji
[1] 6.558423
```

Rezultat: 95% srednjih vrednosti glikemije uzoraka populacije je između 5,97 i 6,56 mmol/l, ili možemo da kažemo da smo 95% sigurni da se stvarna prosečna vrednost glikemije populacije iz koje dolazi uzorak nalazi među 5,97 i 6,56 mmol/l.

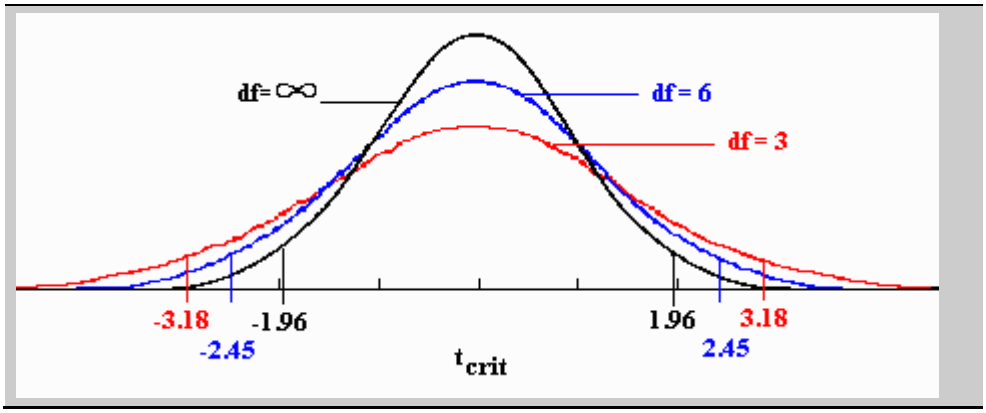
Ovaj interval se zove **intervalsko utvrđivanje** srednje vrednosti ili **interval poverenja** srednje vrednosti ili **confidence interval** uzorka. Kao takav, predstavlja jako bitan deo statističkih procedura. U savremenoj primenjenoj statistici korišćenje intervala poverenja uzorka je jako česta i daje mu se prednost nad uobičajenim opisivanjem osobine uzorka pomoću aritmetičke sredine

i standardne devijacije.

t raspodela

Procenu stvarne srednje vresnosti radili smo na osnovu više pretpostavki: jedna od njih bilo je očekivanje da uzorak dolazi iz normalno distribuirane populacije; u drugom slučaju smo očekivali da poznajemo varijansu populacije σ, a ujedno smo skrenuli pažnju da ovakva situacija nije svakodnevni slučaj i koristili smo je kao model za objašnjenje postupka. 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**. Autor je V. S. Goset, koji je pisao pod pseudonimom „Student" i zato je i ušao u upotrebu takav naziv. Slično kao normalna raspodela i t-raspodela ima srednju vrednost 0, a ujedno je simetrično raspodeljena oko srednje vrednosti. Za razliku od normalne raspodele (koja je imala disperziju jednaku jedan), u Studentovoj raspodeli je disperzija veća od 1; međutim, prilikom većih uzoraka se približava ka njoj i time se pravi cela porodica raspodela, kada je za svaku vrednost $n - 1$ drugačija raspodela. U poređenju sa normalnom distribucijom, nema tako oštar vrh, a ima viši rep (slika 5). Vrednost $n - 1$ se obično zove i **stepen slobode** ili broj stepena slobode, tako da je pri svakom različitom stepenu slobode i t raspodela drugačija. Što viši stepen slobode - to i veća brojnost u uzorka, a tim se i t raspodela više približava prema normalnoj raspodeli.

Slika 5: Studentova t podela prilikom različitih stepena slobode



Korišćenje ove podele prilikom utvrđivanja stvarne srednje vrednosti u populaciji je zasnovano na istom postupku kao pri i korišćenju vrednosti normalne raspodele:

procenitelj ± *(koeficijent poverenja) x (standardna greška u uzorku).*

Razlika je u izračunavanju koeficijenta poverenja, koji se izračunava na osnovu vrednosti t raspodele prema odgovarajućem broju stepena slobode. Prednost ovog postupka je prevashodno činjenica da bolje prikazuje situacije, kada osim srednje vrednosti populacije otkrivamo i disperziju populacije na osnovu njihovih tačaka ocene, koje smo dobili iz uzorka. Uzorak mora da bude dovoljno veliki, obično ne manje od 20 članova. Ujedno je ovaj postupak manje osetljiv na

verovatnoću normalne raspodele, a poneće i manja odstupanja od njega.

Intervalska ocena uzorka populacije

Spojićemo znanje o t raspodeli sa postupkom ocene uzorka i saznaćemo precizniju ocenu stvarne srednje vrednosti populacije (primer 3).

Primer 22: Izračunavanje intervala poverenja za procenu srednje vrednosti populacije korišćenjem t podele

```
> gly<-c(5.3, 6.1, 5.8, 6.2, 6.4, 5.9, 6.7, 7.1, 6.3, 6.6, 7.4, 6.6, 5.7, 6.4, 5.5) # Pravljenje vektora sa podacima iz primera 1
> gly
 [1] 5.3 6.1 5.8 6.2 6.4 5.9 6.7 7.1 6.3 6.6 7.4 6.6 5.7 6.4 5.5
> mean(gly) # Izračunavanje srednje vrednosti uzorka
[1] 6.266667
> sem=sd(gly)/sqrt(length(gly)) # Izračunavanje standardne greške uzorka
> sem
[1] 0.1488581
> alfa<-(1-0.95)/2 # Izračunavanje verovatnoće za koeficijent poverenja α
> alfa
[1] 0.025
> df <- length(gly-1) # Vrednost stepena slobode
> df
[1] 14
> t<- qt(alfa, df) # vrednost t raspodele za vrednost alfa i vrednost stepena slobode
> t
[1] -2.144787
> donji<-mean(gly)-abs(t*sem) # donja granica intervalne ocene srednje vrednosti uzorka populacije
> donji
[1] 5.947398
> gornji<-mean(gly)+abs(t*sem) # gornja granica intervalne ocene srednje vrednosti uzorka populacije
> gornji
[1] 6.585936
```

Rezultat interpretiramo u smislu zadatka, na primer tako da će 95% svih odabranih srednjih vrednosti iz populacije biti između 5.95 i 6.59 mmol/l.

Sigurno ste radoznali da nije neko izmislio jednostavniji proces kao dobiti inerval poverenja za srednju vrednost populacije. Naravno da je izmislio. U R se nalazi komanda kojom se korišćenjem t raspodele osim ostalog računa i interval poverenja. Ova komanda *t.test()* izračunaće, osim intervala poverenja, i neke druge parametre o kojima ćemo govoriti u sledećim poglavljima.

Za ilustraciju ćemo koristićemo podatke o glikemiji iz prošlog primera, a specifikacijom promenljive *gly*, u koju smo je stavili u prethodnim koracima (ako niste isključili u međuvremenu program R) i ujedno određivanjem stepenom spoverenja 95% i zadavanjem *conf.level = 0.95* dobijamo brzo rezultat (primer 4).

Primer 23: Interval poverenja za srednju vrednost populacije sa korišćenjem komande t.test()

```
>t.test(gly, conf.level = 0.95)

        One Sample t-test

data:  gly
t = 42.0983, df = 14, p-value = 3.819e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.947398 6.585936
sample estimates:
mean of x
 6.266667
```

Intervalske ocene proporcije kod populacije

Studenti (ali ne samo oni) rado za izražavanje razlika među merenjima koriste procente. Procenat je deo (proporcija) celine preračunat na 100. Jako često onda interpretiraju razliku između izračunatih procenata kao razliku između merenja, a pri tom nisu svesni da je ovaj zavisan, između ostalog, i od veličine uzorka. Zato je jako značajno izračunati intervalsku ocenu proporcije, isto kao u slučaju srednje vrednosti. Nećemo ga zato ilustrovati prema pojedinim koracima, već ćemo odmah preći na korišćenje komande na primeru studije u kojoj je utvrđeno da je u uzorku 185 mladih ljudi od njih 65, odnosno 35%, redovno pušilo. Pitanje sledi: kakva je stvarna proporcija pušača u populaciji pri verovatnoći 99%? Izračunavanje ilustruje primer 5.

Primer 5: Izračunavanje intervalske ocene proporcije

```
> prop.test(65, 185, conf.level = .99)
> prop.test(65, 185, conf.level = .99)
        1-sample proportions test with continuity
correction

data:  65 out of 185, null probability 0.5
X-squared = 15.7622, df = 1, p-value = 7.182e-
05
alternative hypothesis: true p is not equal to 0.5
```

```
99 percent confidence interval:
 0.2650632 0.4482346
sample estimates:
      p
0.3513514
```

Obratite pažnju da smo u komandi *prop.test()* najpre odredili proporciju 35, a onda celukopan broj u uzorku, a na kraju nivo poverenja, za koji izračunavamo interval poverenja. Rezultat govori da je 99% svih proporcija pušača u populaciji između 26% i 45%.

Sažetak

Pažljiv čitalac naučio je da korsti statističku analizu za procenu stvarne srednje vrednosti populacije na osnovu srednje vrednosti i standardne devijacije uzorka. Takođe je saznao kako otkriti interval u kome se nalaze proporcije kod populacije, na osnovu proporcije odabranog uzorka. Ovo znanje će se stalno ponavljati u raznim primerima u sledećim poglavljima. Ujedno je naučio o suštini dve vrste raspodele, normalnoj i t raspodeli, a takođe i o prednostima za njihovo korišćenje. Opet napominjemo da uvođenje tačkaste ocene srednje vrednosti ili proporcije bez navođenja intervalske procene sa specifikacijom verovatnoće, mnogi nastavnici mogu da kvalifikuju kao grešku.

Primeri[24]

Primer 1.

Istraživač se interesovao za vrednost nekih metala u krvih zdravih davalaca. Slučajnim izborom odabrao je grupu davalaca i dobio je navedene podatke:

Se	**Zn**	**Cu**	**Mg**
55	16,97	10,8	0,84
136,9	18,67	12,4	0,91
110,6	18,16	12,2	0,92
99,4	16,84	16,1	0,83
51,9	17,4	8,8	0,87
100	12,7	8,8	0,85
149,4	12,77	9,2	0,83
167,5	16,1	12,2	0,94
79,6	16,84	11,7	0,96
161,2	12,61	10,8	0,9
149,4	16,32	9	0,89
274,4	21,65	9,7	0,96
33,1	14,11	10,4	0,85
101,1	17,61	15,1	0,97
53,3	16,52	13,2	0,87

Hteo bi da sazna sa 99% verovatnoće, kakva je stvarna vrednost nivoa navedenih metala kod populacije.

Primer 2.

Antropolog je istraživao razmere lobanja u uzorku slučajnog uzorka pojedinaca iz plemena Kikuyu. Merio je sledeće parametre: MB- Maksimalna širina lobanje , BH – Visina lobanje od baze, BL – Visina lobanje od alveola, NH – Nazalna visina lobanje i dobio je ove podatke:

MB	**BH**	**BL**	**NH**
131	138	89	49
125	131	92	48
131	132	99	50
119	132	96	44
136	143	100	54
138	137	89	56
139	130	108	48

[24] Upozorenje: za korektno koišćenje t testa neophodno je imati na raspolaganju barem 20 merenja iz uzorka, u protivnom moramo da koristimo neparametarske metode, jer nemamo dovoljno jaku verovatnoću iz približno normalne raspodele – t. Ako je veličina uzorka veća od 30, možemo da koristimo t raspodelu – zato što u tom slučaju t i N podele su iste. Međutim, nećemo da za provežbavanje postupaka povećavamo broj merenja, da čitalac ne bi bio odvraćen od izračunavanja primera.

125	136	93	48
131	134	102	51
134	134	99	51
129	138	95	50
134	121	95	53
126	129	109	51
132	136	100	50
141	140	100	51

Interesovalo ga je kakve su prosečne vrednosti navedeih parametara lobanja kod populacije sa verovatnoćom 95%.

Primer 3

Student je pratio posećenost fitnes centra u uzorku 150 ljudi slučajno pitanih na ulici. Saznao je da 28 njih posećuje fitnes centar najmanje jednom nedeljno, a 89 ovakav centar nikad nije posetilo. Hteo je da zna, kakva je stvarna proporcija redovnih posetilaca i onih koji još nikada nisu bili, na nivou pouzdanosti 90%, 95% i 99%.

ŠESTO POGLAVLJE

Potvrđivanje hipoteza

Sadržaj poglavlja

Cilj poglavlja

U prethodnom poglavlju smo zavirili u principe statističke analize srednje vrednosti, odnosno proporcija promenljivih u populaciji, a na osnovu poznavanja njihovih tačkastih procena u slučajnom uzorku. Sada ćemo ovaj princip raširiti tako što ćemo potvrđivati hipoteze o osobinama parametara populacije, opet na osnovu poznavanja parametara **slučajnog uzorka**. Uvešćemo proceduru koja omogućava takvo zaključivanje, a ujedno ćemo i prodiskutovati koncept greške u statistici. Ovaj koncept će nam omogućiti ispravnu primenu i naročito interpretiranje p vrednosti u vezi sa značajnoću određene hipoteze.

Tabela 3: Ciljevi poglavlja

●Objasniti pojam hipoteze i njene vrste ●Opisati postupak potvrđivanja hipoteza ●Kako tačno interpretirati rezultate potvrđivanja hipoteza ●Kakve vrste sta stističkih grešaka postoje

Postupak prilikom potvrđivanja hipoteza

Polazeći od koncepta statističke analize, shvatamo postupak potvrđivanja hipoteze kao pomoć istraživaču, lekaru kliničaru ili drugom stručnjaku, prilikom zaključivanja o populaciji na osnovu ispitivanja uzorka date populacije. S obzirom na činjenicu da su ocene i potvrđivanje hipoteza veoma blisko povezane i koriste isti koncept, ovo poglavlje je prirodni nastavak prethodnog. Pokazaćemo i da korišćenje intervalske ocene vodi do istih zaključaka kao korišćenje postupka potvrđivanja hipoteza. Međutim, radi lakšeg razumevanja, prodiskutovaćemo ova dva postupka zasebno.

U uvodnom delu potrebno je razjasniti šta se misli pod pojmom hipoteze. Veoma uopšteno možemo da kažemo da se radi o prosuđivanju o jednoj ili više populacija. Testiranjem hipoteze se utvrđuje da li je data pretpostavka spojiva sa dostupnim podacima. Na primer, specijalista higijene želi da sazna da li je koncentracija metala u uzorcima vode u dozvoljenim granicama. Zato postavlja hipotezu, dakle pretpostavlja, da će intervalska ocena srednje vrednosti Fe u vodi na osnovu slučajno odabranih uzoraka biti u dozvoljenim granicama, odnosno veća od donje granice dozvoljene koncentracije i manja od gornje granice. Možda će ga zanimati da li je ova vrednost jednaka npr. vrednosti 0,03 mg na litar. Postupak potvrđivanja hipoteze predstavljena je postupnim koracima (Tabela 2).

Tabela 2: Postupak prilikom potvrđivanja hipoteza

• Podaci • Pretpostavke • Hipoteze • Statistički test

- Raspodela rezultata testa
- Odlučujuće pravilo
- Rezultati statističkog testa
- Statistička odluka
- Administrativna ili klinička odluka

Pojedine korake ćemo postepeno objasniti. Skrećemo pažnju da je sve korake neophodno uraditi prilikom svakog procesa potvrđivanja hipoteza, iako iskusan statističar ne mora da radi mnoge od njih tako detaljno, kao što ćemo mi sada uraditi.

1. Podaci

Na početku cele procedure moramo da upoznamo suštinu podataka koji čine osnovu testirajućih procedura. Mogu se koristiti mere centralne tendencije, disperzije, kao i grafičko prikazivanje, dakle deskriptivna statistika. Na osnovu ovog poznavanja prelazimo na sledeći korak.

2. Pretpostavke

Pretpostavke se odnose na raspodelu populacije iz koje potiče uzorak. Pretpostavljamo da uzorkujemo iz normalno distribuirane populacije ali, kako će se pokazati u narednim poglavljima, može da se pretpostavi i drugi način raspodele. Ujedno možemo da pretpostavimo da su uzorci nezavisni, ali da postoji jednaka disperzije u uzorcima pri primeni određenog postupka uzorkovanja. Pretpostavki se napravi više, a često se dešava da im istraživač posvećuje malu, ili nikakvu pažnju. Onda se iznenadi da rezultati ne odgovaraju njegovim očekivanjima.

3. Hipoteze

Potvrđivanje hipoteza pretpostavlja uvek dve hipoteze, od kojih jednu zovemo nultom, a drugu alternativnom. Hipoteza koja treba da se potvrđuje je prva od njih, dakle **nulta hipoteza**, označavana kao $\mathbf{H_0}$. Obično se naziva i hipotezom neprisutnosti razlike, zato što se prilikom formulacije statističke hipoteze svako pitanje koje želimo da potvrdimo (ili opovrgnemo) trudimo da preformulišemo u formu tako da pretpostaljamo da nema razlike (drugim rečima, razlika je jednaka nuli) između dva ili više posmatranih parametara. Ova koncepcija traži da budemo svesni cilja statističkog saznanja i njegove formulacije. Pretpostavljamo da postoji razlika između dva uzorka, ali potvrđujemo nultu hipotezu da su njihovi parametri, npr. srednje vrednosti jednake, odnosno da između njih nema razlike. Na primer, potvrđujemo hipotezu da je srednja vrednost jedne promenljive kod populacije jednaka srednjoj vrednosti druge populacije (dakle razlika između njih je 0 – nulta hipoteza), ili da je srednja vrednost jednaka nekoj vrednosti (dakle, razlika srednje vrednosti i pretpostavljene vrednosti jednaka je nuli, ili je disperzija jedne populacije jednaka disperziji druge populacije. To je tvrdnja o saglasnosti (ili nedostatku razlike) sa stvarnim uslovima u populaciji o kojoj se interesujemo. Proces potvrđivanja hipoteza onda vodi ka saznanju da li prihvatamo nultu hipotezu o nepostojanju razlike, odnosno o jednakosti parametara ili je odbacujemo. Tada automatski prihvatamo **alternativnu hipotezu $\mathbf{H_A}$**, koja govori o postojanju

razlike među parametrima.

Pretpostavimo da želimo da potvrdimo da li je prosečan uzrast ispitivane populacije veći ili manji od 25 godina (različit od vrednosti 25 godina). Formulisaćemo obe hipoteze tako da nulta pretpostavlja da je prosečan uzrast ispitivane populacije 25 godina, dok alternativna pretpostavlja da je iznad ili ispod 25 godina. U skraćenoj formi zapisaćemo to ovako:

$$H_0 : \mu = 25 \qquad \text{alternativna hipoteza je} \qquad H_A: \mu <> 25$$

U slučaju kada upoređujemo dve populacije, možemo da potvrdimo da li se srednja vrednost jedne razlikuje od srednje vrednosti druge. Na primer, hoćemo da saznamo da li je savetovalište za smanjenje telesne mase dovelo do smanjenja kod slučajno odabranog uzorka u poređenju sa kontrolnom grupom, koja nije koristila usluge savetovališta. Onda se hipoteze formulišu tako da ako oznakom μ_1 označimo vrednost promenljive telesna masa populacije koja koristi usluge savetovališta (dobijenom na osnovu slučajnog uzorka pojedinaca koji su koristili usluge savetovališta), a sa μ_2 označimo vrednost promenljive telesna masa kod populacije kojoj savetovanje nije bilo pruženo (dobijenu na osnovu slučanog uzorka pojedinaca koji nisu prisustvovali savetovanju), onda hipoteze formulišemo tako da je nulta hipoteza da su obe prosečne vrednosti telesne mase jednake, dok je alternativna hipoteza da nisu jednake.

$$H_0 : \mu_1 = \mu_2 \qquad H_A : \mu_1 <> \mu_2$$

Ako nas interesuje samo efekat savetovanja na redukciju telesne mase, onda ćemo utvrđivati istu nultu hipotezu, ali trudićemo se prihvatimo samo alternativnu hipotezu da će prosečna vrednost telesne mase u grupi sa savetovanjem biti manja nego u grupi bez savetovanja:

$$H_0 : \mu_1 = \mu_2 \qquad H_A : \mu_1 < \mu_2$$

Nulta hipoteza će se ili odbaciti ili prihvatiti. Ako se odbaci, onda možemo da obrazložimo da dobijeni podaci nisu u skladu sa nultom hipotezom, međutim podržavaju alternativnu hipotezu. Ako se nulta hipoteza ne odbaci (prihvati se), onda možemo da obrazložimo da podaci nisu pružili dovoljno informacija za odbacivanje nulte hipoteze. Prihvatanje ili odbacivanje nulte hipoteze još ne znači da je to u stvarnosti dokaz prisutnosti ili neprisutnosti razlike u parametrima. Izražavamo se u smislu da (sa određenom pretpostavkom) ova pojava može da se desi. Ova koncepcija je suštinska za ispravnu interpretaciju rezultata bilo kakvog procesa statističke indukcije.

4. Statistički test

Primena odgovarajućih statističkih testova direktno utiče na prihvatanje ili odbacivanje nulte hipoteze. Statistički test je algoritam ili postupak (većinom izražen formulom), čija ishodišna vrednost određuje da li će nulta hipoteza biti prihvaćena ili odbačena. Ishod statističkog testa je broj koji leži na osi x raspodele populacije. Izbor adekvatnog statističkog testa je vođen karakterom pitanja na koje želimo statistički da damo odgovor, a takođe i karakterom upotrebljenih podataka iz uzorka.

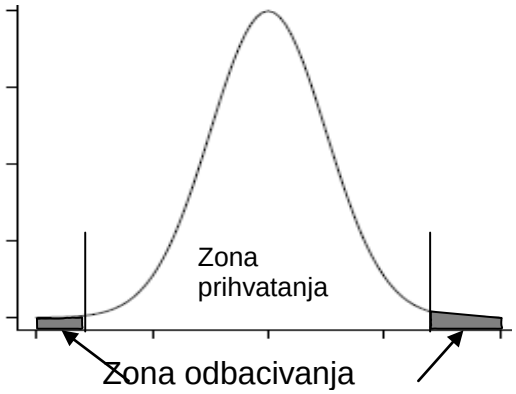
5. Podela statističkih testova

Podela statističkih testova proizilazi iz poznavanja osobina uzorka i formulacije nulte hipoteze. Većinom se navodi kod pojedinih statističkih testova.

6. Odlučujuće pravilo

Odlučujuće pravilo govori o tome šta treba da se desi sa nultom hipotezom. Kada sebi predstavimo da su sve moguće vrednosti rezultata statističkog testa tačke na x osi grafikona njegove distribucije, onda su te tačke, koje se prilikom prihvatanja nulte hipoteze pojave sa malom verovatnoćom, u tzv. **zoni odbacivanja** nulte hipoteze. Tačke koje se pojave sa visokom verovatnoćom su u tzv. **zoni prihvatanja** nulte hipoteze (Slika 1). Prema pravilima odbacujemo nultu hipotezu ako je rezultat statističkog testa, izračunat u uzorku, u rasponu zone odbacivanja. Ako izračunata vrednost nije u zoni odbacivanja, ali jeste u zoni prihvatanja, onda se nulta hipoteza prihvata. Nivo značaja određuje površinu ispod krive razlaganja rezultata statističkog testa iznad zone odbacivanja i govorimo o verovatnoći odbacivanja nulte hipoteze.

Slika 1: Zona prihvatanja ili odbacivanja nulte hipoteze



7. Statistički rezultati

Upoređuju se sa zonom prihvatanja ili odbacivanja nulte hipoteze i sazna se gde se dobijene vrednosti nalaze u pogledu na obe zone.

8. Statističko zaključivanje

U ovom koraku odlučićemo se o sudbini nulte hipoteze. Ako je vrednost statističkih rezultata u rasponu zone prihvatljivosti nulte hipoteze, onda zaključujemo da H_0 ne možemo da odbacimo i prihvatamo je. Ako dobijena vrednost padne u zonu odbacivanja, onda zaključujemo da odbacujemo istinitost H_0 i možemo da prihvatimo alternativnu hipotezu H_A.

9. Naučno zaključivanje

Naučno zaključivanje interpretira statistički zaključak i procenjuje naredni postupak, npr. povećanje veličine uzorka. Jako je bitno i opredeliti se da li je statistički razlika (veličina diferencije) bitna (dovoljno velika) i kod naučnog zaključivanja.

Potvrđivanje hipoteza

Proces potvrđivanja hipoteza približićemo na nekoliko primera. Jedan od njih će biti potvrđivanje hipoteza o srednjoj vrednosti, sledeći će se ticati razlike aritmetičkih sredina, poslednji govori o razlici proporcija.

1. Potvrđivanje hipoteze o aritmetičkoj sredini populacije

Polazimo od opšte pretpostavke da biramo iz populacije sa normalnom raspodelom, iako ovu pretpostavku nekada nije moguće zadovoljiti. Kad se očekuje da je odstupanje od normalne raspodele malo, opravdano je da se koriste ovi postupci potvrđivanja hipoteze. Ako bi se raspodela bitno razlikovala od normalne, onda bi primena ovih postupaka predstavljala veliku grešku i neohodno je koristiti druge postupke, npr. neparametrijske testove (bliže o tome o neparametrijskim testovima u poglavlju 10).

(a) *Obostrani test*

Primer 1: Zadavanje primera za obostran test

U studiji o teškim povredama mozga na odeljenju utvrđeno je da su pacijenti primljeni sa ovom dijagnozom imali sledeći nivo svesti, meren Glazgovskom skalom (GCS):
5, 4, 4, 5, 6, 5, 4, 5, 5, 6, 4, 5, 5, 4, 4, 5, 6, 5, 5, 5, 5, 5, 5, 5, 4, 5, 6.
Lekara je zanimalo da li je moguće da se kaže da su pacijenti primljeni na odeljenje sa prosečnim GCS jednakom vrednosti 5.

Prema postupku potvrđivanja hipoteza najpre steknemo mišljenje o podacima.

Podaci

Podaci dolaze iz bolničkih beleženja o 27 pacijenata, primljenih na odeljenje sa teškom povredom mozga. Prosečna vrednost je 4,9.

Pretpostavke

Pretpostavljamo da su podaci normalno raspodeljeni i njihovu raspodelu saznajemo pomoću standardne devijacije 0,64.

Hipoteze

Pitanje u zadatku je bilo da li je poguće pretpostavljati da je stvarna srednja vrednost populacije jednaka vrednosti 5. Dakle, formulacija hipoteza će biti sledeća:

$H_O: \mu = 5$ $H_A: \mu \neq 5$

Statistički test

Budući da naš uzorak dolazi iz populacije sa normalnom raspodelom, a disperziju ove populacije saznajemo pomoću standardne devijacije, koristićemo test Studentove t

raspodele[25].

Raspodela statističkog testa

Ako je H_0 tačna, onda je testirajuća statistika podeljena prema Studentovoj t raspodeli sa *n – 1* stepena slobode.

Odlučujuće pravilo

Ovo pravilo će nam reći da li da prihvatimo ili odbacimo nultu hipotezu da je srednja vrednost jednaka vrednosti 5. Kada će biti H_0 pogrešna? Onda, kada stvarna srednja vrednost bude manja od 5 ili veća od 5. Kolika treba da bude razlika? Odgovor je u verovatnoći sa kojom želimo da odbacimo istinsku H_0, odnosno da napravimo grešku tipa I. Opredelimo se da će to biti na nivou $\alpha = 0,05$, dakle sa verovatnoćom 5%. Moramo da shvatimo da nas zanimaju vrednosti ne samo veće, nego i manje od 5. Dakle, testiraćemo na dve strane, pa zato moramo da verovatnoću greške razdelimo na dve strane, na jednoj $\alpha/2 = 0,025$ i isto na drugoj strani $\alpha/2 = 0,025$. Pitamo se koje će vrednosti iz naše raspodele odrediti gde leže navedene granice. Te vrednosti ćemo dobiiti iz tabele za t raspodelu, pri broju stepena slobode n - 1, tj. 26 ili korišćenjem komande *qt(),* gde postavimo nivo verovatnoće 0,025, sa kojom smo odlučili da računamo, a takođe i broj stepeni slobode, u ovom slučaju 14. Zadavanjem *> qt(0.025, 26),* R će nam dati *[1] -2,055529.* Ovu vrednost ćemo dalje zvati **kritičkom vrednošću statistikog testa**. Budući da proveravamo na obe strane da li je stvarna vrednost manja ili veća od broja 5, onda govorimo o obostranom testu. U ovakvom slučaju zona prihvatanja nulte hipoteze biće između brojeva -2.055529 i 2.055529, a zona odbacivanja biće vrednost izračunatog statističkog testa manja od -2.055529 ili veća od 2.055529 (slika 2).

Izračunata statistička vrednost

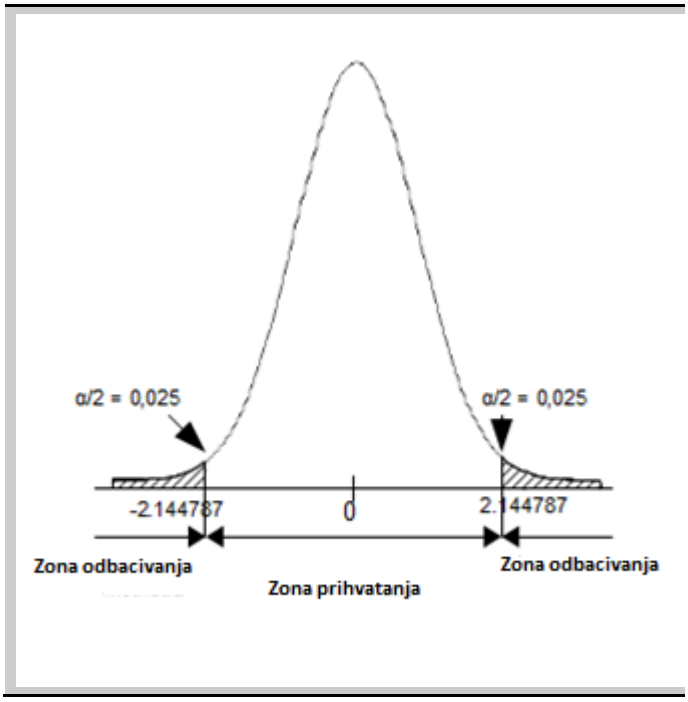
Statističku vrednost izračunaćemo pomoću već poznate komande *t.test(gcs,mu=5, conf.level = 0.95),* gde je *gcs* vektor vrednosti, *mu=5* govori da testiramo nultu hipotezu o jednakosti stvarne srednje vrednosti sa vrednošći 5, a *conf.level = 0.95* određuje vrednost verovatnoće. Rezultat

data: gcs
t = -0.9014, df = 26, p-value = 0.3757
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
4.635511 5.142267

govori da je t test dao vrednost -0,9014.

[25] Ne možemo da koristimo normalnu raspodelu kad ne poznajemo stvarnu varijansu populacije. Zato je tražimo pomoću standardne devijacije i tada moramo da koristimo Studentovu t -raspodelu.

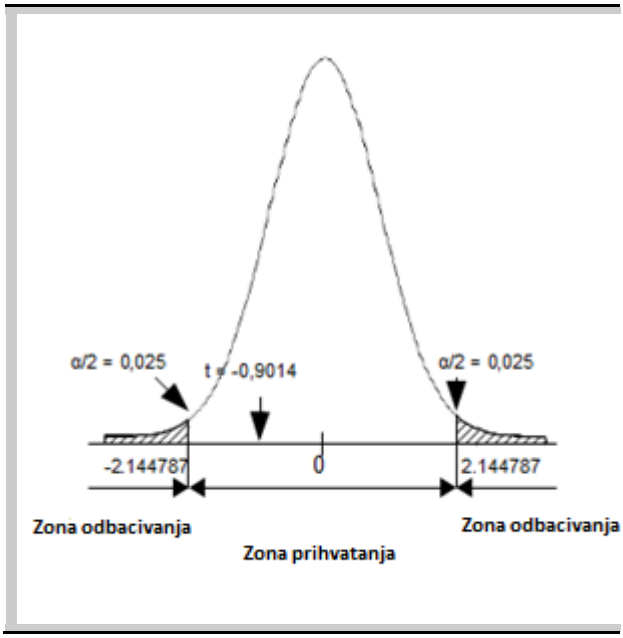
Slika 2: Odlučujuće pravilo o prihvatanju ili odbacivanju nulte hipoteze



Statistička odluka

Na osnovu primene odlučujućeg pravila, prihvatamo H_0 onda kada je rezultat statističkog testa između kritičnih vrednosti, odnosno između brojeva -2.144787 i 2.144787. Budući da njena vrednost leži stvarno među njima (Slika 3), možemo da prihvatimo nultu hipotezu da je stvarna srednja vrednost GCS jednaka vrednosti 5.

Slika 3: Postavljanje rezultata t testa prema kritičkim vrednostima t raspodele



Stručna odluka

Zaključujemo da je stvarna srednja vrednost GCS kod pacijenata primljenih na odeljenje posle teške povrede mozga u nivou vrednosti 5.

Kod detaljnije studije komande *t.test()* vidimo da izračunati interval poverenja *95 percent confidence interval: 4.635511 5.142267* ima vrednost 5. To znači, samo da podsetimo, koliko su usko povezane koncepcije otkrivanja i testiranja hipoteza.

(b) Jednostrani test

Prethodni test nazivamo **obostran**, zato što je verovatnoća odbacivanja rastavljena na obe strane raspodele rezultata statističkog testa. Ako je zona odbacivanja samo na jednoj strani raspodele, onda govorimo o **jednostranom** testu. Na istom primeru kao i u prethodnom slučaju, pokazaćemo kako da napravimo ovakav jednostran test.

Primer 2: Zadavanje

U studiji o teškim povredama mozga, na odeljenju su dobijeni podaci da su pacijenti primljeni sa ovom dijagnozom imali sledeći nivo svesti, meren Glazgovskom skalom (GCS):
5, 4, 4, 5, 6, 5, 4, 5, 5, 6, 4, 5, 5, 4, 4, 5, 6, 5, 5, 5, 5, 5, 5, 5, 4, 5, 6.
Lekara je zanimalo da li je moguće reći da su pacijenti primani na odeljenje sa prosečnim GCS manjim od 5.

Postupak će biti isti, samo umesto pitanja da li je stvarna vrednost jednaka vrednosti 5, pitaćemo da li je ona manja od 5.

Podaci: isti kao u prethodnom primeru.
Pretpostavke: iste kao u prethodnom primeru.
***Hipoteze*:** potvrđivaćemo da je stvarna srednja vrednost manja od 5. Zato nulta hipoteza mora da pretpostavlja da je stvarna srednja vrednost jednaka ili veća od 5.

$H_O: \mu \geq 5$ $H_A: \mu < 5$

Nejednačina veće ili jednako u nultoj hipotezi predstavlja, zapravo, bezbroj hipoteza.

Mi ćemo testirati samo deo koji izražava jednačinu, zato što je moguće dokazati da odluka važi za bilo koju proizvoljnu hipotezu iz datog skupa.

Statistički test: ista kao u prethodnom primeru.
Raspodela rezultata statističkog testa: ista kao u prethodnom primeru.
***Odlučujuće pravilo*:** Ostaćemo na nivou verovatnoće 0,05, isto kao u prethodnom primeru. Kada ćemo, međutim, zaključiti da prihvatamo, odnosno odbacujemo nultu hipotezu? Onda, kada je reč o kritičnoj vrednosti levo od sredine, dakle u smeru gde su jako mali brojevi. Zona prihvatanja biće desno od kritične vrednosti, a zona odbacivanja levo (slika 4). Odgovarajuću kritičnu vrednost izračunaćemo komandom t podele > *qt(0.05, 26)*, sa rezultatom -1.705618.
Izračunavanje statističkog testa:

```
> t.test(gcs, alternative = c("less"), mu = 5, conf.level = 0.05)
        One Sample t-test
data:  gcs
t = -0.9014, df = 26, p-value = 0.1878
alternative hypothesis: true mean is less than 5
```

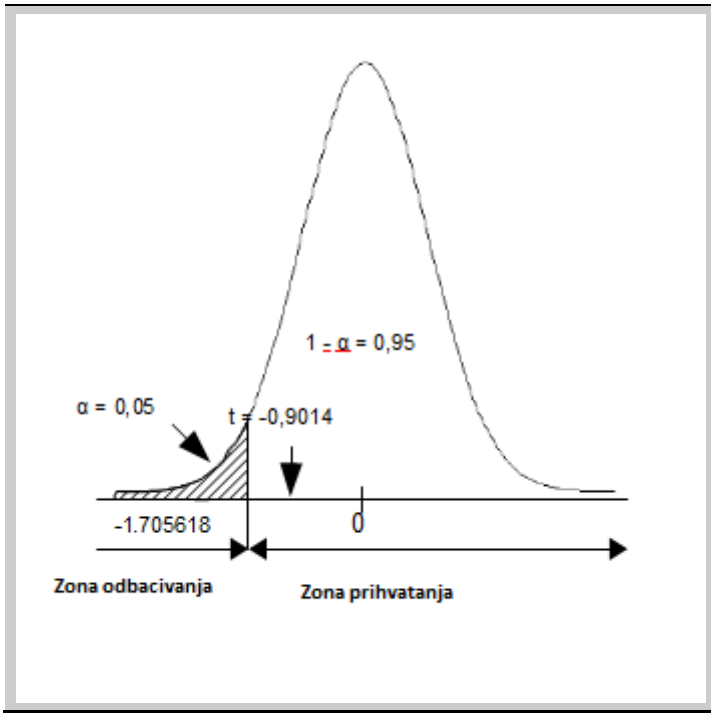
Obratite pažnju da smo u komandi zadali specifikaciju alternativnog testa *alternative = c("less")*, koja govori da potvrđujemo alternativnu hipotezu da je srednja vrednost manja od 5. Dobijeni rezultat ima vrednost -0,9014.

***Statistička odluka*:** Ne možemo da odbacimo nultu hipotezu, budući da je izračunata vrednost -0,9014 veća od kritične vrednosti -1,705618 ili -0,9014 > -1,705618.

***Stručna odluka*:** Stvaran prosek GCS kod pacijenata primljenih na odeljenje može da bude jednak ili veći od 5, na nivou značaja 0,05.

Ovim smo završili testiranje hipoteza o jednoj aritmetičkoj sredini. U sledećem delu ćemo se koncentrisati na potvrđivanje razlika između dve aritmetičke sredine. Ovo je veoma često korišćena procedura, ali ne i svaki put ispravno ni primenjena, niti su rezultati uvek ispravno interpretirani.

Slika 4: Ilustracija zone prihvatanja i odbacivanja nulte hipoteze za srednju vrednost



2. **Potvrđivanje hipoteze za razliku aritmetičkih sredina**

Najčešće želimo da pokažemo da se dve aritmetičke sredine različite. Potvrđujemo da li je neka aktivnost dovela ili nije dovela do rezultata u formi razlike aritmetičkih sredina merenja promenljive. U ovim slučajevima potvrđujemo neku od tri nulte hipoteze:

H_O: $\mu_1 - \mu_2 = 0$ H_A: $\mu_1 - \mu_2 \neq 0$ ili u formatu H_O: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$

H_O: $\mu_1 - \mu_2 \geq 0$ H_A: $\mu_1 - \mu_2 < 0$ ili u formatu H_O: $\mu_1 \geq \mu_2$ H_A: $\mu_1 < \mu_2$

H_O: $\mu_1 - \mu_2 \leq 0$ H_A: $\mu_1 - \mu_2 > 0$ ili u formatu H_O: $\mu_1 \leq \mu_2$ H_A: $\mu_1 > \mu_2$

Dok je prvi primer definicije H_0 obostrano potvrđivanje hipoteza, druga dva primera predstavljaju jednostrane testove. Njihovo korišćenje ilustrovaćemo na primeru (primer 3).

Primer 3: Zadavanje

U studiji o teškim povredama mozga, na odeljenjima dve bolnice upoređivana je dužina

boravka bolesnika u danima na jedinici intenzivne nege, sa ovim rezultatima. Bolnica A za praćeni period primila je 16 bolesnika sa teškom povredom mozga, a broj dana hospitalizacije je bio:
6, 5, 6, 5, 8, 14, 17, 6, 10, 10, 3, 4, 13, 4, 11, 17.
Bolnica B u praćenom periodu je primila 22 bolesnika sa teškom povredom mozga, sa sledećom dužinom hospitalizacije:
44, 8, 12, 24, 5, 20, 3, 20, 6, 22, 4, 3, 91, 6, 4, 4, 13, 9, 18, 10, 1, 40.
Manadžment obe bolnice je zanimalo da li je na osnovu navedenih merenja moguće reći da postoji razlika između dužine hospitalizacije u navedenim bolnicama.

Rešenje zadatka ćemo prikazati prema navedenoj šemi potvrđivanja hipoteze.

***Podaci**:* Podaci dolaze iz dve ustanove, u bolnici A je bilo u praćenom periodu hospitalizovano 16 bolesnika sa teškom povredom mozga, a u bolnici B ih je bilo 22.

Pretpostavke: Pretpostavljamo da je dužina hospitalizacije normalno raspodeljena u populaciji, zatim da su oba uzorka bila uzajamno nezavisna. Raspodelu svake populacije utvrđujemo na osnovu standardnih devijacija uzoraka.

Hipoteze: Menadžement se pita da li je postoji razlika u prosečnoj dužini hospitalizacije među ove dve ustanove. Zato nulta hipoteza mora da pretpostavlja da ovaka razlika ne postoji. Formalno ćemo ti zapisati kao:

$$H_0: \mu_1 - \mu_2 = 0 \quad H_A: \quad \mu_1 - \mu_2 \neq 0$$

ili na drugi način

$$H_0: \mu_1 = \mu_2 \qquad H_A: \quad \mu_1 \neq \mu_2$$

Obe varijante su po značaju jednake, samo ćemo stvarnost jednakosti aritmetičkih sredina jednom izraziti kao razliku koja je jednaka nuli, a drugi put ćemo koristiti znak jednakosti između aritmetičkih sredina obe populacije.

Statistički test: biće to t raspodela sa zajedničkom disperzijom dobijenom na osnovu standardnih devijacija oba uzorka.

***Raspodela rezultata testa*:** Ako je nulta hipoteza tačna, onda je raspodela prema Studentovoj t raspodeli, sa izračunatim brojem stepena slobode 23,939.

Odlučujuće pravilo: Izračunamo kritične vrednosti > *qt(0.01, 23.939)* sa rezultatom -2.49261. Pri nivou verovatnoće 0,01 kritička vrednost za odluku iznosiće t = ±2,49261. Nultu hipotezu moramo da odbacimo ako izračunata t vrednost bude mimo intervala poverenja, sačinjenog od od granica - 2,49261 do +2,49261.

***Izračunavanje statističkog testa**:* Opet ćemo koristiti komandu *t.test()* sa sledećim parametrima: > *t.test(Hosp_A, Hosp_B, conf.level = 0.01)*, gde ćemo zadati oba proveravana uzorka i nivo spoverenja 0,01. Rezultat

data: Hosp_A and Hosp_B
t = -1.7899, df = 23.939, p-value = 0.08612

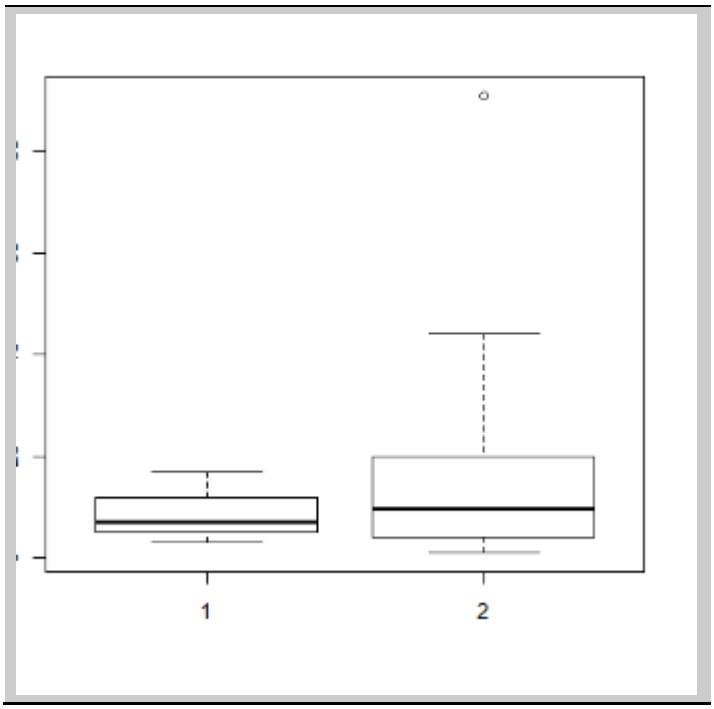
obrazložiće ne samo statistički test, nego i broj stepena slobode (nije potrebno izračunavati ga posebno).

***Statistička odluka*:** Rezultat nas primorava da prihvatimo nultu hipotezu o neprisutnosti razlike među aritmetičkim sredinama obe populacije, budući da -2.49261 < -1.7899 < 2.49261.

Stručna odluka: Menadžment može da zaključi da, uz navedenu verovatnoćom i na osnovu podataka, nije utvrđena ispitivana razlika između prosečnog vremena hospitalizacije u obe ustanove. U posmatranom primeru nismo uspeli da ignorišemo nultu hipotezu, što znači da postoji šansa da prilikom proširenja posmatranja ova razlika može da se pojavi.

Relultat možemo da potvrdimo i pomoću oslikavanja u boksplotu (slika 5).

Slika 5: Boksplot upoređivaja dužine hospitalizacije u obe ustanove pomoću > boxplot(Hosp_A, Hosp_B)



Iz slike vidimo da između obe aritmetičke sredine nema razlike, međutim u drugoj bolnici je disperzija podataka veća nego u prvoj.

3. **Parno upoređivanje**

U istraživanju delotvornosti lekova ili eksperimentalnih procedura često radimo sa nepoznatim uzorcima. Cilj ovakvih upoređivanja je minimalizovanje eksternog izvora disperzije približavanjem sličnosti parova za što više promenljivih. Takođe kod svakog pojedinca radimo dva merenja, većinom prvo pre intervencije ili davanja lekova, a drugo sledi posle intervencije ili posle uzimanja leka. Često se dešava da material, koji treba da se analizira, podeli na dve polovine i svaka se meri drugom metodom. Nekada se parovi naprave na osnovu odgovarajuće karakteristike, npr. ljudi istog uzrasta i pola. Postupak analize je umeren na modifikovanje tako što se, umesto analize svakog posmatranja, analiziraju razlike u parovima. Demonstriraćemo ovaj postupak na primeru gde se potvrđivao učinak dijete na visinu sistolnog krvnog pritiska kod 10 pojedinaca. Pritisak je bio meren pre početka dijete i posle 20 dana nakon započinjanja iste. Rezultati merenja nalaze se u sledećoj tabeli.

Primer 4: Zadavanje primera za parno upoređivanje

Sistolni krvni pritisak kod 12 pojedinaca pre i posle dijete

Pojedinac	**Sistolni pritisak**		**Razlika**
	Pre (X1)	**Posle (X2)**	
1	125	120	-5
2	150	145	-5
3	140	120	-20
4	160	140	-20
5	135	150	15
6	145	125	-20
7	130	160	30
8	160	125	-35
9	155	135	-20
10	140	120	-20
11	135	155	10
12	135	150	15
13	145	125	-20

Podaci: U tabeli je 12 merenja u parovima, pre i posle intervencije. Poslednja kolona predstavlja razlike među vrednostima u paru.

Pretpostavke: Posmatrane razlike predstavljaju slučajan uzorak iz populacije sa normalnom raspodelom i logično je da je neophodno imati isti broj posmatranja kod obe grupe. U slučaju da to nije tako, moramo se zadovoljiti samo upoređivanjem srednjih vrednosti – aritmetičkih sredina, kao u primeru 3.

Hipoteze: Pretpostavljamo da ćemo uspeti da dokažemo da je dijeta učinkovita, a između dve aritmetičke sredine postoji razlika. Zato nultu hipotezu formulišemo u smislu nepostojanja razlike između aritmetičkih sredina ili je razlika veća od 0. Alternativna hipoteza pretpostavlja mogućnost da je razlika manja od 0. Ako označimo razliku između aritmetičkih sredina kao $d = \mu_2 - \mu_1$, onda možemo da formulišemo hipotezu na sledeći način:

$H_0: \mu_2 - \mu_1 \geq 0$ odnosno $H_0: \mu_2 \geq \mu_1$

$H_A: \mu_2 - \mu_1 < 0$ odnosno $H_0: \mu_2 < \mu_1$

Statistički test: Dok je nulta hipoteza tačna, t test je raspodeljen sa n – 1, to jest 11 stepeni slobode. U prošlom slučaju imali smo dve disperzije, dok sada, kada radimo sa istim pojedincem, imamo jednu zajedničku disperziju.

Odlučujuće pravilo: Prema konstrukciji nulte hipoteze koristićemo jednostran test na nivou 0,01, koji govori o odbacivanju nulte hipoteze ako je rezultat testa manji od -2,718079 (> *qt(0.01, 11))*.

Izračunata vrednost testa: 1,1703

```
> t.test(X1, X2,  alternative = c("less"), paired=T, conf.level = 0.01)
        Paired t-test
data:  X1 and X2
t = 1.1703, df = 12, p-value = 0.8677
alternative hypothesis: true difference in means is less than 0
```

Obratite pažnju da smo koristili specifikaciju *paired=T* za određivanje načina izračunavanja pomoću parnog t-testa.

***Statistička odluka*:** Prihvatiti nultu hipotezu o nepostojanju razlike između parova, budući da je -2,718079 < 1.1703.

Stručna odluka: Do sada se nije potvrdio učinak proveravane dijete. Problem može da bude i mali broj posmatranja.

Parna posmatranja nisu uvek povoljna. Jedan od nedostataka je gubljenje stepena slobode, a time i potreba za većim uzorkom. To može da se poveća troškove realizacije istraživanja.

Greške u procesu potvrđivanja hipoteza

Posle savladavanja postupka potvrđivanja hipoteza ukazaćemo na neke moguće posledice ovog procesa. Jedna od njih je pogrešan zaključak, koji ne napravi istraživač, već je posledica situacije koja je nastala u procesu uzorkovanja. Konstatujemo da u procesu zaključivanja može da se desi greška prouzrokovana karakterom verovatnoće procesa. Dok bacamo kocku i 10 puta uzastopno padne na 6, može se pomisliti da je kocka nameštena i da smo dobili pogrešan rezultat. Ali to ne mora da bude tačno, i sa normalnom kockom može da se desi da će više puta uzastopno pasti isti taj broj. To je slučajnost. Prilikom većeg broja bacanja može se pre li kasnije sve ispraviti, a učestalost svih brojeva može biti jednaka. To je problem i sa malim uzorcima. To nam ukazuje na činjenicu da statističar mora da zna da se nosi i sa ovakvom situacijom. Nazivamo je **greškom tipa I**, to jest onda kada odbacimo pravila nulte hipoteze. U slučaju bacanja kocke, nulta hipoteza pretpostavlja da svi brojevi na kocki imaju istu verovatnoću. Kad, međutim, 10 puta padne na šest, pogrešno zaključimo da to nije tačno i da broj šest ima veću verovatnoću da se pojavi nego ostali brojevi. Ovakav zaključak (kod nenameštene kocke) je pogrešan, a to je greška tipa I, odnosno odbacujemo tačnost nulte hipoteze. Verovatnoća da se napravi ovakva greška je jednaka vrednosti α, što je ista alfa vrednost kakvu smo koristili kod otkrića stvarne vrednosti kod populacije. Zato kad je proizvođač prilikom provere leka na 100 pacijenata saznao da ovaj smanjuje temperaturu za jedan stepen za 30 minuta sa standardnom devijacijom tri minuta, znamo da izračunamo da će stvarna prosečna vrednost biti u rasponu 29,4 minuta do 30,6 minuta. Kod dva pacijenata smo, međutim, saznali, da se period učinka leka produžio na 45 minuta i zato smo zaključili da informacija koju je dao proizvođač nije tačna. Time smo u naš zaključak uveli grešku tipa I, odnosno odbacili smo tačnu hipotezu.

Šta je onda greška tipa II? Ona nastaje onda, kada bismo bacali lažno nameštenu kocku, ali rezultati bacanja na to ne bi ukazivali (tome se ne bi obradovali ni lažan igrači). Statistički rečeno, prihvatili smo lažnu, netačnu nultu hipotezu, da svako bacanje ima jednako verovatnoću za sve brojeve. Ovako smo izgubili mogućnost da otkrijemo lažnu kocku, a isto tako i lažnu hipotezu. Pojava ove greške ima uticaj na više faktora. U prvom redu je veličina uzorka. Ako bismo, u slučaju kocke, bacali češće od 10 puta, onda bi se sigurno pokazalo da je neki od brojeva preferiran, a kocka je lažna, kao i hipoteza o jednakosti brojeva. Isto tako bitna je i veličina razlike koju želimo da odredimo: što manja razlika, to je potreban veći uzorak i obrnuto.

Snaga testa

Greška tipa I, dakle odbacivanje tačne hipoteze, ukazala nam je na slabost procedure testiranja hipoteza. Kada pogledamo na ovu slabost optimističnim pogledom, onda je njena suprotnost snaga testa da otkrije stvarnu razliku. Ovo se zove snaga testiranja hipoteze. Ona se meri verovatnoćom, koja je jednaka verovatnoći da dođe do greške tipa II, ali sa suprotnim predznakom. Drugim rečima, snaga testa je verovatnoća odbacivanja H_0 za situacije kada je H_0 netačna. Bitno je

imati u vidu da snaga testa zavisi od veličine uzorka i stvarne vrednosti parametra, kako smo o tome već ranije govorili.

Iako računanje snage testa prevazilazi mogućnosti i potrebe ovog teksta, pokazaćemo da R pruža komande koje će omogućavaju takvo izračunavanje. Za komandu *t.test()*, sa kojom smo se već upoznali, to je komanda *power.t.test()*. Zadavanjem raznih parametara možemo da otkrijemo snagu testa u raznim situacijama.

Ako ne želimo da izračunavamo snagu testa za pojedina zadavanja, internet nudi brojne grafikone, iz kojih lako možemo da izvedemo vrednost, koja odgovara datoj situaciji. Ovde dolazimo do sledeće mogućnosti praktične primene ovog postupka, a to je procena veličine uzorka. Kad ponavljamo postupak iz primera, onda lako dolazimo do toga da za određenu granicu više nije potrebno povećavati uzorak, zato što snaga testa ostaje na 100 procenata. Zato svako bespotrebno merenje može da nam znači trud ili novac koji nisu morali da budu utrošeni, jer nam ne donose ništa novo, nikakvu veću tačnost.

Primer 5: Snaga komande t.test()

Na osnovu podataka o glikemiji, pokazaćemo izračunavanje snage t.testa

```
> gly
 [1] 5.3 6.1 5.8 6.2 6.4 5.9 6.7 7.1 6.3 6.6
7.4 6.6 5.7 6.4 5.5
> power.t.test(n=length(gly), delta = 1, sd
= sd(gly), sig.level = 0.05)

     Two-sample t test power calculation

              n = 15
          delta = 1
             sd = 0.5765249
      sig.level = 0.05
          power = 0.9956279
    alternative = two.sided

> power.t.test(100, delta = 1, sd = 0.577,
sig.level = 0.05)

     Two-sample t test power calculation

              n = 100
          delta = 1
             sd = 0.577
      sig.level = 0.05
          power = 1
    alternative = two.sided

 NOTE: n is number in *each* group

> power.t.test(5, delta = 1, sd = 0.577,
```

Podatke o glikemiji imamo iz prethodnih primera

komandom power.t.test() zadajemo broj merenja u uzorku veličine n, delta = 1 govori da želimo da testiramo stvarnu razliku između aritmetičkih sredina; zadajemo takođe i standardnu devijaciju i nivo verovatnoće sig.level

Rezultat je visoka verovatnoća da greška tipa I neće biti dostignuta

Probamo šta će se desiti kada veličinu uzorka povećamo na 100, pri zadržavanju ostalih parametara

Pokazalo se da je snaga testa dostigla 1,

sig.level = 0.05) Two-sample t test power calculation n = 5 delta = 1 sd = 0.577 sig.level = 0.05 power = 0.6713634 alternative = two.sided NOTE: n is number in *each* group	dakle maksimum. Drugim rečima, test ima maksimalnu snagu da spreči grešku tipa I Prilikom smanjenja veličine uzorka na 5, snaga testa se će se značajno sniziti na 0.67, dakle na 67%

p vrednost

U zaključku ovog poglavlja smatramo da je bitno ukratko govoriti o koncepciji *p* vrednosti, ili tzv. „pe“ i njenoj ulozi prilikom interpretacije značajnosti ili signifikantnosti rezultata testa. Standardno se navodi da je npr. razlika dva merenja značajna, sa $p < 0,05$. Pokušaćemo da odgonetnemo ovu izreku u svetlu do sada neizrečenog. Jasno nam je da testiranje hipoteza dovodi do potvrđivanja hipoteza o prisustvu ili odsustvu razlike, npr. između aritmetičkih sredina. Ovu činjenicu onda izražavamo prihvatanjem ili odbacivanjem nulte hipoteze, a u drugom slučaju prihvatanjem alternativne hipoteze. Uvek odredimo na kom nivou želimo da potvrđujemo H_0, a ovu verovatnoću onda koristimo pri izračunavanju kritične vrednosti. Moguć je i drugi način i to takav, da kompjuter sam izračunava kada bi bilo dobro odbaciti nultu hipotezu, a prihvatti alternativnu. A to je to naše *p*. Dakle, ako vidimo u rezultatu $p = 0,123$, jasno nam je da će kritična vrednost za odbacivanje H_0 biti na nivou 12%, šta se iz pogleda uobičajenog postupka interpretacije rezultata u biološkim naukama smatra za neprihvatljivo i zato se prednost daje prihvatanju H_0. Pogledajte rezultate naših primera, tamo gde smo prihvatili nultu hipotezu – tamo je i komanda *t.test()* ispisala vrednost p, označenu kao *p-value* veću od 0,5. Ovu vrednost moramo i da tačno interpretiramo. Ako na njenoj osnovi odbacujemo H_0, mora da nam bude jasno da to još ne znači da će sve razlike biti različite, to samo znači da radimo sa određenom greškom, a to je verovatnoća njene pojave. Kakva je to greška? To je greška iz odbacivanja tačne nulte hipoteze, dakle tipa I. Kakav zaključak možemo da damo na osnovu rezultata utvrđivanja razlika dve aritmetičke sredine, kada program izračuna $p = 0,004$? U prvom redu moramo da budemo svesni da u biološkim naukama radimo sa uobičajenom vrovatnoćom i to 0,05, 0,01, 0,001 i 0,0001. Sve izračunate verovatnoće uporedićemo sa ovim vrednostima tako da zapisana vrednost verovatnoće iz skupa uobičajenih bude veća nego izračunata. Tako ćemo izračunatu vrednost $p = 0,004$ navoditi kao $p < 0,01$. Nije praksa da se navodi sama izračunata vredenost. Interpretirati je možemo tako da je stvarna razlika aritmetičkih sredina verovatno na nivou 1% ili sa greškom 1%. Ovaj podatak nam je značajan i prilikom shvatanja da kod svakog slučaja razlika mora da da bude jednoznačna. Zato podatak tipa „pokazali smo da je signifikantna razlika između aritmetičkih sredina“ ne bi trebalo da se pojavljuje u stručnoj literaturi (ni kod studenata). Šta sa slučajevima kad je vrednost p veća od 0,05? Ni u takvom slučaju ne navodimo njenu vrednost, već navedemo n.s. kao znak da razlika nije bitna (signifikantna). Moguće su i druge oznake - često se koriste * (zvezdice) za označavanje nivoa verovatnoće, npr. $p<0,05$ se označava *, $p < 0,001$ kao **, $p < 0,0001$, kao *** i tako dalje.

Sažetak

U ovom poglavlju smo naveli postupak pri potvrđivanju hipoteza, koji je deo statističke analize. Ovaj postupak, iako ne ovako detaljno, koristićemo i u narednim poglavljima, ali nećemo navoditi sve korake. Čitalac će ih pretpostavljati iza svake realizovane analize. Naveli smo i koncepciju greške testiranja, a takođe i snagu testa. Iz praktičnih razloga ovde smo svrstali i razmišljanje o tačnoj interpretaciji izračunatog nivoa verovatnoće.

Vežbe

Zadatak 1.

Ispitivač je posmatrao 24-časovni utrošak energije u MJ u grupi mršavih i gojaznih. Interesovalo ga je da li može da zaključi da se prosečne vrednosti utroška energije razlikuju. (Podaci su u *IswR*, uzorak *energy*).

Zadatak 2.

Ispitivač je posmatrao energetski unos u KJ kod 11 žena pre i posle menstruacije. Da li možemo da zaključimo, da se prosečan energetski unos razlikovao? Možemo da kažemo da je unos pre menstruacije većo nego posle nje? (Podaci su u *IswR*, uzorak *intake*).

Zadatak 3.

Kod praćenja radnika na izradi kadmijuma upoređivan je celokupni vitalni kapacitet kod dve grupe: 1 – izloženih više od 10 godina, 2 – neizloženih. Da li je moguće zaključiti da je vitalni kapacitet pluća u prvoj grupi bio veći nego u drugoj? (Podaci su u *IswR*, uzorak *vitcap*).

SEDMO POGLAVLJE

Analiza varijanse

Sadržaj poglavlja

Cilj poglavlja

U prethodnim poglavljima informisali smo se o načinima utvrđivanja razlike između dve aritmetičke sredine. Sigurno ste mogli da predstavite sebi situaciju kad treba da međusobno uporedimo tri ili više aritmetičkih sredina. Jedan od mogućih odgovora bi bio da ponovimo test upoređivanja dve aritmetičke sredine procedurom *t.test()*. Za tri aritmetičke sredine bi morali da uradimo tri upoređivanja: prvu aritmetičku sredinu sa drugom, drugu sa trećom i prvu sa trećom. To nije tako komplikovano. Međutim, šta u situaciji kad treba da upoređujemo više aritmetičkih sredina, recimo šest? U ovakvoj situaciji broj kombinacija biće mnogo veći, čak 30. To bi već zahtevalo jako veliki napor. Pritom bi nam se javila još jedna prepreka, a to je narastuća greška. Zamislite da stvarno napravimo šest uzoraka i krenemo da ih upoređujemo pri prihvatanju verovatnoće greške na nivou 0,05. Pri dvadeset ponavljanja, greška se između sebe množi i na kraju bismo dobili grešku 0,64, odnosno verovatnoća da se odbaci tačna hipoteza (greška tipa I) bi bila 0,64, tj. 64%. Priznajte da veran korisnik statistike ovo sebi ne može da dozvoli. Osim toga, bilo bi jako teško zanemariti uticaj načina izbora uzorka. Rešenje pruža postupak koji zovemo analizom disperzije ANOVA (Analysis of Variance), a za razliku od prethodnog postupka, analiziraćemo razlike u disperziji između pojedinih uzoraka. U ovom poglavlju pokazaćemo kako ANOVA radi, kako možemo da interpretiramo rezultate i kako da je koristimo u praksi.

Tabela 1: Ciljevi poglavlja

- Objasniti osnovne postupke prilikom analize varijanse ANOVA
- Interpretacija rezultata

Postupak

Prvi korak čine podaci koje želimo da analiziramo. Zadajemo u primeru 1.

Primer 24: Zadavanje primera i pravljenje uzorka podataka

Zadavanje:
Ispitivač se interesovao za dužinu hospitalizacije (DH-broj dana provedenih u bolnici) operisanih bolesnika na odeljenju u tri bolnice. Želimo da znamo da li postoji razlika između bolnica.
Postupak:
Svaku bolnicu je označio sa velikim slovom od A do C.
Vektor označene bolnice A: 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A'
Vektor vrednosti DH bolnice A: 9, 3, 14, 13, 14, 5, 3, 5, 11, 3, 8, 6, 5, 8
Vektor označene bolnice B: 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B'
Vektor vrednosti DH bolnice B: 4, 4, 7, 7, 7, 4, 4, 3, 3, 4
Vektor označene bolnice C: 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C'
Vektor vrednosti DH bolnice C: 3, 6, 4, 9, 6, 4, 5, 5, 6, 3, 5
Napravio je dva vektora:
U vektor ABC je stavio podatke DH, u abc je stavio vektor označenih bolesnika:
> ABC<-c(9, 3, 14, 13, 14, 5, 3, 5, 11, 3, 8, 6, 5, 8, 4, 4, 7, 7, 7, 4, 4, 3, 3, 4, 3, 6, 4, 9, 6, 4, 5, 5, 6, 3, 5)
> abc<-c('A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C')
Za razaznavanje označenih podataka potrebno je konvertovati vektor *abc* u vektor označen

pomoću komande *label():*

```
> fact_abc <- factor(abc)
> fact_abc
 [1] A A A A A A A A A A A A A A B B B B B B B B B B B C C C C C C C C C C C C
Levels: A B C
```

Za spajanje podataka i njihovo označavanje koristi se šablon podataka, dakle *data frame*, a ovo spajanje se dobije komandom:

```
> bolnice<- data.frame(ABC, fact_abc)
> bolnice
   ABC fact_abc
1    9        A
2    3        A ......
```

Sada u okviru bolnica imamo promenljivu *ABC* sa vrednostima broja dana hospitalizacije i *fact_abc* sa oznakama. Pošto smo pripremu podataka završili, možemo još da prekontrolišemo da li je sve u redu, a koristićemo za to komandu *summary().*

```
> summary(bolnice)
      ABC      fact_abc
 Min.   : 3   A:14
 1st Qu.: 4   B:10
 Median : 5   C:11
 Mean   : 6
 3rd Qu.: 7
 Max.   :14
```

Ovaj način zadavanja ulaznih podataka za više grupa koristićemo i u narednim slučajevima. Obratite pažnju da je svaka bolnica primila različit broj bolesnika: A – 14, B – 10 i C – 11. I iz ovog vidite da nije moguće za zadavanje podataka koristiti običnu matricu.

Podaci: predstavljaju uzorak iz tri bolnice, a u svakoj je bio zabeležen različit broj slučajeva. Karakteristike prikupljenih podataka su u tabeli 1.

Tabela 2: Podaci iz primera 1

Bolnica A	Bolnica B	Bolnica C
9	4	3
3	4	6
14	7	4
13	7	9
14	7	6
5	4	4
3	4	5
5	3	5
11	3	6
3	4	3
8		5
6		
5		

	8		
Broj	14	10	11
Zajedno *sum()*	107	47	56
Prosek (aritmetička sredina)	7,6	4,7	5,1
Za sve kolone	Broj =35	Zajedno dana = 210	

Pretpostavke: Odabrali smo tri grupe podataka slučajnim izborom iz populacije sa normalnom raspodelom. Moraju da ispunjavaju sledeće pretpostavke:

- Populacija je nezavisna (drugačije rečeno: ne utiču jedna na drugu);
- Svaka od njih ima normalnu raspodelu;
- Disperzije su jednake.

Hipoteze: Isto kao u slučaju prilikom potvrđivanja hipoteza o razlici aritmetičkih sredina i u ovom slučaju potvrđujemo hipotezu da su aritmetičke sredine u populaciji jednake.
Nultu hipotezu za korišćenje ANOVA procedure onda formulišemo uopšteno kao:
H_0: $\mu_1 = \mu_2 = . . . = \mu_k$, gde je μ_1 aritmetička sredina prve populacije, a broj svih aritmetičkih sredina ne znamo, pa zato koristimo indeks *k,* iza koga možemo da stavimo stvaran broj uzoraka. U našem slučaju, dakle, nulta hipoteza će glasiti:

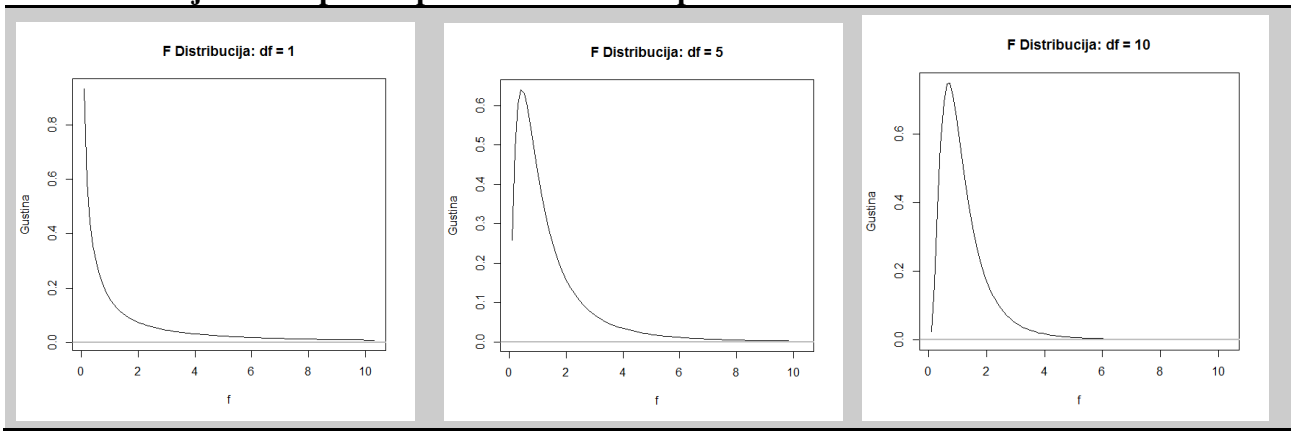
$$H_0: \mu_1 = \mu_2 = \mu_3$$

Šta ćemo sa alternativnom hipotezom? Na prvi pogled može da nam se čini da bi to trebalo da bude hipoteza o nejednakosti svih aritmetičkih sredina. Zapravo, to nije tako; morali bi da imamo jako striktno pravilo, ukoliko bismo želeli da imamo sve aritmetičke sredine različite, a u većini slučajeva to nam nije ni cilj. Tako da je alternativna hipoteza pretpostavka da najmanje dve aritmetičke sredine nisu jednake. Dakle, napisaćemo
H_A: *Najmanje dve od aritmetičkih sredina* $\mu_1, \mu_2, , , \ldots, \mu_k$ *je su različite.*
U našem slučaju to će biti za $k = 3$.
H_A: *Najmanje dve od aritmetičkih sredina* μ_1, μ_2, μ_3 *su različite.*

Za razliku od testa razlike dve aritmetičke sredine, u ANOVA je moguća samo jedna varijanta, a to je razlika aritmetičkih sredina; ni veća, ni manja se ne testira. To važi i za rezultat, jedina informacija koju će ANOVA dati je informacija da postoje najmanje dve aritmetičke sredine koje su različite. Za objašnjenje koje su to dve aritmetičke sredine, nephodno je koristiti druge testove. Nasuprot tome, rezultat ANOVA je jako bitan za predstojeći rad istraživača – statističara.

Statistički test. Od početka poglavlja naglašavamo da je ANOVA koncipirana na potvrđivanju razlika između varijansi. Prisetimo se da je t raspodela bila korišćena onda kada smo se interesovali za razlike u srednjim vrednostima i otkrivali smo disperziju kod populacije pomoću standardne devijacije uzorka. Iz razloga statističke korektnosti (objašnjenje bi tražilo korišćenje matematičkog aparata i dublje znanje matematike) neophodno je koristiti druge raspodele, odnosno F raspodelu.

Slika 1: Slučajevi F raspodele prilikom raznih stepena slobode



Ova raspodela nije simetrična oko aritmetičke sredine, ali jeste određena sa dva stepena slobode. Sve F podele počinju nulom i zakrivljene su na desno. Stepen zakrivljenosti i njena visina su dati sa dva stepena slobode, datim $k - 1$ i $n - k$. Dok prvo govori o broju uzoraka minus jedan, drugo govori o broju svih merenja kod svih uzoraka minus broj uzoraka.

***Izračunavanje statističkog testa*:** radi se u tri koraka. Izračunavanje zbira drugih korenova razlika, izračunavanje njihovih aritmetičkih sredina i formiranje F testa. Uprkos tome što R program izračunava direktno kritične vrednosti i vrednosti testa, navodimo ukratko način izračunavanja, iz razloga objašnjenja postupka. Prvi korak je podela celokupne disperzije istraživane promenljive y (slovom y označićemo sve vrednosti dobijene u uzorku) u delove, koji omogućavaju dobijanje uzroka disperzije. Usled navedenog, saznaćemo koliko disperzije se nalazi u celom uzorku. Ako su podaci jako heterogeni, onda će i ova disperzija biti velika i suprotno. Ovu meru disperzije nazivamo **celokupan zbir kvadrata** i označavamo kao **CZK**. Njega delimo na dva dela: disperzija između grupa, koju ćemo nazvati **zbir kvarata za kolone ZKZK** i drugi, koji **zovemo zbir kvadrata za grešku ZKG**. Onda je **celokupan zbir kvadrata** jednak zbiru kvadrata za kolone i za grešku, skraćeno zapisano kao CZK=ZKZK+ZKG. Postupak izračunavanja pojedinih zbirova je sledeći (s obzirom na obavezu da ne koristimo formule, opis postupaka zahteva više reči):

- **Celokupan ZK, CZK** se računa tako da se od svakog merenja u svakoj koloni oduzme celokupna aritmetička sredina, onda se ove razlike dižu na kvadrat, čime se odstranjuju znakovi, i na kraju se saberu po kolonama, a takođe se saberu sve kolone i dobije se krajnji zbir. Za one koje se ne plaše matematičkog zapisivanja postupak se zapiše kao: CZK = $\Sigma\Sigma\left(x_{ij}-\overline{x}\right)^2$.
- **Zbir kvadrata za kolone, ZKZK** se računa kao razlika aritmetičke sredine kolone i celokupne aritmetičke sredine, on se podigne na kvadrat, sve se sabere i zbir se pomnoži sa brojem kolona. ZKZK = $n_i \Sigma\left(\overline{x}_i-\overline{x}\right)^2$.
- **Zbir kvadrata za grešku ZKG** se izračunava tako da svakoj vrednosti u koloni oduzmemo aritmetičku sredinu kolone, razlika se diže na kvadrat i sabere se po kolonama i na kraju se saberu zbirovi svih kolona. ZKG = $\Sigma\Sigma\left(x_{ij}-\overline{x}_i\right)^2$.

Navedene operacije ilustrovaćemo na našem primeru, ali u R bi bilo jako puno posla, pa za takvu namenu radije koristimo EXCELL, imajući u vidu i da niko u realnom životu neće raditi ove korake (Tabela 1).

Tabela 4: Izračunavanje zbirova kvadrata za rastavljanje disperzije

	Bolnica A			Bolnica B			Bolnica C		
	x_A	$(x_A - avrg_A)^2$	$(x_A - avrg_T)^2$	x_B	$(x_B - avrg_B)^2$	$(x_B - avrg_T)^2$	x_C	$(x_C - avrg_C)^2$	$(x_C - avrg_T)^2$
	9	1.841836735	9	4	0.49	4	3	4.41	9
	3	21.55612245	9	4	0.49	4	6	0.81	0
	14	40.41326531	64	7	5.29	1	4	1.21	4
	13	28.69897959	49	7	5.29	1	9	15.21	9
	14	40.41326531	64	7	5.29	1	6	0.81	0
	5	6.984693878	1	4	0.49	4	4	1.21	4
	3	21.55612245	9	4	0.49	4	5	0.01	1
	5	6.984693878	1	3	2.89	9	5	0.01	1
	11	11.27040816	25	3	2.89	9	6	0.81	0
	3	21.55612245	9	4	0.49	4	3	4.41	9
	8	0.12755102	4				5	0.01	1
	6	2.698979592	0						
	5	6.984693878	1						
	8	0.12755102	4						
Količina	14			10			11		
Zajedno	107	211.2142857	249	47	24.1	41	56	28.91	38
$avrg_j$	7.64	15.08673469		4.7	2.41		5.1	2.6281818	
$(avrg_j - avrg_T)^2$		82.5687474			12.8881			11.369158	
Za sve kolone	Količina	35		Zbir	210		$avrg_T$	6	

Rezultat

CSŠ	328	**SŠMS+SŠC=**	327.82
SŠMS	63.59571429		
SŠC	264.2242857		

Izračunali smo pojedine zbirove kvadrata, a sada moramo da izračunamo samu disperziju. Prva mera će govoriti o disperziji u kolonama - zvaćemo je **prosečan kvadrat u koloni.** Izračunava se kao aritmetička sredina zbira kvadrata za grešku, gde odgovarajući zbir kvadrata podelimo stepenima slobode za datu kolonu. Druga mera će biti **prosečan kvadrat među kolonama,** a izračunaćemo ga kao srednju vrednost ZKZK, odnosno ZKZK podelimo celokupnim brojem stepena slobode.

Ako bi važila H_0 za jednakost atiritmetičkih sredina, trebalo bi da su obe otkrivene disperzije jednake. Ako su jednake, odnosno ako nisu jednake, dobićemo ih odnosom vrednosti prosečnog kvadrata između kolona i u kolonama.

Rezultat se zove **odnos varijansi (OV)**, a u našem slučaju je jednak količniku 31,8 i 8,26, što iznosi 3,85.

Odluka. Za odluku moramo da uporedimo izračunatu vrednost OV sa kritičnom vrednošću F raspodele. Ako OV bude mnogo manja od kritične vrednosti, onda ćemo prihvatiti H_0, a ako bude veća, onda ćemo odbaciti H_0 i prihvatićemo H_A. Kompjuter će nam pomoći da nađemo kritične vrednosti za stepene slobode između kolona = 3 – 1 = 2 i u kolonama 35 – 3 = 32. Za nivo poverenja koristićemo verovatnoću 0,05. Komanda > *qf(0.05, 2, 32, lower.tail = FALSE)* daće nam rezultat 3.294537. Dakle, rezultat testa OV = 3.85 je manji od kritične vrednosti F raspodele za navedene stepene slobode = 3,29. Zato se odlučujemo da odbacimo H_0, gde smo pretpostavljali da su sve aritmetičke sredine iste, a prihvatićemo H_A, sa pretpostavkom da se barem jedna aritmetička sredina iz populacije razlikuje od ostalih.

Stručna odluka: Istraživač može da zaključi da se broj dana provedenih u bolnici razlikuje barem u jednoj od njih. U kojoj, to moramo da saznamo u narednom postupku.

Da ne bismo morali da radimo ovako komplikovana računanja, u R postoji više mogućnosti kako izračunati ANOVA test. Pokazaćemo korišćenje najjednostavnije komande *anova()*, koja

pretpostavlja pripremu podataka pomoću komande *lm(),* (Tabela 4).

Tabela 4: Izračunavanje ANOVA za primer 1

> anova(lm(ABC ~ abc)) # Komanda *lm(ABC ~ abc)* priprema ulazeće podatke za komandu *anova*(). Argument komande *ABC ~ abc* govori da su vrednosti promenljive ABC označene prema pripadanju grupi pomoću promenljive abc.

Analysis of Variance Table

Response: ABC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
abc	*2*	*63.777*	*31.888*	*3.862*	*0.03145*	*
Residuals	*32*	*264.223*	*8.257*			

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

*Komentar: red*r abc se odnosi na disperziju između kolona, red označen kao Residuals se odnosi ZKG. Izračunata vrednost F raspodele i verovatnoća alternativne hipoteze vode do identičnog zaključka do kakvog smo došli u našem izlaganju.

Poređenje višestrukih srednjih vrednosti

Za utvrđivanje koja od srednja vrednost se razlikuje koristićemo **Dankanov test**. On analizira razlike između promenljivih u grupama prilikom analize varijanse. Zasnovan je na različitim principima od ANOVA-e, pa rezultati ne moraju da budu jednaki rezultatima ANOVA-e. Obično se zove i poređenje višestrukih srednjih vrednosti. Njegovo korišćenje pokazaćemo na novom primeru (Primer 2).

Primer 25: Korišćenje Dankanovog testa za višestrukih raspona

Zadavanje:

U četiri naselja slučajno je bilo odabrano 10 pojedinaca i kod njih je meren, između ostalog, i indeks telesne mase BMI. Ispitivači su se interesovali da li postoje razlike između aritmetičkih sredina BMI kod populacije ispitivanih naselja.

Rešenje:

U prvom koraku ćemo zadati podatke u promenljive, koje ćemo nazvati imenima naselja:

```
> Horna <- c(26, 27, 26, 30, 21, 22, 38, 20, 22, 20)
> Dolna <- c(20, 19, 24, 23, 22, 21, 20, 19, 22, 23)
> Vysna <- c(28, 32, 34, 32, 29, 41, 42, 33, 35, 28)
> Nizna <- c(20, 25, 28, 21, 22, 35, 29, 21, 26, 29)
> NASELJA <-c(Horna, Dolna, Vysna, Nizna)
> NASELJA
 [1] 26 27 26 30 21 22 38 20 22 20 20 19 24 23 22 21 20 19 22 23 28 32 34 32 29
[26] 41 42 33 35 28 20 25 28 21 22 35 29 21 26 29
```

Pripremimo vektor označen tako da nam pokazuje pripadanje prema pojedinim naseljima.

```
> naselja = factor(rep(letters[1:4], each = 10))
> naselja
 [1] a a a a a a a a a a b b b b b b b b b b c c c c c c c c c c d d d d d d d d d d
[39] d d
```

Levels: a b c d
> levels(naselja) <- c("Horna", "Dolna", "Vysna", "Nizna")
> obce
[1] Horna Horna Horna Horna Horna Horna Horna Horna Horna Horna Dolna Dolna
[13] Dolna Dolna Dolna Dolna Dolna Dolna Dolna Dolna Vysna Vysna Vysna Vysna
[25] Vysna Vysna Vysna Vysna Vysna Vysna Nizna Nizna Nizna Nizna Nizna Nizna
[37] Nizna Nizna Nizna Nizna
Levels: Horna Dolna Vysna Nizna

Imamo isti broj merenja u svakoj grupi: 10, zato možemo da koristimo komandu *rep(letters[1:4], each = 10)*, koja će napraviti lanac sa prva četiri slova abecede 10 puta. Komandom *factor* iz lanca napravićemo uzorak oznaka. Ovaj postupak napravimo tako da umesto slova imamo nazive i to pomoću komande *levels()*. Imamo pripremljene podatke za puštanje procedure ANOVA. Ova se sastoji iz dva koraka. U prvom koraku pomoću komande *lm*() pripremimo model, a u drugom pustimo samu komandu *anova()*.

> fit = lm(formula = NASELJA ~ naselja)
> anova(fit)
Analysis of Variance Table

Response: NASELJA
Df Sum Sq Mean Sq F value Pr(>F)
*obce 3 770.87 256.958 12.629 8.585e-06 ****
Residuals 36 732.50 20.347

Rezultat omogućava odbacivanje H_0 o nepostojanju razlike između aritmetičkih sredina sa verovatnoćom $p < 0{,}0001$.

Pomoću Dankanovog testa ćemo pokušati da saznamo gde se nalaze ti parametri koji se značajno razlikuju. Ovaj test se, međutim, ne nalazi među standardnim datotekama R, pa je neophodno instalirati novu datoteku, sa nazivom *DTK* (koristiti pritom okolinu R: *Packages; Install Package(s)...; Load Package*). Sada možemo da pustimo komandu *DTK*:

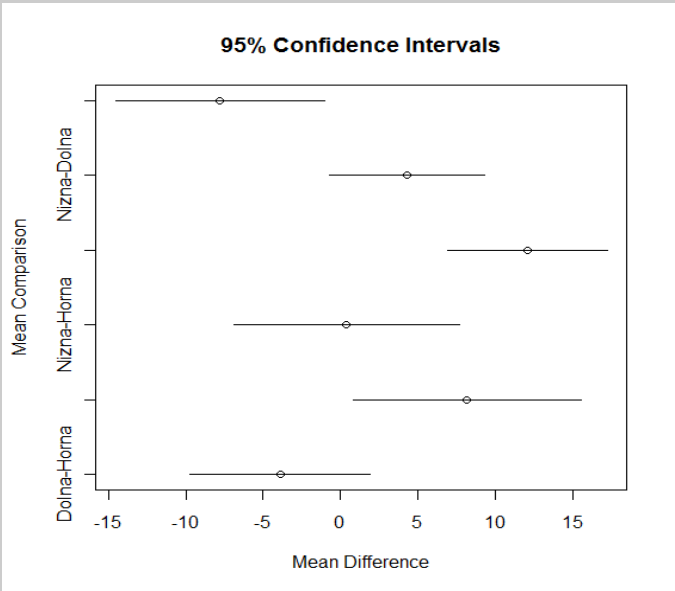
> razlike <- DTK.test(NASELJA,naselja,a=0.05)

Komandu smo spustili sa dve promenljive i sa traženim nivoom značajnosti 0,05.

> razlike
Diff Lower CI Upper CI
Dolna-Horna -3.9 -9.7096778 1.909678
Vysna-Horna 8.2 0.8418498 15.558150
Nizna-Horna 0.4 -6.8693162 7.669316
Vysna-Dolna 12.1 6.9545534 17.245447
Nizna-Dolna 4.3 -0.7175895 9.317589
Nizna-Vysna -7.8 -14.5502757 -1.049724

Rezultat je tabela razlika. Vidimo da neke razlike u rasponu ograničenom donjim i gornjim intervalom poverenja (Lower CI a Upper CI) sadrže nulu i neke ne. Prisetimo se da potvrđujemo H_0 o nepostojanju razlike između aritmetičkih sredina kada je njihova razlika nula. Zato tamo gde se nalazi nula, H_0 je tačna. To su razlike Dolna-Horna, Nizna-Horna i Nizna-Dolna. Ostali intervali ne sadrže nulu, pa ovde možemo da govorimo da postoji stvarna razlika između aritmetičkih sredina na nivou 0,05. To su *Vysna-Horna i Vysna-Dolna.* Rezultat možemo da prikažemo i grafički.

> DTK.plot(razlike)



Bivarijantna ANOVA

U prethodnom odeljku smo analizu varijanse koristili za objašnjenje razlike aritmetičkih sredina kvantitativne promenljive *y* posmatranih na više grupa označenih *x*. Obično se promenljiva *y* naziva **zavisna promenljiva,** a *x* **nezavisna,** zato što se vrednost *y* menja u odnosu na pripadanje određenoj vrednosti *x*. ANOVA se često koristi za proveru da li različiti nivoi nezavisne promenljive x imaju odjek na različite nivoima zavisne promenljive y. U ovakvom slučaju vrednost nazivamo x faktor.

Pretpostavite da farmakološka kompanija istražuje učinak novog leka na nivo holesterola, time da pretpostavlja da taj lek snižava nivo holesterola u zavisnosti od njegove doze. U ovom slučaju doza leka biće faktor. Iz tog razloga zabeležićemo četiri različite doze: 5 mg dnevno, 10 mg dnevno, 15 mg dnevno i 20 mg dnevno. Pretpostavljamo da je kompanija ujedno odlučila da proveri da li lek ima različit efekat ukoliko se daje jednom dnevno ili dva puta. U ovom slučaju javlja se sledeći faktor koji ima dva nivoa, a to je davanje jednom ili dva puta dnevno. Pretpostavimo da želite da to proverite odjednom. Ovakav predlog studije se zove **bivarijantna ANOVA**, koja ispituje zajedničko delovanje dva faktora na prosečni odgovor. Naravno, da bi eksperiment mogao da se napravi kao dve jednostrane ANOVA-e, ali time bi se poništo uticaj zajedničkog delovanja faktora.

U našem primeru imaćemo, dakle, više kombinacija (Tabela 5).

Tabela 5: Predlog eksperimenta i razlaganje prema faktorima

Faktor		
A	**B**	
Davanje	**Jednom dnevno**	**Dva puta dnevno**
● 5 mg	Terapija 1	Terapija 2
● 10 mg	Terapija 3	Terapija 4
● 15 mg	Terapija 5	Terapija 6
● 20 mg	Terapija 7	Terapija 8

Ovaj predlog eksperimenta je svakodnevan u kliničkim studijama i zato se obično govori o **lečenju (treatment)** i u situacijama kada se ne radi o studiji korišćenja lekova. U našem slučaju, dakle, imamo četiri nivoa faktora A i dva nivoa faktora B. Njihovom zajedničkom kombinacijom dobijamo osam raznih vrsta lečenja.

Dvostrana ANOVA daje tri rezultata:

1. Glavni efekat faktora A
2. Glavni efekat faktora B
3. Interakciju A i B: efekat kombinacije faktora A i B.

U jednostranoj ANOVA-i razložili smo celokupnu varijansu na dve komponente CZK = ZKMK + ZKG. CZK predstavlja celokupnu varijansu zavisne promenljive *y*, ZKMK predstavlja varijansu u grupama lečenja, a ZKG je varijansa napravljena greškom. Potvrđivali smo hipotezu da faktor A uzrokuje, odnosno ne uzrokuje razliku u prosecima grupama-lečenja.

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_i$$

Slovom *i* smo označili razne nivoe faktora A. Pri nejednakosti aritmetičkih sredina možemo da odbacimo H_0 i konstatujemo da je uticaj faktora A bio značajan.

Prilikom dodavanja narednog faktora u dvostranoj ANOVA-i, situacija se komplikuje. Moramo da dodamo novi faktor B, a time i njihovo zajedničko delovanje AB. Onda se celokupan zbir kvadrata razloži na više podzbirova:

$$ZKG = ZKA + ZKB + ZKAB + ZKG$$

ZKG predstavlja celokupnu varijansu u promenljivoj *y*, ZKA je zbir kvadrata za faktor A i analogno za faktor B je ZKB. ZKAB je varijansa objašnjena interakcijom faktora A i B. ZKG je varijansa, koja je ostala neobjašnjena i zato je smatrana za grešku. S obzirom na kompleksnost izračunavanja konstatovaćemo samo da je analogno jednostranoj ANOVA-i i zaključićemo da komanda sadržana u R izračunava pojedine koeficijente. Postupak ćemo pokazati na primeru, koji smo opisali u uvodu u ovom delu (Primer 3).

Primer 3: Izračunavanje dvostrane ANOVA-e

Zadavanje:
Prilikom kliničkog ispitivanja leka za snižavanje holesterola u krvi bile su korišćene četiri različite doze aktivne supstance: 5, 10, 15 i 20 mg, a ujedno su bile davane ili jednom dnevno ili u dve doze. Nas zanima da li možemo da zaključimo da je neki od faktora bio delotvoran.

Podaci:

```
> L51 <- c(5.4, 5.8, 6.0, 4.5, 7.9, 8.9, 9.9, 6.8, 8.3, 10.1) #doza 5mg 1x
> L101 <-c(6.2, 5.4, 8.9, 9.9, 6.5, 10.1, 12.5, 9.9, 5.4, 12.3) #doza 10mg 1x
```

```
> L151 <-c(4.5, 6.8, 7.3, 6.7, 4.5, 8.0, 9.0, 5.6, 5.3, 4.5) #doza 15mg 1x
> L201 <-c(5.1, 5.8, 5.4, 6.0, 4.8, 5.3, 6.0, 4.6, 5.0, 4.4) #doza 20mg 1x
>  L52 <- c(6.4, 6.0, 7.0, 4.7, 6.9, 9.1, 9.8, 8.8, 8.1, 6.1) #doza 5mg 2x
> L102 <-c(5.2, 5.2, 9.9, 9.8, 5.5, 11.1, 10.5, 9.8, 6.4, 11.3) #doza 10mg 2x
> L152 <-c(5.5, 6.6, 8.3, 6.9, 5.5, 8.4, 7.0, 5.8, 6.3, 8.1) #doza 15mg 2x
> L202 <-c(4.6, 4.8, 5.6, 6.3, 5.8, 5.2, 6.4, 4.8, 5.1, 4.8) #doza 20mg 2x
> terapija <- c(L51, L101, L151, L201, L52, L102, L152, L202) #jedna promenljiva sa svim vrednostima
> dosage <- gl(4,1,80) # promenljiva označena sa četiri vrste doza
> levels(dosage)<- c("5mg", "10mg", "15mg", "20mg") #dodali smo vrednost doza
> factor(dosage) #oblikovana kao faktor
> admin <- gl(2, 40, 80) # druga promenljiva označavanja administracije doza
> levels(admin)<- c("1x", "2x")
> factor(admin)
```

Izračunavanje ANOVA:

```
> fit <- lm(formula = terapija ~ dosage + admin)
> anova(fit)
Analysis of Variance Table

Response: terapija
          Df Sum Sq Mean Sq F value Pr(>F)
dosage     3  10.28  3.4255  0.7799 0.5088
admin      1   0.21  0.2101  0.0478 0.8275
Residuals 75 329.39  4.3919
```

Rezultat:

Nije se pokazao uticaj ni doze ni davanja, zato prihvatamo nultu hipotezu, da se između faktora nisu pronašle razlike u njihovom delovanju.

Kada nismo našli značajne razlike među delovanjima faktora, ne moramo da tražimo koje su kombinacije faktora različite. Ako bi se našle, onda bismo primenili Dankanov test za traženje razlika.

Sažetak

U ovom poglavlju smo se posvetili analize varijanse u slučaju kada treba da upoređujemo aritmetičke sredine u više grupa. Korišćenje ANOVA-e smo demonstrirali na dva primera eksperimenta i to jednofaktorijalnom i dvofaktorijalnom ANOVA-om. Naučili smo da pripremimo podatke za ovaj tip zadavanja, kao i da formulišemo model rešavanja. Pokazali smo da ANOVA još ne donosi definitivno rešenje, već da moramo da ga tražimo korišćenjem naredih postupaka. U sledećim poglavljima pokazaćemo korisnost ovog koncepta kod rešavanja pitanja zavisnosti promenljivih.

Vežbe

1. Studija uticaja tri različita načina anestezije na sadržaj folne kiseline u krvi (u μg/l) dala je rešenja iskazana u uzorku *red.cell.folate* iz datoteke *ISwR*. Istraživača je zanimalo da li je moguće izjaviti da su različiti načini anestezije (promenljiva *ventilation*) povezani sa različitim nivoom folne kiseline (promenljiva *folate*) u krvi pacijenta.

2. U kliničkoj studiji uticaja enalaprila na frekvenciju srca, koji je bio dat pre (nulti minut) i posle (30, 60 i 120 minuta) devetorici pacijenata sa kongestivnom bolešću srca. Istraživač se pitao da li može da kaže da postoji razlika u frekvanciji srca pri davanju leka u pojedinim vremenskim davanjima (uzorak *heart.rate* iz datoteke *ISwR).*

3. Prilikom studije farmakodinamike indomecithina, lek je bio davan intravenski šestorici dobrovoljaca. Naknadno su bile posmatrane koncentracije leka u krvi posle 0.25, 0.50, 0.75, 1.00, 1.25, 2.00, 3.00, 4.00, 5.00, 6.00, 8.00 sati. Istraživač je želeo da sazna da li su postojale razlike u aritmetičkim sredinama koncentracije indomecithina u posmatranim vremenskim intervalima (uzorak *Indometh* datoteke *datasets*).

OSMO POGLAVLJE

Regresija i korelacija

Sadržaj poglavlja

Cilj poglavlja

Do sada smo se interesovali naročito za to da li su jedna ili više aritmetičkih sredina jednake ili se razlikuju. Koristili smo različite metode za potvrđivanje hipoteza o ovom odnosu. Bila su to posmatranja, do kraja značajno komplikovana uticajem jednog ili više faktora i njihovim zajedničkim interakcijama. Da li može da se unese dinamika promena u odnosu promenljivih? Šta ako su promenljive od sebe zavisne tako da vrednost jedne zavisi od vrednosti druge, a ova zavisnost ima različite oblike? U ovom poglavlju objasnićemo kako da se ravnamo sa ovakvim stavom u slučaju jedne ili više zavisnih promenljivih, čija će vrednost zavisiti od aktuelne vrednosti nezavisne promenljive. Za saznanje ovog tipa zavisnosti koristimo postupak koji obično zovemo regresija, a prema vrsti odnosa govori se o raznim vrstama regresije. U ovom poglavlju posvetićemo se isključivo linearnoj regresiji, odnosno odnosu koji u ovom najjednostavnijem odnosu dve promenljive možemo izraziti pravom. Pojam regresije je bio uveden od lorda Galtona (1822 - 1911), britanskog naučnika koji je proučavao odnos između visine dece i njihovih roditelja. Saznao je da su deca ekstremno niskih ili ekstremno visokih roditelja pokazala tendenciju vraćanju prema prosečnim vrednostima. Zato je postupak kojim je ovu pojavu kvantifikovao najpre nazvao reverzijom, a kasnije mu je promenio naziv u regresiju. Danas ovaj naziv koristimo za sve postupke koji bezleže odnose između zavisne (krajnje) promenljive Y i nezavisne (kovarijantne ili objašnjavajuće) promenljive X. Cilj regresije je da se ustanovi da li su X i Y u nekom sistemskom odnosu i da se pretpostave ili otkriju vrednosti Y, koje odgovaraju vrednostima X.

Analiza korelacije, zasnovana na sličnom postupku kao analiza regresije, meri snagu odnosa među promenljivima. I ovaj termin je uveo lord Galton. Pojednostavljeno možemo da kažemo da su dve promenljive u korelaciji ako su promene u jednoj promenljivoj praćene promenom u drugoj, bilo u jednakom, bilo u suprotnom smislu promene. Na primer, incidencija ishemijske bolesti srca je u pozitivnoj korelaciji sa mekošću vode za piće, kako su to pokazale brojne epidemiološke studije. Dakle, što je voda tvrđa, time je veća pojava IBS.

Tabela 1: Ciljevi poglavlja

1. Objasniti postupak linearne regresije
2. Biti svestan osobina pojedinih parametara
3. Pokazati načine merenja kvaliteta regresije
4. Proširiti prosta linearnu regresiju na multiplu
5. Pokazati mogućnosti predikcije
6. Uvesti korelaciju kao meru snage odnosa

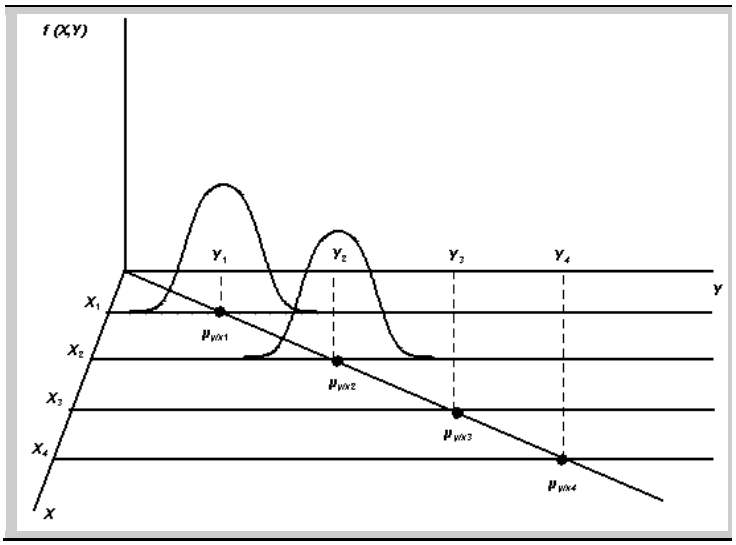
Prosta linearna regresija

I u slučaju regresije rešićemo osnovno pitanje statističke analize, odnosno kakve zaključke možemo da donesemo o situaciji u populaciji na osnovu podataka iz uzorka. U svakom slučaju, neophodno je računati sa greškom, kako smo to naveli u prethodnim poglavljima. Regresija će nam pružiti model odnosa koji je retko kad perfektan, a otežan je određenom greškom. Veličina ove greške može da presudi da li je model dovoljno tačan prikaz situacije u populaciji i da li ga je moguće koristiti i za predikciju vrednosti.

Regresiju možemo da sprovedemo pri ispunjenju određenih pretpostavki o zavisnoj

promenljivoj *Y* i nezavisnoj promenljivoj X^i. U prvom redu, pretpostavljamo da je nezavisna promenljiva *X* nepromenljiva, dakle da se u toku posmatranja njene osobine ne menjaju. Takav slučaj može da bude vreme, uzrast, udaljenost i slično. Kada odlučimo da će nezavisna promenljiva biti uzrast i posmatramo je u godišnjim intervalima, ne smemo da ovu odluku menjamo tokom istraživanja. Ujedno moramo da pretpostavljamo da je zavisna promenljiva izmerena bez greške, iako znamo da to nije uvek tako. Pretpostavljamo da iako postoji greška merenja, pa je smatramo tako malom da je ne moramo uzimati u obzir. Sledeća pretpostavka je da svakoj vrednosti *X* odgovara skup vrednosti Y, koje su normalno raspodeljene. Zato pretpostavljamo da su vrednosti Y normalno raspodeljene. Ujedno pretpostavljamo da aritmetičke sredine svih subpopulacija *Y* leže na jednoj pravoj i zato mogu biti prikazane jednakošću prave. Na kraju, pretpostavljamao da su vrednosti *Y* statistički nezavisne, odnosno vrednosti *Y* izabrane za jedno *X* ne utiču na vrednosti Y za drugo X.

Slika 1: Linearna regresija i uzorak populacije. Svakoj vrednosti X odgovara skup vrednosti Y, koje su normalno raspodeljene i njihove aritmetičke sredine leže na pravoj



Iz srednje škole se možda sećate da je jednačina prave $a = i + k*b$, gde kada stavimo za *i* tačku u kojoj preseca nultu liniju, za razne vrednosti b dobićemo pojedine tačke prave. I u našem slučaju linearna regresija je prikazana pravom

$$y = \alpha + \beta x + \varepsilon$$

gde su α i β koeficijenti regresije, iz kojih α *izražava tačku* u kojoj seče pravu os y, a β predstavlja ugao koji prava regresije zatvara sa osom x. Parametar ε predstavlja grešku.

Cilj regresivne analize je pronaći ovakvu pravu, otkriti njene parametre u populaciji na osnovu podataka iz uzorka i zaključiti koliko ona odgovara (objašnjava) stvarnosti. Postupak izračunavanja regresije predstavićemo na primeru (Primer 1).

[i] Velikim slovima označavamo promenljive kod populacije, dok malim slovima označavamo promenljive u uzorku.

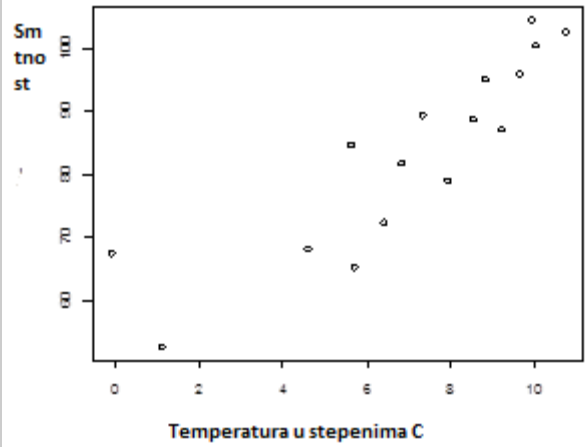
Primer 1: Zadavanje zadatka

Epidemiolozi iz više evropskih zemalja su pokušali da odgovore na pitanje da li može da se utvrdi odnos između prosečne godišnje temperature i pojave raka dojke kod žena. Na raspolaganju su imali podatke iz određenih oblasti Velike Britanije, Norveške i Švedske. Dobili su sledeće podatke:
Smrtnost: 102.5, 104.5, 100.4, 95.9, 87, 95, 88.6, 89.2, 78.9, 84.6, 81.7, 72.2, 65.1, 68.1, 67.3, 52.5
Temperatura: 10.7, 9.9, 10, 9.6, 9.2, 8.8, 8.5, 7.3, 7.9, 5.6, 6.8, 6.4, 5.7, 4.6, -0.1, 1.1

```
> rak <- c(102.5, 104.5, 100.4, 95.9, 87, 95, 88.6, 89.2, 78.9, 84.6, 81.7, 72.2, 65.1, 68.1, 67.3, 52.5)
> temperatura <- c(10.7, 9.9, 10, 9.6, 9.2, 8.8, 8.5, 7.3, 7.9, 5.6, 6.8, 6.4, 5.7, 4.6, -0.1, 1.1)
```

Rešenje zadatka počećemo time što ćemo temperaturu okoline smatrati nezavisnom promenljivom (ona se menja nezavisno od naše volje i zato je precizno data za svaku zemlju), a smrtnost od raka ćemo smatrati zavisnom promenljivom. Ujedno možemo da konstatujemo da su zadovoljeni svi uslovi da bismo verovali da su pretpostavke za regresiju bile ispunjene. Prvi korak će biti da prikažemo rezultate formom rasutog dijagrama (Slika 2).

Slika 2: Rasuti dijagram ulaznih podataka



Već na prvi pogled na dijagram vidimo da sa povećavanjem temperature raste i smrtnost. Čitalac bi mogao da uzme lenjir i olovku i da nacrta pravu koja bi išla između podataka tako da bude približno jednako udaljena od svakog merenja. Kako to uraditi tačno, da bi što tačnije bio prikazan odnos između obe promenljive? Jedan od mogućih kriterijuma za unošenje prave je postizanje što manjeg odstupanja prave od izmerenih podataka. S obzirom na to da podaci mogu da odstupaju od prave u pozitivnom ili negativnom smeru, odstupanja se dižu na kvadrat (time se gubi negativna vrednost i pozitiva odstupanja). Metoda izračunavanja koja vodi ka određivanju prave se naziva **metoda najmanjih kvadrata.** Za čitaoca koji bi hteo da se bliže upozna sa ovom metodom, na internetu se može pronaći dovoljno informacija o ovoj metodi. Mi ćemo iskoristiti računar, koji

će da izračuna oba parametra ovakve prave, odnosno intercept – tačku preseka sa osom y, a time i odstupanje od nule, a takođe i nagib prave.

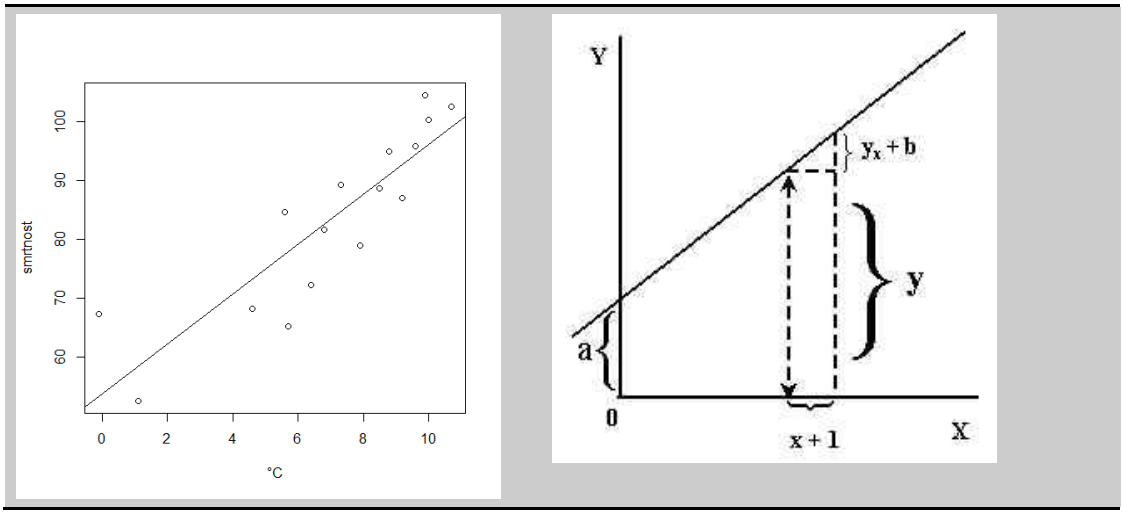
Slika 3: Prelazake prave kroz podatke

```
> fit <- lm(formula = rak ~ temperatura)
> fit
Call:
lm(formula = rak ~ temperatura)

Coefficients:
(Intercept)     temperatura
     53.611          4.248
```

Objasnićemo šta znače pojedini koeficijenti u ovom primeru. Kao što smo već rekli, jednačina će imati oblik *y = 53,6 + 4,25*x*. U našem slučaju x označava prosečnu temperaturu u °C. Pretpostavimo da ćemo vrednost x postepeno povećavati za 1. Onda će za vrednost 0 rezultat biti *y = 53,6 + 4,25*0 = 53,6*. To je intercept ili presek sa nulom. Kad je povećamo za 1, dobićemo *y = 53,6 + 4,25*1 = 57,85* i ako idemo dalje, dobićemo *y = 53,6 + 4,25*2 = 62,1*, povećanje na tri doneće *y = 53,6 + 4,25*3 = 66,35* i tako dalje. Svako povećanje x za jednu jedinicu rezultovaće porast vrednosti y tačno za vrednost koeficijenta. Ovaj koeficijent zovemo **jedinični priraštaj** (smanjenje, ako je vrednost negativna) ili **priraštaj kod jedinične promene**. On definiše nagib prave. Rezultat je moguće prikazati slikom (Slika 4).

Slika 4: Prikazivanje rezultata linije regresije i pokazivanje koeficijenta prilikom jedinične promene

```
> koef <- coef(fit <- lm(formula = rak ~ temperatura)) # u promenljivu koef
smo stavili koeficijente iz regresije pomoću komande coef, koja ih ekstrahuje
u formi promenljivih
> plot(temperatura, rak, xlab="°C", ylab="smrtnost") # nacrtali smo
pojedine tačke
> abline(coef=koef) # i dodali pravu za korišćenje koeficijenata linije
regresije
```



Slika desno prikazuje kako zapravo funkcioniše jedačina koju smo preveli podacima pomoću metoda najmanjih kvadrata (odstupanja). Malo *a* predstavlja pomeranje prave nasuprot nule, malo *b* izražava priraštaj prilikom jedinične promene nezavisne promenljive; drugim rečima, *b* izražava brojnost promene vrednosti zavisne promenljive za jednu jedinicu, prikazane kao x+1, malo *e* predstavlja grešku sa kojom izražavamo sve podatke. Prisetimo se ujedno osnovnog principa statističke indukcije. Koeficijenti koje smo izračunali na osnovu vrednosti uzorka su tačka otkrića stvarnih koeficijenata prave u populaciji. Zato je cilj regresije da se otkrije gde se nalaze stvarne vrednosti koeficijenata *a* i *b* i na osnovu njih da se otkriju stvarne vrednosti zavisne promenljive *y*. Zapis jednačine prave, koju se trudimo da otkrijemo kod populacije glasiće

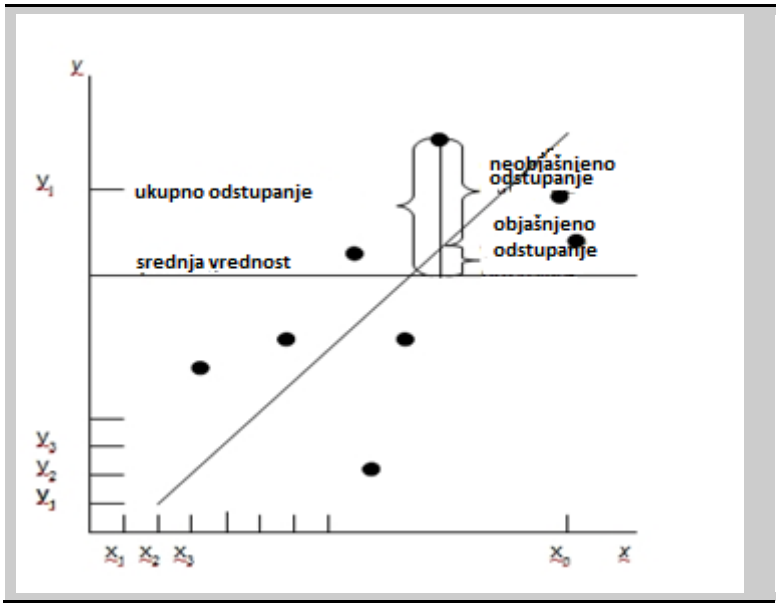
$$\mathbf{y = \alpha + \beta * x + \varepsilon}$$

U narednom delu posvetićemo se kvalitetu odgovaranja regresivne prave podacima i vrednovanju koeficijenata.

Kvalitet regresivne prave

Kako možemo da vidimo na slici, prava retko kad prolazi kroz sva merenja. Pojedine tačke odstupaju od prave i prosečne vrednosti u smislu plus ili minus. Što više podaci odstupaju od prave, to manje prava objašnjava njihov odnos. Snaga odnosa između nezavisne i zavisne promenljive je srazmerna tome koliko prava (ili kriva) odgovara merenjima. Tada govorimo o **tačnosti odgovaranja**. Kada prava odgovara? Što više se približava tačkama merenja, a takođe i što je manja disperzija tačaka. Zato slično kao u ANOVA-i, pokušaćemo da otkrijemo vrednost disperzije. Slika 5 beleži odnos merenja prema pravoj i prema aritmetičkoj sredini vrednosti.

Slika 5: Ukupno, objašnjeno i neobjašnjeno odstupanje



Kao i u slučaju računanja disperzije kod ANOVA-e, i u ovom slučaju izračunavamo drugi koren odstupanja sa ciljem da odstranimo zajedničko delovanje pozitivnih i negativnih vrednosti. Računamo sledeće zbirove odstupanja:

1. **Ukupno odstupanje** predstavlja udaljenost tačke od srednje vrednosti svih tačaka;
2. **Objašnjeno odstupanje** je razlika vrednosti pripadajuće tačke na pravoj i srednje vrednosti. Objašnjenim se naziva zato što pokazuje koji deo odstupanja snižava regresivna prava.
3. **Neobjašnjeno odstupanje** je udaljenost merenja od regresivne prave.

Kada ova odstupanja dignemo na kvadrat i saberemo, dobijemo zbir kvadrata, odnosno celokupan zbir kvadrata, objašnjen i neobjašnjen zbir kvadrata.

Logično je da zbir objašnjenog i neobjašnjenog zbira kvadrata bude jednak ukupnom. Ove vrednosti onda koristimo kao mere disperzije. Ukupan zbir kvadrata predstavlja ukupnu disperziju vrednosti zavisne promenljive *Y*. Objašnjen zbir kvadrata predstavlja disperziju posmatrane vrednosti *Y* objašnjen regresivnom pravom. I, obrnuto, neobjašnjen zbir kvadrata predstavlja deo disperzije koju prava nije u stanju da objasni. Jasno je da što više regresivna prava uspe da objasni disperziju, to je odgovaranje bolje. U idealnom slučaju trebalo bi vrednost objašnjene disperzije približavati ka celokupnoj disperziji. Iz toga se izvela i mera, pomoću koje vrednujemo kvalitet regresije, a zovemo je **koeficijentom određenosti** ili **determinacije** i označavamo je kao r^2 . Objašnjavamo ga kao meru koliko blizu je regresivna prava pojedinim tačkama merenja.

Slika 6: Detaljan ispis rezultata regresije pomoću komande summary()

```
> summary(fit)

Call:
lm(formula = rak ~ temperatura)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-12.7219  -5.7121   0.3592   4.3806  14.1140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.6107     4.7845  11.205 2.24e-08 ***
teplota       4.2476     0.6282   6.762 9.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 7.542 on 14 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:
0.7488
F-statistic: 45.72 on 1 and 14 DF,  p-value: 9.148e-06
```

Iz cele tabele rezultata za nas je trenutno najzanimljiva vrednost označena kao *Multiple R-squared* sa vrednošću 0,77. Dakle, više od ¾ disperzije je objašnjeno regresivnom pravom. Sledeća slika 7 demonstrira razne slučajeve sa kojima možemo da se sretnemo u praksi.

Do sada smo se bavili vrednošću koeficijenta determinacije izračunatog iz merenja koja direktno govore o odgovaranju prave pojedinačnim merenjima, kao i o disperziji vrednosti u uzorku. Kako smo već naučili u prethodnim poglavljima, potrebno je još da saznamo kakva je situacija kod populacije, kako možemo da otkrijemo koliko su promenljive *X* a *Y* u linearnoj zavisnosti. To zahteva postupak potvrđivanja hipoteze i definiciju nulte i alternativne hipoteze.

Kako smo već naznačili u postavljenom pitanju, nulta hipoteza će biti o nepostojanju pravolinijskog odnosa između promenljivih *X* i *Y*, koje su bile otkrivene na osnovu odabranog uzorka i glasiće:

H_0: *X* i *Y* nisu u pravolinijskoj zavisnosti

H_A: *X* i *Y* jesu u pravolinijskoj zavisnosti

Za potvrđivanje H_0 koristićemo isti postupak kao u slučaju ANOVA-e. Prisećamo se da smo testirali **odnos disperzija** (varijansi) OV, koji je jednak odnosu prosečnog zbira kvadrata objašnjenog i neobjašnjenog odstupanja. Za izračunavanje prosečnih zbirova potrebno je još odrediti stepen slobode. Važi pravilo da je broj stepena slobode za zbir kvadrata objašnjen regresijom jednak broju konstatnti (koeficijenata) snižen za jedinicu. U slučaju jednostavne linearne regresije, imamo dve konstante i to intercept i koeficijent nagiba prave. Iz toga proizilazi da ćemo za objašnjen deo zbira kvadrata koristiti broj stepena slobode 2 – 1 = 1. Za neobjašnjen deo koristićemo *n – 2* stepena slobode, dakle u ovom slučaju 16 – 2 = 14. Sam odnos izračunaćemo kao odnos prosečnog zbira kvadrata objašnjenog i neobjašnjenog (Tabela 1)

Slika 7: Razne vrednosti odgovaranja regresivne prave vrednostima merenja

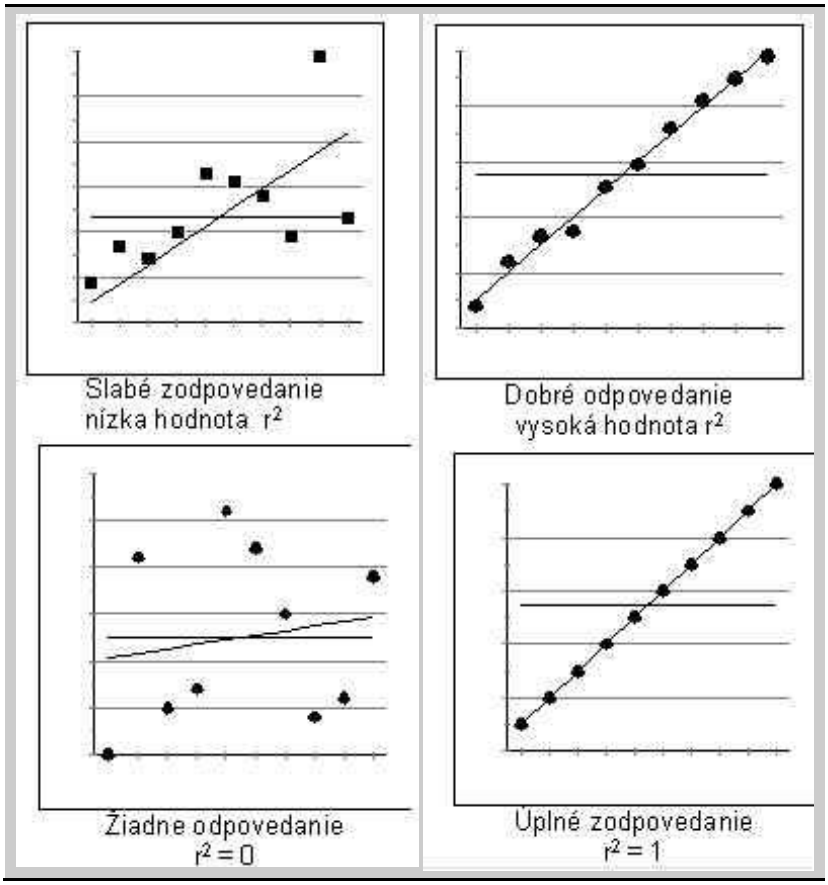


Tabela 5: Odnos disperzije za izračunavanje ANOVA za linearnu regresiju

Izvor varijanse	**Zbir kvadrata ZK**	**Stepen slobode d.f.**	**Kvadrat aritm. sredine KAS**	**Odnos varijansi OV**
Linearna regresija	$ZK_{objašnjen}$	*1*	$ZK_{objašnjen}/1$	$ZK_{objašnjen}/$
Neobjašnjen (rezidua)	$ZK_{neobjašnjen}$	*n - 2*	$ZK_{neobjašnjen}/(n - 2)$	$ZK_{neobjašnjen}$
Zajedno	ZK_{ukupno}	*n - 1*		

O prihvatanju ili odbacivanju hipoteze H_0 odlučićemo na osnovu F testa. Iako u tabeli rezultata nećemo naći vrednost OV, program nam pruža direktnu vrednost verovatnoće za prihvatanje, odnosno odbacivanje H_0. Pogledajmo, šta nam o tome govori rezultat naše statistike:

F-statistic: 45.72 on 1 and 14 DF, p-value: 9.148e-06

Broj 9,1e-06 nam dozvoljava odbacivanje H_0, a prihvatanje H_A sa nivoom poverenja $p < 0,0001$, šta je veoma visoka vrednost. Dakle, možemo da zaključimo da je linearni odnos između prosečne temperature i smrtnosti od raka značajan sa verovatnoćom $p < 0,0001$.

Interval poverenja za koeficijent β regresivne prave

Regresivni koeficijent β predstavlja stvarnu vrednost nagiba prave u populaciji. U našem primeru vrednost koeficijenta *b* smo izračunali iz uzorka. Dakle, možemo da izjavimo da je *b* tačka otkrića β, a za njegovo otkriće potrebno je saznati da li se razlikuje od nule i kakav je odgovarajući interval poverenja. Zato što i njegovu varijansu kod populacije otkrivamo na osnovu standardne devijacije izračunate iz uzorka, moramo da koristimo Studentovu t-raspodelu. Ujedno pretpostavljamo da su svi uzorci ovog koeficijenta u normalno raspodeljenoj populaciji.

Pokušaćemo da saznamo da li se stvaran koeficijent u populaciji razlikuje od nule. Drugim rečima, potvrdićemo nultu hipotezu da je vrednost β jednaka nuli.

$$H_0 = 0 \qquad H_A \neq 0$$

Potvrđivanje ove hipoteze sprovodi se istim postupkom kao pri potvrđivanju hipoteze o aritmetičkoj sredini. Nećemo prelaziti celo računanje, ali ćemo pogledati šta nam je izračunao statistički program (Slika 5). Iz njega proizilazi da je vrednost tačke otkrića koeficijenta β 4,25, standardna greška otkrića je 0,63, kritična vrednost t raspodele za odgovarajući broj stepena slobode je 6,76, a pritom dodata vrednost verovatnoće odbacivanja H_0 jeste 0,000009. Dakle, možemo da zaključimo da verovatnoća da je nagib prave jednak nuli tako mala, da možemo da kažemo, da sa verovatnoćom $p < 0,0001$ nagib jeste različit od nule. Isti postupak možemo da koristimo i za tačku preseka sa nultom osom, a i tamo vidimo da se ovaj razlikuje od nule sa verovatnoćom $p < 0,0001$.

Konstrukcija intervala poverenja za utvrđivanje koeficijenta β je zasnovana na istoj koncepciji kao utvrđivanje aritmetičke sredine. Dakle, moraju da nam budu poznate tačke otkrića a i b, kao i standardna greška i onda određujemo na kom nivou poverenja želimo da računamo. Ove vrednosti su nam poznate, ali da ne moramo da izračunavamo konkretne vrednosti, program će ih izračunati za nas. Koristićemo komandu *confint(object, parm, level = 0.95, ...)*. Za *object* unesemo promenljivu u kojoj smo zamenili rezultat regresije, u *parm* možemo detaljisati za koje koeficijente želimo da računamo; ako to ne navedemo, onda će program da izračuna za sve. Na kraju, odredićemo nivo signifikantnosti za koji izračunavamo odgovarajuće intervale poverenja. Ako ga ne navedemo, automatski se primeni vrednost 0.95. Rezultat je ste u sledećoj slici (Slika 8).

Slika 8: Izračunavanje intervala poverenja za koeficijent regresije

```
> ci <- confint(fit)
> ci
                2.5 %    97.5 %
(Intercept) 43.348947 63.872553
temperatura      2.900218  5.594925
```

Iz slike vidimo da stvarna vrednost koeficijenta regresije nazvanog temperatura leži između 2,9 i 5,6 smrtnosti. To znači da pri povećanju temperature za jedan stepen možemo da očekujemo povećanje smrtnosti u rasponu 2,9 do 5,6 smrtnosti.

Verovatnoća ili predikcija

Jedna od praktičnih primena prave regresije je verovatnoća. Ona služi za dobijanje odgovora na pitanje kakvu vrednost dobija *Y* u tački *X* koju nismo izmerili. Zaista je široko korišćenje predikcije u slučajevima kada želimo da pronađemo odgovore na pitanja iz svakodnevnog života. Na primer, znamo da je visina deteta do određene mere pravolinijska u određenom dobu. Mi je merimo, recimo, u dva mesečna intervala. Postavljamo pitanje kakva bi bila vrednost posle mesec dana. Ili imamo povećanje oboljenja (ili pad) i želimo da kažemo šta će se desiti za mesec ili za godinu. U slučaju predviđanja, međutim, moramo da budemo svesni da je predviđanje važeće samo ako ostanu očuvani uslovi u kojima je regresija bila izračunata.

Samu predikciju uradićemo veoma jednostavno sa korišćenjem komande *predict()*. U slučaju da želimo predikciju za tačke koje se nalaze između izmerenih tačaka, koristićemo pozivanje ove komande sa argumentom u formi izvršene regresije. Kada želimo da izračunamo regresiju za određene vrednosti nezavisne promenljive, moramo je specifikovati kao parametar, sa istim nazivom kakav je imala nezavisna promenljiva, u našem slučaju to je bila *temperatura*. Da bi bio u okviru promenljive, i za sam naziv je neophodno koristiti okvir za podatke, dakle *data.frame*. Postupak izračunavanja i rezultat je u sledećoj slici (Slika 9).

Slika 9: Predikcija vrednosti pomoću linearne regresije

```
> temperatura <-seq(1,20,1)  #napravili smo promenljivu, koja sadrži sekvenciju brojeva od 1 do 20
> temperatura <- data.frame(teplota) # promenljivu smo transformisali u okvir za podatke
> pred <- predict.lm(fit, newdata=teplota) #pozivamo komande predikcije i samu predikciju zamenimo u promenljivu
> pred
       1        2        3        4        5        6        7        8
 57.85832  62.10589  66.35346  70.60104  74.84861  79.09618  83.34375  87.59132
       9       10       11       12       13       14       15       16
 91.83889  96.08646 100.33404 104.58161 108.82918 113.07675 117.32432 121.57189
      17       18       19       20
125.81946 130.06704 134.31461 138.56218
```

Predikcija nam je dala seriju vrednosti koje odgovaraju vrednostima nezavisne promenljive temperatura i izvan okvira izvorno izmerenih vrednosti. Izvorno najviša temperatura je bila 10,7°C. Mi smo napravili predikciju koja nam govori da kad bi se smrtnost od raka i prilikom viših prosečnih temperatura održala onako kako smo izračunali, onda bi, na primer, pri temperaturi od 17°C bila 125,82. Ali, uvek je to procena tačke, a kako onda da dobijemo procenu intervala? Jednostavno dodamo argumente koji omogućavaju izračunavanje procene intervala, odnosno intervala poverenja (Slika 10).

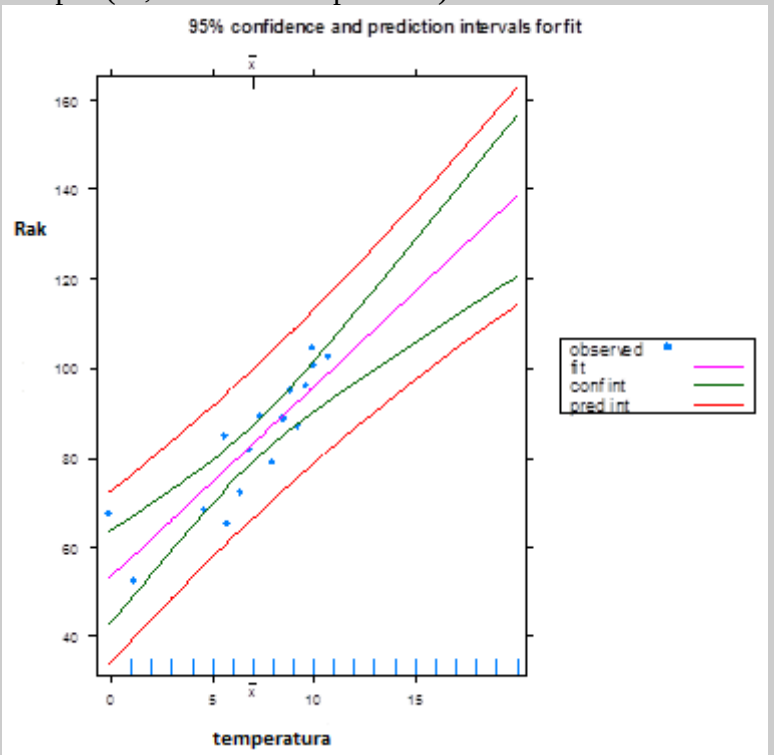
Slika 10: Izračunavanje intervala poverenja za predikciju

```
> pred <- predict.lm(fit, newdata=temperatura, interval="confidence")
> pred
       fit      lwr      upr
1  57.85832  48.81925  66.89740
2  62.10589  54.24867  69.96311
3  66.35346  59.61569  73.09124
4  70.60104  64.88351  76.31856
5  74.84861  69.98925  79.70796
6  79.09618  74.83387  83.35848
7  83.34375  79.30001  87.38749
8  87.59132  83.32902  91.85363
9  91.83889  86.97954  96.69825
10 96.08646  90.36894 101.80399
11 100.33404 93.59626 107.07181
12 104.58161 96.72439 112.43883
13 108.82918 99.79010 117.86825
14 113.07675 102.81495 123.33855
15 117.32432 105.81194 128.83671
16 121.57189 108.78924 134.35454
17 125.81946 111.75220 139.88673
18 130.06704 114.70440 145.42967
19 134.31461 117.64836 150.98086
20 138.56218 120.58586 156.53850
```

Na kraju ostaje samo da nacrtamo sliku. Pre toga moramo da objasnimo razliku između intervala poverenja i intervala predikcije. Intervali poverenja govore o verovatnom mestu gde se nalazi parametar populacije. Predikcije vremenskih intervala govore gde možemo da očekujemo naredne podatke tačke uzorka. Nasuprot ovome, interval prognoze, odnosno predikcije, govori o rastavljanju vrednosti, ne o nesigurnosti prilikom određivanja vrednosti iz populacije. Zato je uvek širi od intervala poverenja. Pokazaćemo to na slici (Slika 11).

Slika 11: Iscrtavanje 95% intervala poverenja i intervala predikcije

```
> ci.plot(fit, newdata=temperatura)
```



Korelacija

Regresivni model kojim smo se do sada bavili pretpostavlja da je zavisna promenljiva (*Y*) slučajna, normalno distribuirana promenljiva. Nezavisna promenljiva nije slučajna, odnosno fiksna je. Zadatak regresivne analize je traženje odgovarajuće vrednosti zavisne promenljive prema izabranim tačkama nezavisne. Često se javi situacija da tražimo da li postoji ili ne postoji odnos između dve slučajne promenljive. Ovakav odnos onda zovemo **korelacija**. Pod njim podrazumevamo odnos dve ravnopravne promenljive i korelacija izražava snagu tog odnosa. Ako je odnos linearan, za njegovo izračunavanje koristimo postupak istovetan sa regresijom. U tom slučaju pretpostavljamo da su obe promenljive rastavljene prema zajedničkoj varijansi, koju nazivamo **zajednička podela** ili distribucija. Kada se radi o normalnoj distribuciji, govorimo o **bivarijantnoj normalnoj raspodeli**.

Snagu odnosa merimo analogno linearnoj regresiji, gde smo koristili koeficijent određivanja r^2 za izražavanje kvaliteta odgovaranja na tačke pravih. U slučaju korelacije koristićemo **koeficijent korelacije**, koji je poznat pod nazivom Pirsonov koeficijent korelacije. Njegova vrednost se kreće u rasponu od -1 do +1. Kada ima pozitivnu vrednost onda je reč o odnosu koga iskazuje rastuća prava (sa povećavanjem vrednosti x povećava se i vrednost y). Kad je njegova vrednost negativna, onda se prava spušta (sa povećavanjem vrednosti x vrednost y opada). Ako je koeficijent korelacije bliži ili jednak nuli, govorimo da odnosa, odnosno korelacije između dve promenljive nema. Objasnićemo bliže vrednosti koeficijenta korelacije. Ako je ovaj jednak

broju 1 ili -1, onda govorimo o jakoj povezanosti, a sve tačke leže u jednoj ravni. Ako izračunat koeficijent korelacije nije jednak ± 1, ali je dovoljno blizu, možemo da kažemo da ova koleracija ukazuje na jaku vezu između dve promenljive. Ako se koeficijent korelacije približava ± 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. Izračunavanje koeficijenta korelacije ćemo pokazati u sledećem primeru (Primer 2).

Primer 2: Utvrđivanje snage korelacije između procenta Meticilin rezistentnog *Staph. aureusa* MRSA na odeljenju bolnice i dužine hospitalizacije bolesnika

Istraživača je zanimalo da li je moguće govoriti o odnosu između učestalosti MRSA i dužine hospitalizacije bolesnika sa određenom dijagnozom na odeljenju intenzivne nege. Prošao je 20 odeljenja i saznao je sledeće podatke:

Odeljenje	% MRSA	Prosečna dužina hospitalizacije
1	12	6
2	14	6
3	27	8
4	24	9
5	36	7
6	32	6
7	44	12
8	52	11
9	48	13
10	38	11
11	45	10

Odlučio je da izračuna koeficijent korelacije za određivanje tipa i snage odnosa.

Postupak:

```
> MRSA <-c(12, 14, 27, 24, 36, 32, 44, 52, 48, 38, 45)
> hosp <-c(6, 6, 8, 9, 7, 6, 12, 11, 13, 11, 10)
> cor(MRSA, hosp)
[1] 0.8012835
```

Izračunata vrednost koeficijenta korelacije jeste visoka, ali bismo trebali da potvrdimo da li je njegova stvarna vrednost u populaciji nije jednak nuli potvrđivanjem nulte hipoteze o jednakosti koeficijenta sa nulom.

```
> cor.test(MRSA, hosp)

        Pearson's product-moment correlation

data:  MRSA and hosp
t = 4.0179, df = 9, p-value = 0.003028
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3878236 0.9463004
sample estimates:
      cor
```

0.8012835

Interpretacija: Potvrđivanje je pokazalo da možemo da odbacimo nultu hipotezu sa verovatnoćom manjom od 0,01. Ujedno smo dobili procenu intervala stvarnog koeficijenta korelacije. Zaključujemo da postoji srednje jak, pa čak i jak odnos između MRSA i dužine hospitalizacije, takav da sa rastućom učestalošću MRSA se produžava i vreme hospitalizacije.

Sažetak

Bavili smo se upoznavanjem odnosa dve promenljive: nezavisne i zavisne. Objasnili smo način izračunavanja prave rezultata merenja pomoću metode najmanjih kvadrata. Ujedno smo objasnili kako se meri kvalitet rezultata postupkom sličnim kao kod ANOVA-e. Takođe smo govorili o korišćenju regresije za pretpostavljanje ili predikciju vrednosti. Na kraju poglavlja smo naveli koncepciju merenja snage odnosa dve promenljive, što smo nazvali korelacija.

Ovo poglavlje daje samo uvod u širu problematiku regresije i korelacije. Njen prirodan sled su metode multiple regresije i korelacije, kao i primena i drugih, nelinearnih regresija. S obzirom da ova publikacija ima za cilj da navede samo osnove biostatistike, od svih navedenih mogućnosti u posebnom poglavlju ćemo se baviti logističkom regresijom (Poglavlje 9). Čitaoca koga interesuju i druge vrste regresije upućujemo na izvore napredne statističke literature.

Vežbe

- U studiji odnosa visine sistolnog krvnog pritiska i telesne mase pregledana je 102 slučajno odabranih muškaraca i žena u malom mestu u SAD. Istraživača je zanimalo da li se povećavanjem vrednosti aritmetičke stvarne telesne mase prema idealnoj povećava i sistolni pritisak (uzorak *bp.obese* iz datoteke *ISwR*). Pokušajte da nacrtate i regresivne odnose.

- Na osnovu podataka iz prethodnog zadatka odredite korelaciju promenljivih za muškarce i žene.

DEVETO POGLAVLJE

Logistička regresija

Sadržaj poglavlja

Cilj poglavlja

Logistička regresija kao statistička metoda je jedna od najčešće primenjivanih u oblasti epidemiologije, kao i u biomedicinskim naukama. Postoji nekoliko razloga zašto je to tako, ali najbitnija je činjenica da pomoću ove metode možemo da detaljno analiziramo uticaj nekoliko promenljivih (istovremeno) na verovatnoću nastajanja posmatrane pojave. Ova premisa omogućava modeliranje multikauzaliteta nastajanja određene pojave (u slučaju epidemiologije radi se o oboljenju, smrtnosti, remisiji i slično). Cilj ovog poglavlja je da predstavi logističku regresiju, objasni njene rezultate i praktičnu primenu.

Regresivna analiza se karakteriše, između ostalog, i time što ima jasno definisanu takozvanu zavisnu i nezavisnu promenljivu. Slično kao i kod linearne, i kod logističke regresije možemo nezavisnu promenljivu da nazovemo prediktorom, koji će da na određeni način (tačno definisan smerom i jačinom) utiče na zavisnu promenljivu. Kao kod drugih tipova regresivne analize, i u slučaju logističke analize označavamo zavisnu promenljivu Y, a nezavisne promenljive $X_1,X_2....X_n$. Logistička regresija, međutim, osim sličnosti sa drugim tipovima regresije, ima i svoje specifičnosti. O njima ćemo govoriti u narednim delovima ovog poglavlja.

Tabela 1: Ciljevi poglavlja

1. Uvesti koncepciju logističke regresije i njene primene u biomedicinskim istraživanjima
2. Pojasniti računanje parametara u sredini R
3. Tačkaste i intervalske procene regresivnih koeficijenata i odnos šansi
4. Navesti prognostičke modele
5. Pristrasnost i interakcije.

Zavisna i nezavisna promenljiva u logističkoj regresiji

Zavisne promenljive

U slučaju regresivne analize, uopšteno zavisnom promenljivom zovemo onu koja izražava pojavu koju posmatramo, a koja je u vezi sa izloženošću. Klasičan primer iz epidemiologije je odnos između izloženosti faktoru rizika i bolesti. Bolest je u ovom slučaju zavisna promenljiva, ona zavisi od izloženosti, što predstavlja nezavisnu promenljivu.

U logističkoj regresiji zavisna promenljiva je uvek dihotomna (binarna), što znači da može da dobije samo dve moguće vrednosti: 0 ili 1. Uobičajeno se brojem 1 označava prisutnost određene pojave (dakle oboljenje, smrtnost, remisija i slično), a sa 0 se označava odsustvo određene pojave (dakle zdrav čovek, preživljavanje, odsustvo remisije i slično). Sama regresivna analiza onda ima za cilj da sazna uticaj jedne ili više nezavisnih promenljivih na verovatnoću da se data pojava (označena kao 1 u okviru zavisne promenljive) javi.

Veoma je bitno ispravno „kodirati“ zavisnu promenljivu, odnosno ispravno označiti pojavu jedinice ili nule. Statistički programi standardno daju rezultat pojave označene brojem jedan. Zato kad jedinicom označimo bolesne ljude (a nulom zdrave), modeliramo verovatnoću da će se čovek razboleti, a ako jedinicom označimo zdrave ljude (a nulom bolesne), modeliramo verovatnoću da se čovek neće razboleti. Između ovih situacija postoji suštinska razlika, pa je potrebno povesti pažnju već pri pripremanju baze podataka, koju ćemo posle koristiti prilikom analize.

Nezavisne promenljive

Uopšteno u regresivnoj analizi nezavisim zovemo promenljive koje analiziramo kao faktore koji utiču na posmatranu pojavu. Klasičan primer iz epidemiologije su opet izloženost i bolest. Izloženost je u ovom slučaju nezavisna promenljiva koja utiče na nastanak posmatrane pojave.

Logistička metoda je metoda koja omogućava analiziranje istovremenog uticaja nekoliko promenljivih na posmatranu pojavu. Zato se smatra adekvatnom metodom za multivarijantnu analizu u slučajevima kada je zavisna promenljiva binarna (bolest/zdravlje, smrt/život, i sl.). Dobra strana logističke regresije je što može da otkrije uticaj faktora (nezavisnih promenljivih) reprezentovanih mnogim drugima promenljivima – možemo da koristimo kontinuirane i kategorijalne promenljive u proizvoljnim kombinacijama (npr. kombinacija kontinuiranih faktora kao što su uzrast, visina, telesna masa i kategorijalnih faktora kao pol i sl.).

Primena logističke regresije

Logistička regresija olakšava izražavanje kompleksnih odnosa između uzoraka nezavisnih promenljivih i zavisnih promenljivih i identifikovanje nezavisne promenljive sa značajnim, signifikantnim uticajem na zavisnu promenljivu u okviru analiziranog uzorka.

Ujedno, na osnovu analize uticaja uzoraka nezavisne promenljive na zavisnu promenljivu u okviru analiziranog uzorka, logistička regresija omogućava predviđanje verovatnoće pojave, izražavajući je zavisnom promenljivom kod pojedinca sa poznatom kombinacijom nezavisnih promenljivih. To znači da omogućava izračunavanje verovatnoće smrtnosti, preživljavanja ili druge pojave kod individue sa poznatim parametrima koji čine nezavisnu promenljivu, na osnovu prethodnog računanja koeficijenata logističke regresije.

Za konkretniju ilustraciju navedimo primer niske porođajne mase i faktora koji su povezani sa ovom pojavom. Uzmimo zato populaciju majki i odaberimo iz nje uzorak 500 majki, kod kojih smo saznali da li je dete rođeno od date majke imalo normalnu ili nisku telesnu masu pri rođenju. Isto tako zabeležimo etničku pripadnost porodilje, njenu starost i da li je pušila tokom trudnoće.

Zavisna promenljiva, u ovom slučaju promenljiva koja izražava da li se dete rodilo sa normalnom ili sa niskom porođajnom masom, gde brojčani kod „1“ označava dete sa niskom porođajnom masom, a brojčani kod „0“ označava dete rođeno sa normalnom porođajnom masom. Nezavisnim promenljivim biće dodeljene informacije koje smo dobili o porodiljama, dakle etnička pripadnost, uzrast i pušenje.

Primenom prvog načina pravljenja logističke regresije možemo da analiziramo na koji način pojedine nezavisne imaju veze sa zavisnom promenljivom (na koji način etnička pripadanost, uzrast i pušenje majke utiču da li će se dete roditi sa niskom porođajnom masom), kvantifikovaćemo ove

uticaje i njihov statistički značaj.

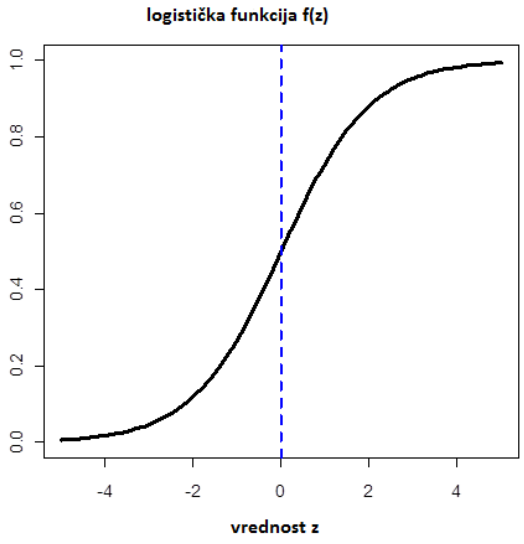
Primenom drugog načina izvršenja logističke regresije možemo da kod slučajno odabrane majke (iz populacije iz koje smo odabrali spominjan 500-člani uzorak) predvidimo sa kojom verovatnoćom će se njeno dete roditi sa niskom porođajnom masom. Pritom ćemo iskoristiti informacije dobijene prvim načinom primene logističke regresije i informacije o posmatranju nezavisnih parametara kod izabrane majke. Drugim rečima, ako kod izabrane majke saznamo uzrast, etničku pripadnost i pušenje, moći ćemo da predvidimo da li će se njeno dete roditi sa niskom ili normalnom porođajnom masom.

U nastavku ćemo objasniti na koji način primenjujemo i interpretiramo logističku regresiju postizanjem ova dva cilja.

Logistička funkcija i logistički model

Jedna od osnovnih premisa logističke regresije je logistička funkcija. Pojednostavljeno rečeno, logistička funkcija, koju označavamo sa f(z), je matematička osnova za izražavanje kompleksnih odnosa među nezavisnim promenljivima i zavisnim promenljivima, veza koje analiziramo. Slika prikazuje sigmoid logističke funkcije, gde se na osi x nalaze vrednosti Z, a osa Y predstavlja vrednost funkcije. Kako već znamo iz grafikona (Slika 1), vrednost f(z) pašće uvek u interval 0-1.

Slika 1: Grafička reprezentacija krive logističke regresije za jednu zavisnu promenjlivu



Iz rečenog proizilazi da za bilo koju vrednost Z, ona će biti u logističkoj funkciji u rasponu 0 - 1. Ovu činjenicu u epidemiologiji znamo da iskoristimo za izražavanje verovatnoće, koja se takođe nalazi u intervalu 0 - 1, respektivno 0 - 100%.

Od logističke funkcije ka logističkom modelu

Pomoću logističke funkcije f(z) možemo da saznamo verovatnoću posmatrane pojave sa pretpostavkom specifične kombinacije nezavisnih promenljivih izraženih vrednošću Z (primenom logističke regresije za predikciju verovatnoće određene pojave). Učinićemo to korišćenjem sledećeg odnosa, koji predstavlja matimatički zapis logističkog modela:

$$f(z)= 1 / 1-e^{-z}$$

Iz navedenog vidimo da je vrednost Z u logističkoj regresiji ključna. Pokazaćemo šta tačno predstavlja i kako ćemo doći do nje. Z predstavlja brojčano izražavanje kombinovanog uticaja uzorka nezavisnih promenljivih na verovatnoću pojave posmatrane zavisne promenljive. Ova vrednost izražava uticaj svake nezavisne promenljive i sa svakom promenom bilo koje od njih se menja. Uticaj pojedinih nezavisnih promenljivih na vrednost Z može se izraziti sledećim odnosom:

$$Z= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3 X_3 +.... \beta_nX_n$$

U ovom odnosu β_0 predstavlja tačku preseka, dakle vrednost Z u slučaju da sve nezavisne promenljive imaju vrednost nula (tj. vrednost Z u slučaju da dati čovek nema ni jedan od faktora rizika korišćeih u analizi). Vrednosti od β_1 do β_n su koeficijenti regresije, koji su brojčano izraženi uticajem pojedinih nezavisnih promenljivih od X_1 do X_n na rezultativnu vrednost Z. Drugim rečima, regresivni koeficijenti su brojčane konstante koje izražavaju u kom pravcu i kakvim intenzitetom će data promenljiva učestvovati u pojavi posmatrane pojave (npr. da će se čovek razboleti). Pod intenzitetom se u ovom kontekstu misli na to koliko promena intenziteta vrednosti date nezavisne promenljive promeni verovatnoću da se pojavi posmatrana pojava. Pod smerom uticaja se u ovom kontekstu misli da li povećanje vrednosti nezavisne promenljive poveća ili smanji verovatnoću da se pojavi posmatrana pojava. Vrednost Z zato detaljno uzima u obzir kakav je intenzitet i smer uticaja nezavisnih promenljivih na verovatnoću pojave posmatrane pojave.

Primer porođajne mase i njenih prediktora

Jednostavnije je zapisati i objasniti korake logističke regresije na konkretnom primeru.

Sa ciljem objašnjenja pojedinih koraka pri aplikaciji i interpretaciji logističke regresije, koristićemo datoteku *birthwt*, koja je deo programa R. Ova datoteka sadrži podatke o porođajnoj masi 189 dece i podatke o rizičnim faktorima majki, koji mogu da utiču na rođenje deteta sa masom manjom od 2500g, što jeste uopštena granica za nisku porođajnu masu (deca sa masom manjom od 2500g rođena su sa niskom porođajnom masom).

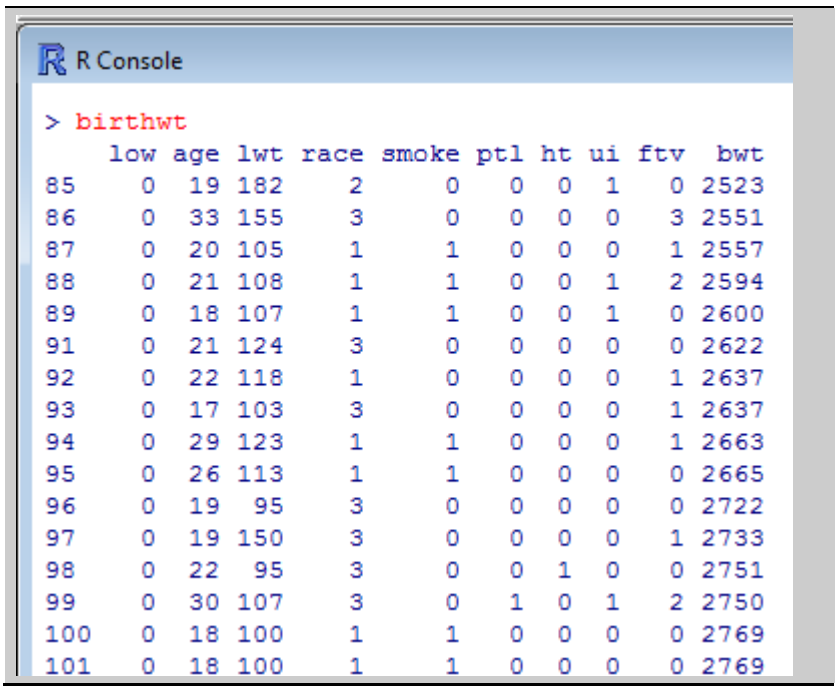
Zavisna promenljiva u ovom slučaju je faktor da li se dete rodilo sa niskom porođajnom masom ili sa normalnom porođajnom masom. U datoj bazi podataka zavisna promenljiva je promenljiva sa nazivom „low“, koja ima vrednost 1 ako je dete imalo porođajnu masu manju od 2500g, a vrednost 0, ako je imalo normalnu porođajnu masu, dakle iznad 2500g.

Datoteka sadrži i druge promenljive koje beleže prisutnost rizičnih faktora kod majki. Uzmimo u obzir uzrast majki, označen kao promenljiva sa nazivom „age“ (kontinuirana promenljiva u godinama), zatim pušenje majke tokom trudnoće, sadržano u promenljivoj sa

nazivom „smoke“ (kategorijalna promenljiva 1-da , 0-ne) i etnička pripadnost majke, iz promenljive sa nazivom „race“ (majka belkinja -1, majka crnkinja -2, majka druge rase -3). Obratite pažnju da kao predikatore u ovom slučaju koristimo kontinuiranu promenljivu (promenljivu „age“) i kategorijalnu promenljivu (promenljive „race“ i „smoke“). Logistička regresija omogućava analiziranje i kontinuiranih i kategorijalnih nezavisnih promenljivih.

Posle zadavanja na ekranu dobijamo sledeći prikaz u programu R library(MASS), data(birthwt) i list(birthwt), (Slika 2).

Slika 2: Ulazni podaci primera porođajne mase



Ako želimo da budemo sigurni da će promenljive, koje su kategorijalne, i biti analizirane kao kategorijalne (takva nesigurnost se pojavi ukoliko su kategorije izražene brojčanim kodom), transformišemo ih u R pomoću sledećih komandi (u našem primeru reč je o promenljivim „race“ i „smoke“):

birthwt$smoke=as.factor(birthwt$smoke) i

birthwt$race=as.factor(birthwt$race).

U našem slučaju posmatrana pojava je rođenje deteta sa niskom porođajnom masom. Saznaćemo, dakle, da li pojedini prediktori povećavaju ili smanjuju mogućnost da se dete rodi sa niskom telesnom masom, kao i intenzitet uticaja pojedinih promenljivih.

Za ovu namenu koristićemo komandu *glm* u sledećoj formi :

```
regresija  = glm (low ~ age+ smoke+ race, family=binomial, data=birthwt)
summary (regresia)
```

Komanda će u prvom redu napisati rezultate logističke regresije analize u objekat sa nazivom "regresija" (ovaj naziv je proizvoljan i možete ga staviti prema potrebi). Komanda *summary* onda odštampa sadržaj datog objekta na ekranu (Tabela 2).

Tabela 6: Rezultat logističke regresije

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.00755    0.86166  -1.169  0.24228
age         -0.03488    0.03340  -1.044  0.29634
smoke1       1.10055    0.37195   2.959  0.00309 **
race2        1.01141    0.49342   2.050  0.04039 *
race3        1.05673    0.40596   2.603  0.00924 **
---
```

Prva kolona prikazuje spisak prediktora (nezavisnih promenljivih), zajedno sa tačkom preseka (intercept). Druga kolona prikazuje koeficijente regresije (estimate), treća kolona prikazuje standardnu grešku koeficijenata regresije (std.error), četvrta kolona prikazuje rezučtat statističkog testa (Valdov test ili Z-test) za date prediktore, a poslednja kolona prikazuje p-vrednosti za odgovarajuće prediktore.

U slučaju kategorijalnih promenljivih „smoke" i „race" koeficijent je izračunat za sve kategorije date promenljive. Prilikom izračunavanja uticaja pojedinih kategorija, jedna od njih se automatski odredi kao referentna u pogledu prema kojoj se izračuna koeficijent regresije uticaja preostalih kategorija. Koeficijent regresije uticaja referentne kategorije je 0 (nula). Program R prilikom korišćenja komandi *glm* ili *lrm* ne piše u tabeli referentne kategorije. Zato je uvek ta kategorija nezavisne promenljive, koja se ne nalazi u tabeli, referentna. Program R određuje, kao referentnu, onu kategoriju koja se nalazi prva u redosledu. U slučaju da su kategorije označene brojčanim kodovima, reč je o kategoriji sa najmanjim brojčanim kodom. U slučaju da su kategorije označene rečima, referentna će biti ona kojoj se početno slovo nalazi prvo u abecedi. Ako želimo da promenimo ovaj redosled, potredno je promenljive odgovarajuće svrstati. Ovom aspektu bliže ćemo se posvetiti u delu o interpretaciji rezultata.

Predmet našeg interesovanja iz pogleda interpretacije rezultata su, iz navedene tabele, najpre koeficijenti regresije i odgovarajuće p-vrednosti.

Koeficijent regresije uopšteno određuje pravac i intenzitet uticaja datog prediktora na razvoj posmatrane pojave. Ako koeficijent regresije pozitivnu vrednost (u ovom slučaju, na primer, promenljiva „smoke"), u interpretaciji možemo navesti da pripadanje date majke kategoriji „1" promenljive „smoke" (dakle, ta majka je pušač) povećava verovatnoću da će se dete roditi sa niskom porođajnom masom. Ukoliko koeficijent regresije ima negativnu vrednost (u ovom slučaju, na primer, promenljiva „age"), uticaj je suprotan, odnosno ako se poveća vrednost promenljive, snižava se i verovatnoća da se dete rodi sa niskom porođajnom masom.

P-vrednost Valdovog testa ukazuje na to da li je uticaj date nezavisne promenljive na zavisnu promenljivu, iskazan koeficijentom regresije, statistički signifikantan ili ne. Data p-vrednost se odnosi na test hipoteze o nultom uticaju, odnosno o nezavisnosti datih promenljivih. U našem slučaju p-vrednost pripadajuća promenljivoj „age" odnosi se na testiranje hipoteze o

nezavisnoj promenljivoj „age“ kao prediktora (nezavinle promenljive) i zavisne promenljive „low“, koja sadrži informaciju o niskoj, odnosno normalnoj porođajnoj masi. Nećemo se posebno zadržavati na interpretaciji p-vrednosti, zato što je ona opisana u prethodnim poglavljima.

Koeficijenti regresije, u obliku kako smo ih do sad izračunavali, su dosta komplikovani za interpretaciju. Jednostavnije je da se dati koeficijenti regresije preračunaju na vrednosti Odds Ratio, odnosno unakrsni odnos šansi, parametra koji se u epidemiologiji svakodnevno koristi. Odnos šansi (OR) dobićemo na sledeći način:

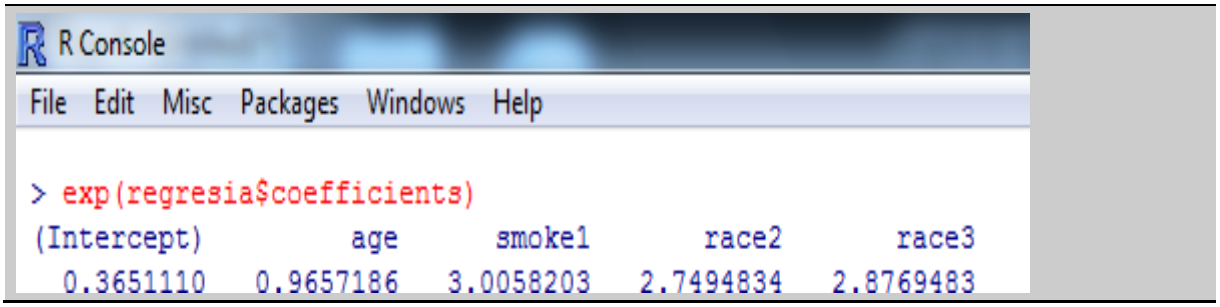
$OR_{\beta 1} = e^{\beta 1}$

Odnos šansi je jednak *e* podignutim na kvadrat odgovarajućim koeficijentom regresije. Na sličan način izračunamo OR za sve koeficijente regresije u tabeli. U programu R za ovo možemo da koristimo komandu *exp*. Kad bismo želeli da izračunamo OR iz naših koeficijenata koji su zapisani u objektu „regresija“ (pogledaj sliku gore), primenjujemo komandu *exp* na sledeći način:

```
exp(regresija$coefficients)
```

Rezultat će biti sledeće vrednosti OR (Tabela 3).

Tabela 3: Koeficijenti regresije

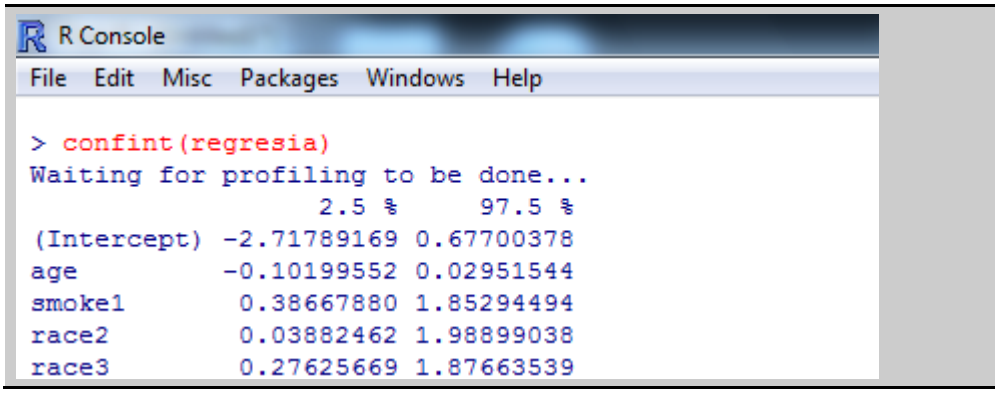


U interpretaciji koeficijenata regresije, odnosno OR, moramo da vodimo računa ne samo sa njihovim tačkastim ocenama, koje smo dobili korišćenjem logističke regresije, već su nam potrebne i njihove intervalske ocene, izražene na osnovu intervala poverenja. Budući da ih sama komanda *glm()* neće izračunati automatski, moramo da koristimo dodatnu komandu *confint()*. Nju ćemo primeniti na objekat „regresija“, napravljen u prethodnom koraku, koji sadrži rezultate logističke regresivne analize. Zadaćemo ovaj prikaz:

```
confint(regresija)
```

a rezultat su intervali poverenja za koeficijente regresije pojedinih nezavisnih promenljivih (Tabela 4).

Tabela 4: Intervali poverenja za koeficijente regresije



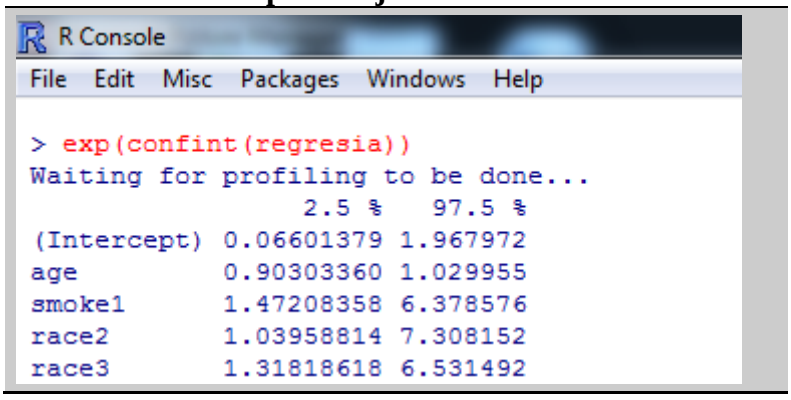
Data tabela sadrži kolone, pri čemi prva sadrži nazive promenljivih, druga sadrži donju granicu intervala poverenja, a treća kolona gornju granicu.

Za izračunavanje intervala poverenja za OR koristićemo postupak kao pri transformisanju koeficijenata regresije za OR, dakle

exp(confint(regresija))

i dobićemo sledeću tabelu sa nazivom promenljivih i pripadajućim gornjim i donji granicama intervala poverenja, (Tabela 5).

Tabela 5: Intervali poverenja za unakrsni odnos šansi OR



Za zbirno zapisivanje svih koeficijenata regresije, pripadajućim OR i intervala poverenja oko OR, u našem konkretnom slučaju koristićemo sledeće funkcije u R:

```
sum.coef<-summary(regresija)$coef
est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+1.96*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-1.96*sum.coef[,2])
summary(regresija)
```

cbind(regresija$coef,est,lower.ci,upper.ci)

Rezultat ovog uzorka je sledeća pregledna tabela, koja predstavlja ključne informacije rezultata (prva kolona prikazuje nazive promenljivih, druga kolona koeficijente regresije, kolona označena „est“ sadrži vrednosti OR, a kolona označena „lower.ci“ i „upper.ci“ sadrži donje, odnosno gornje granice intervala poverelja OR), (Tabela 6).

Tabela 6: Prikazivanje rezultata u jednoj tabeli

		est	lower.ci	upper.ci
(Intercept)	-1.00755379	0.3651110	0.06744702	1.976456
age	-0.03488277	0.9657186	0.90451953	1.031058
smoke1	1.10055049	3.0058203	1.44996262	6.231164
race2	1.01141304	2.7494834	1.04529923	7.232053
race3	1.05673010	2.8769483	1.29829507	6.375154

Interpretacija rezultata logističke regresije

Prilikom interpretacije rezultata logističke regresije jako je bitno biti svestan činjenice da logistička regresija analizira kompleksan odnos između skupa nezavisnih promenljivih i jedne zavisne promanljive. Omogućavanje ovakve kompleksne analize odnosa je razlog što je logistička regresija metoda izbora pri analizi uticaja više rizičnih faktora na jedno oboljenje - dakle prilikom tipične epidemiološke analize.

Praktičan značaj ima činjenica da logistička regresija omogućava analiziranje uticaj jedne nezavisne promenljive na zavisnu promenljivu, pri čemu uzimamo u obzir (izvagamo) uticaj drugih nezavisnih promenljivih koje, i kada nisu primarni predmet našeg interesovanja, mogu da budu sa zavisnom promenljivom statistički značajno povezane. Time što ćemo ove nezavisne promenljive uključiti u našu analizu, „odfiltriraćemo“ njihov skriveni uticaj i više ćemo se približiti stvarnom izražavanju odnosa između nezavisne promanljive čiji uticaj nas primarno zanima i između zavisne promenljive. Na ovakav način radimo takozvanu izvaganu analizu i dobijamo izvagano OR (u statističkoj i stručnoj literaturi se ovakav OR obično označava kao „adjusted OR“). Bitno je biti svestan da ako u uzorak nezavisnih promenljivih dodamo sledeću nezavisnu promenljivu (ili više nezavisnih promenljivih odjednom), ili ako iz uzorka nezavisnih promenljivih jednu ili više nezavisnih promenljivih izbacimo, mogu da se promene koeficijenti regresije i statistička značajnost njihovog uticaja.

Ako se vratimo na naš primer sa prediktorima niske porođajne mase, onda OR u tabeli za pojedine nezavisne možemo da interpretiramo rečima za pojedine promenljive. Nezavisna promenljiva „age“ potvrđuje da kada se poveća starost majke za jedinicu, snižava se šansa da će se njeno dete roditi sa niskom porođajnom masom 0,96 puta, sa pretpostavkom da ostale nezavisne promenljive ostanu konstantne. Na osnovu toga što interval poverenja uključuje i broj 1, odnosno da je p-vrednost, koja pripada koeficijentu regresije od kojeg je dato OR izvedeno, veća od 0,05, ovaj uticaj nije statistički signifikantan.

Nezavisna promenljiva promenljiva „smoke“ je kategorijalna, pa se njena interpretacija pomalo razlikuje. Setimo se takozvane referentne kategorije, o kojoj smo govorili u prethodnim delovima poglavlja. Ponovimo da je uticaj pojedinih kategorija date nezavisne promenljive relativan prema referentnoj kategoriji. U slučaju ove promenljive referentna kategorija je označena u bazi podataka brojčanim kodom „0“, odnosno majka-nepušač. OR pripadajuće kategorije označene

brojčanim kodom „1“, dakle „majka pušač“, interpretiramo na sledeći način: ako majka na osnovu promenljive „smoke“ spada u kategoriju pušača, u poređenju sa kategorijom nepušača u našoj bazi podataka, šansa da će se njeno dete roditi sa niskom porođajnom masom će se povećati se za 2,98 puta – sa pretpostavkom da ostale nezavisne promenljive ostanu konstantne. Na osnovu toga što interval poverenja uključuje broj 1, odnosno da p-vrednost pripadajuća regresivnom koeficijentu, iz kojeg je bio izveden dati OR, je manja od 0,05, ovaj uticaj je statistički signifikantan.

Nezavisna promaneljiva promenljiva „race“ je takođe kategorijalna, pa je interpretiramo slično kao promenljivu „smoke“. U slučaju ove promenljive, referentna kategorija je ona označena označena u bazi podataka brojčanim kodom „1“, odnoson majka kavkaske rase (referentna kategorija u tabeli nije prisutna). OR koji pripada kategoriji označenoj brojčanim kodom „2“, odnosno majka crnkinja, interpretiramo na sledeći način: ako majka na osnovu promenljive „race“ spada u kategoriju crnkinja, u poređenju sa kategorijom kavkaske rase u našoj bazi podataka, šansa da njeno dete bude rođeno sa niskom porođajnom masom se povećava 2,75 puta – sa pretpostavkom da ostale nezavisne promenljive ostanu konstantne. OR koje pripada kategoriji označenoj brojčanim kodom „3“, dakle majke druge rase, interpretiramo na sledeći način: ako majka na osnovu promenljive „race“ spada u kategoriju druge rase, u poređenju sa kategorijom majki kavkaske rase u našoj bazi podataka, šansa da dete bude rođeno sa niskom porođajnom masom povećava se 2,88 puta – sa pretpostavkom da ostale nezavisne promenljive ostanu konstantne.

Na osnovu činjenice da interval poverenja u slučaju kategorije „2“ i „3“ ne uključuje broj 1, odnosno da p-vrednost pripada koeficijentima regresije, od kojih su bili izvedeni dati OR, je <0,05, ovi uticaji su statistički signifikantni.

Primena logističkog modela na predikciju

Nakon što smo kompleksno analizirali odnose između uzorka nezavisnih promenljivih i zavisne promenljive u našem uzorku, sada ćemo govoriti o primemi logističke regresije za predikciju verovatnoće da će nastati posmatrana pojava. Ako ostanemo kod našeg primera, onda govorimo o predikciji da se majci sa određenom kombinacijom posmatranih nezavisnih promenljivih (prediktora) rodi dete sa normalnom ili sa niskom porođajnom masom.

Analiziraćemo konkretan primer majke, kod koje smo saznali sledeće vrednosti prediktora:

X_1, dakle „age“ = 21 godina

X_2, dakle „smoke“ = “1“ ili „pušač“
X_3, dakle „race“ = “3“ ili „druga rasa “

Za predikciju koristimo nama već poznatu komandu logistčkog modela

$f(z)=1/1+e^{-z}$

Kako bismo odredili verovatnoću pojave f(z), što je u ovom slučaju niska porođajna masa izražena brojem „1“ u okviru promenljive „low“, moramo da izračunamo vrednost Z i ubacimo je u ovaj odnos. Vrednost Z dobijemo nama već poznatim načinom pomoću odnosa:

$Z= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, gde je
β_0 = vrednost koeficijenta regresije za intecept,
β_1= vrednost koeficijenta regresije za promenljivu „age“,
β_2= vrednost koeficijenta regresije za kategoriju promenljive „smoke1“, budući da je reč o pušaču,
β_3= vrednost koeficijenta regresije za kategoriju promenljive „race3“, budući da je reč o majci druge rase,

i zato je
$Z= -1.00755 + (-0.03488) X_1 + 1.10055 X_2 + 1.05673 X_3$

gde je posle umetanja vrednosti X_1, odnosno utrasta:
$Z= -1.00755 + (-0.03488) * 21 + 1.10055 X_2 + 1.05673 X_3$
$Z= 0.42$

Posle zamene vrednosti Z u odnos
$f(z)=1/1+e^{-z}$

dobijamo
$f(z)=1/1+e^{-(0,42)}$

a posle
f(z)=0,6 , odnosno
f(z)=60%.

Kod 21-godišnje majke – pušača „druge rase“, verovatnoća rođenja deteta sa niskom porođajnom masom prema modelu nepravljenom u našem uzorku je 60%.

Validnost i dijagnostika logističkog modela

Dijagnostika ili vrednovanje modela je skup naziva za kompleksne pokazatelje koji govore o tome koliko je validan model koji smo napravili. Kako smo definisali dve vrste primene logističkog modela, tako ćemo sada da objasnimo koji atributi su bitni pri obe primene.

Signifikantnost uticaja prediktora na zavisnu promenljivu

Ovaj vid validnosti modela se pre svega odnosi na primenu logističke regresije na kompleksnu analizu odnosa između uzoraka nezavisnih promenljivih i zavisne promenljive. Odlučujući atribut za vrednovanje signifikantnosti uticaja pojedinih nezavisnih promenljivih na zavisnu promenljivu je rezultat Valdovog testa hipoteze o nezavisnosti promenljive od date nezavisne promenljive. Ako p-vrednost pripadajuća Valdovom testu za datu nezavisnu promenljivu ukazuje na odbacivanje ovakve hipoteze, data promenljiva ima statistički signifikantan uticaj na zavisnu promenljivu.

U optimalnom modelu sve obuhvaćene nezavisne promenljive imaju signifikantan uticaj na zavisnu promenljivu. Obično se, međutim, srećemo sa slučajem gde su u modelu obuhvaćene i

nezavisne promenljive sa statističkim nesignifikantnim uticajem. U ovakvom slučaju potrebno je model promeniti izbacivanjem, odnosno dodavanjem nezavisnih promenljivih sve dok ne postignemo takav model, gde će sve nezavisne promenljive imati statistički signifikantan uticaj.

Za traženje optimalnog modela možemo da iskoristimo tri načina. Prvi je takozvana napredna selekcija, gde pravljenje modela počinjemo univarijantnom analizom uticaja pojedinih nezavisnih promenljivih na zavisnu promenljivu, a do konačnog modela obuhvatimo one promenljive kod kojih se sazna statistički signifikantni uticaj pri univarijantnoj analizi. Drugi način je takozvana povezana selekcija, prilikom koje najpre u modelu obuhvatimo sve nezavisne promenljive, koje imamo na raspolaganju. Naknadno iz nje izbacimo promenljive sa nesignifikantnim uticajem i ponovimo analizu. Ovu proceduru ponavljamo sve dok ne dobijemo model gde sve nezavisne promenljive imaju signifikantni uticaj na zavisnu promenljivu. Treći način je kombinacija elemenata napredne i povezane selekcije.

Ako se vratimo na naš zadatak, model koji smo napravili sadrži tri nezavisne promenljive, od kojih jedna promenljiva „age“, ima nesignifikantan uticaj. Za optimalizaciju modela radili bismo sledeće:

Naš prvobitan model sadržao je tri nezavisne promenljive: „age“, „smoke“ i „race“.

U programu R smo komandom:

```
regresija  = glm (low ~ age+ smoke+ race, family=binomial, data=birthwt)
summary (regresija)
```

dobili model sa sledećim parametrima (Tabela 7).

Tabela 7: Koeficijenti regresije i njihov značaj

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.00755    0.86166  -1.169  0.24228
age         -0.03488    0.03340  -1.044  0.29634
smoke1       1.10055    0.37195   2.959  0.00309 **
race2        1.01141    0.49342   2.050  0.04039 *
race3        1.05673    0.40596   2.603  0.00924 **
---
```

Na osnovu Valdovog testa se, dakle, nije pokazao statistički signifikantan uticaj promenljive „age“, zato ćemo koristiti metodu povezane selekcije i izrazićemo ovu nezavisnu promenljivu iz modela. Onda analizu ponovimo pomoću sledeće komande:

```
regresija  = glm (low ~ smoke + race, family=binomial, data=birthwt)
summary (regresija)
```

i dobijemo model sa novi m parametrima (Tabela 8).

Tabela 8: Novi parametri posle povezane selekcije

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8405     0.3529  -5.216 1.83e-07 ***
smoke1        1.1160     0.3692   3.023  0.00251 **
race2         1.0841     0.4900   2.212  0.02693 *
race3         1.1086     0.4003   2.769  0.00562 **
---
```

Iz navedene tabele proizilazi da je uticaj obe nezavisne promenljive koje su u modelu statistički signifikantan. Obratite pažnju na ključnu pojavu: posle izbacivanja jedne od promenljivih i ponavljanja analize, vrednost i signifikantnost uticaja promenljivih, koje su ostale u modelu, se zamenila.

Validnost modela kao celine

Ova vrsta validnosti se pre svega odnosni na primenu logističke regresije na predikciju. Logistički model ovde vrednujemo kao celinu, za razliku od individualnog vrednovanja zavisnih promenljivih, koje smo spomenuli u prethodnom delu. Problematika prognostičkog modelovanja i dijagnostičkih modela je kompleksan problem, koji prevazilazi ciljeve ove publikacije, pa ćemo zato govoriti samo o nekoliko osnovnih aspekata.

Osnovni atributi validnosti prognostičkog modela su AUC (Area Under the Curve), odnosno površina ispod krive ROC i Nagelkerkeovog R^2. AUC, ponekad označen i sa C, je vrednost koja se uvek nalazi u rasponu od 0,5 do 1. Vrednost AUC=1 ukazuje na model sa visokom diskriminativnom sposobnošću, odnosno sposobnošću da ispravno odredi obolele i zdrave. Vrednost AUC=0,5 ukazuje na model sa praktično nultom diskriminativnom sposobnošću. Ovaj parametar proizilazi iz specifičnosti i senzitivnošću modela, odnosno sposobnosti predviđanja stvarno pozitivnih i stvarno negativnih slučajeva.

Nagelkerkeov R^2 je jedan od pseudokoeficijenata determinacije. Njegov značaj je sličan značaju stvarnog koeficijenta determinacije, međutim metod njegovog izračunavanja je drugačiji (kada pri logističkoj regresiji nije reč o linearnoj zavisnosti, nije moguće odrediti koeficijent determinacije, kako ga poznajemo u linearnoj regresiji – zato i ima naziv pseudokoeficijent determinacije). Ovde govorimo o parametru koji dobija vrednost od 0 do 1. Što se vrednost više približava 1, time je model validniji. Izračunamo vrednost AUC i Nagelkerkeovog R^2 za model u našem primeru.

Oba navedena parametra, kao i sledeće parametre korišćene pri prognostičkom modelovanju (što nije predmet ove publikacije), dobijamo, na primer, korišćenjem komande *lrm* iz paketa *Design* u programu R. Komanda ima sledeći oblik:

```
library(Design)
data(birthwt)
lrm (low ~ smoke + race, data=birthwt)
```

Rezultat su odgovarajući parametri prikazani u Tabeli 9.

Tabela u donjem delu se slaže sa onom koju smo dobili komandom *glm*. Iznad nje se nalazi nekoliko sledećih parametara koji se odnose na model, među kojima je i AUC (u ovom slučaju označen kao C) i Nagelkerkeov R^2, koji je u ovom slučaju izražen kao R^2.

U našem zadatku je C=0,65, što govori o umerenoj do dobroj diskriminatornoj sposobnosti modela. Model koji smo napravili će umereno do dobro dokazati i ispravno odrediti da li će se dete određene majke roditi sa normalnom ili sa niskom porođajnom masom. R^2 ima u ovom konkretnom slučaju 0,105, što je vrednost relativno bliska nuli. Ovo ukazuje na nisku validnost modela. Nagelkerkeov R^2 ne bi trebalo da se uzima za odlučujući kriterijum prilikom vrednovanja validnosti modela, već pre kao podržavajuća informacija za interpretaciju.

Tabela 9: Rezultat dijagnostičkog modela

```
Logistic Regression Model

lrm(formula = low ~ smoke + race, data = birthwt)


Frequencies of Responses
  0   1
130  59

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy
       189      1e-10       14.7          3     0.0021       0.65       0.29
     Gamma      Tau-a         R2      Brier
     0.378      0.129      0.105        0.2

          Coef   S.E.   Wald Z P
Intercept -1.841 0.3529 -5.22  0.0000
smoke=1    1.116 0.3692  3.02  0.0025
race=2     1.084 0.4900  2.21  0.0269
race=3     1.109 0.4003  2.77  0.0056
```

Pridruženost (confounding) i interakcije

Jedna od ključnih uloga epidemiologije je sposobnost da se objasni odnos između određenih definisanih faktora i zdravstvenog efekta. Kada bismo odnos koji želimo da objasnimo postavili kao jednostavan kauzalni odnos između izloženosti faktoru rizika R i efekta E, morali bi da uzmemo u obzir kompleks sledećih faktora i okolnosti, koji na ovaj jednostavno postavljen odnos mogu da utiču i menjaju na različite načine. U epidemiologiji definišemo dva osnovna tipa ovakvih uticaja. Govorimo o pridruženosti i interakciji.

Pridruženost

Pridruženost bismo mogli jednostavno da definišemo kao kombinaciju uticaja rizičnog faktora, čiji uticaj na zdravstveni efekat opisujemo sa efektom drugih rizičnih faktora na nama posmatran efekat. Ova premisa proizilazi iz kompleksnosti odnosa u prirodi i iz činjenice da ni

jedan zdravstveni efekat nije prouzrokovan isključivo uticajem jednog faktora, nego može da bude rezultat uticaja više faktora. U slučaju pridruženosti razmišljamo o efektu faktora rizika R1 na zdravstveni efekat E i o efektu faktora rizika R2 na taj isti zdravstveni efekat E kao o dva moguća načina pojave posmatranog zdravstvenog stanja.

Interakcije

Interakcije možemo da opišemo kao uticaj faktora R2, koji nije neophodno primarni predmet našeg interesovanja, na efekat proučavanog faktora rizika R1 na posmatrani zdravstveni efekat E. Dakle, radi se o modifikaciji određenog faktora na posmatrani zdravstveni efekat drugim faktorom. Međutim, ujedno dati faktor R2 može da bude u određenoj kauzalnoj vezi sa posmatranim zdravstvenim efektom. Ovde govorimo o međunezavisnom delovanju, gde faktor R2 sam po sebi utiče na zdravstveni efekat i ujedno je u interakciji sa faktorom R1. Promena faktora R2 može da izazove promenu uticaja faktora R1 na zdravstveni efekat E.

Kako izaći na kraj sa pridruženošću i interakcijama

Primarno bi sve snage na eliminisanje uticaja na moguće pridruženosti i interakcije trebalo koncentrisati na fazu dizajna epidemiološke studije. Posle definisanja kauzalnog odnosa, koji je u studiji primarni predmet našeg interesovanja, trebali bismo da se trudimo da identifikujemo faktore koji mogu u datom slučaju da deluju kao pridruženosti ili da izazovu interakcije. Dobar način kako da se dizajnom epidemiološke studije minimiziraju ovakvi uticaji je randomizovan izbor iz populacije ili mečovanje (sparivanje). Obe ove procedure su predmet bazične epidemiologije.

U praksi, međutim, nije moguće sprečiti pojavu pridruženosti i interakcije samo dizajnom studije, tako da je potrebno posvetiti im pažnju i u fazi statističke analize podataka. Najbolja strategija kako analizirati pridruženost i interakcije je stratifikovana statistička analiza i upotreba specifičnih statističkih testova (Mantel-Hencelove procedure i sl.).

Koncept pridruženosti i interakcije detaljno je prodiskutovan u mnogim udžbenicima epidemilogije, a opis statističkih procedura, koje interakcije i pridruženosti analiziraju je kompleksan i nije predmet ovog udžbenika. Faktori spomenuti u ovom poglavlju trebalo bi da budu odgovorno izvagani i uzeti u obzir prilikom bilo kojeg postupka multivarijantne analize, uključujući i logističku regresiju.

Vežbe

1. Prema zadacima i uputstvima u ovom poglavlju napravite sledeću analizu: analizirajte na koji način oboljenje mokraćne bešike, uzrast, hipertenzija, gojaznost i trajanje estrogene terapije potenciraju razvoj endomerijalnog karcinoma. Izračunajte pojedine koeficijente regresije, unakrsni odnos šansi (OR) i njihov interval poverenja. Pokušajte da napravite optimalan model za predikciju endometrijalnog karcinoma na osnovu dostupnih promenljivih i interpretirajte da li je uticaj pojedinih prediktora u modelu signifikantan. Vrednujte model kao celinu. Koristite datoteku „*bdendo*“ u okviru paketa „*Epi*“ u programu R.

2. Razmislite o pridruženostima i interakcijama u okviru prethodnog zadatka o endometrijalnom karcinomu. Pokušajte da predložite optimalan dizajn epidemiološke studije za saznanje datih faktora za razvoj endometrijalnog karcinoma.

DESETO POGLAVLJE

Hi-kvadrat i neparametarske statistike

Sadržaj poglavlja

Cilj poglavlja

Do sada smo govorili o statističkom pristupu koji se odnosio na izmerene podatke, a oni su bili normalno raspodeljeni, odnosno približno normalno raspodeljeni. Ali, u svakodnevnom životu, u kliničkim i epidemiološkim studijama, često se srećemo sa podacima u numeričkom obliku, u obliku učestalosti (broj oboljenja, broj slučajeva sa simptomima i bez njih, broj pušača i slično). Srećemo se i sa situacijom gde izvorno izmerene kvantitativne podatke (uzrast, nivo šećera – glikemije u krvi) menjamo u kategorije i zanima nas zastupanje u ovim kategorijama (starosne kategorije, glikemija povišena i dotična vrednost). Često se srećemo sa situacijom kada uslove normalne raspodele iz raznih razloga ne možemo da ispunimo, naročito usled malog uzorka. U ovom poglavlju bavićemo se ovakvom vrstom problema i navešćemo osnovne postupke prilikom njihovog rešenja i interpretacije.

Tabela 1: Ciljevi poglavlja

1. Objašnjenje osobina hi-kvadrat raspodele
2. Primena hi-kvadrat testa za 2x2 i m puta n tabele
3. Osnovi neparametrijski testovi

Osobine hi-kvadrat raspodele

Hi-kvadrat raspodelu (χ^2) možemo da izvedemo iz normalne raspodele. Za svaki stepen slobode raspodela ima drugačiji umeren oblik krive (Slika 1).

Obratite pažnju da je oblik raspodele za prva dva stepena slobode značajno različit od ostalih.

Hi kvadrat test i tabele kontigencije

Tabele kontigencije (contingency table) omogućavaju tabelarnu ukrštenu klasifikaciju podataka. Predstavljaju kombinaciju dve (ili više) tabela frekvencije, tako da svaka unutrašnja ćelija predstavlja jednoznačajnu kombinaciju specifičnih vrednosti (ovde nazivanu i kategorija) ukršteno tabelarnih promenljivih. Dakle, omogućava saznanje frekencije, broja ispitanika, koji odgovaraju specifičnoj kategoriji za više od jedne promenljive. Ispitivanje ovih frekvencija omogućava saznanje relacija, odnosa između promenljivih. Tabela kontigencije odgovara samo nominalnim promenljivim ili brojčanim promenljivim, koje dostižu relativno mali broj skupova vrednosti. 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). Najjednostavnija tabela kontigencije je ona koja ima dve kolone i dva reda (Tabela 1). Dodavanjem kolona i redova tabela se komplikuje i govorimo o *r x c* ili *m x n* tabelama. Slova *r* ili *m* označavaju broj redova, a slova *c* ili *n* govore o broju kolona.

Slika 1: Hi kvadrat raspodela za različite stepene slobode

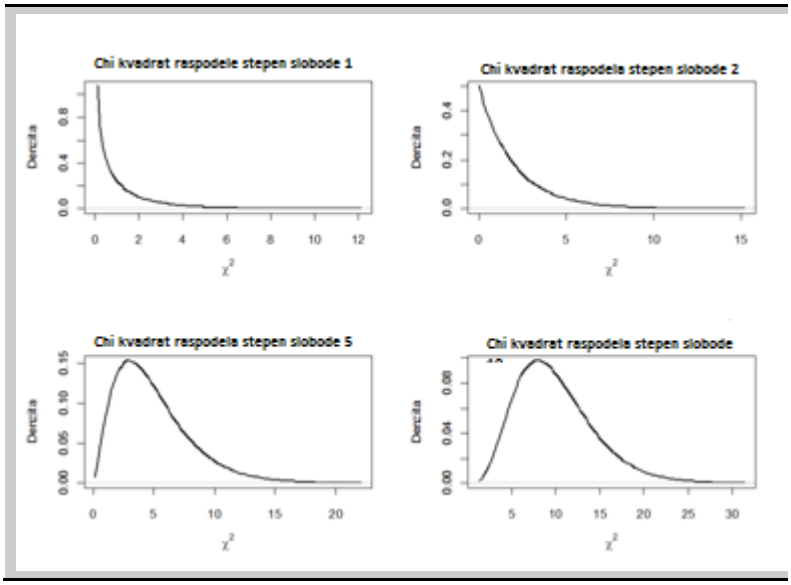


Tabela 7: Tabela kontigencije 2x2

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

Hi-kvadrat test jes zasnovan na upoređivanju testiranih podataka koji imaju hi-kvadrat raspodelu. Najčešće korišćeni testovi služe za saznanje da li su distribucije – raspodele poreklom iz dve ili više populacija uzajamno različite. Ovi testovi obično koriste zbirne podatke (podatke o kvalitativnim vrednostima). Može se raditi o upoređivanju uzoraka iz posmatrane populacije sa teorijski očekivanom distribucijom.

2 x 2 tabele

Najjednostavniji oblik tabele kontigencije je 2x2 tabela, gde su obe promenljive binarne, dajući samo dve moguće vrednosti, postavljene tako da za potkategorije jedne karakteristike uvode horizontalne (u redovima -r), a potkategorije druge karaktristike se navode vertikalno (u kolonama -c). Testovi zavisnosti među karakteristikama u kolonama i u redovima mogu se uraditi onda pomoću hi-kvadrat testa. Konstrukciju tabele kontigencije pokazaćemo na primeru (Primer 1). Tabela ima dva reda i dve kolone, obe promenljive su binarne, dakle tipa ili/ili. Ili je imao učesnik nesreće na glavi kacigu, ili je nije imao i ili je završio u komi ili bez kome. To su slučajevi koji se uzajamno isključuju, dakle niko ne može da i ima i nema kacigu. Na drugoj strani, pogledajte da

četiri polja nude sve moguće kombinacije koje stvarno mogu da se dese: sa kacigom i bez kome, sa kacigom i sa komom, bez kacige i bez kome, bez kacige i sa komom.

Primer 1: Primer pravljenja tabele kontigencije

Primer: U studiji uzroka povreda tražio se učinak zaštitne kacige kod motociklista, koji su imali ozbiljnu saobraćajnu nezgodu. Saznalo se da od 254 učesnika nezgoda na motorima kacigu je koristilo 176, a od njih je preživelo ozbiljnu povredu praćenu komom 23, dok oje onih koji nisu koristili kacigu 56 osoba posle nezgode ostalo u komi. Da li je moguće reći da kaciga štiti učesnike od ozbiljnih povreda mozga praćenih komom?

Tabela kontigencije: Prema gore navedenom slikovitom primeru i na osnovu zadavanja zabeležićemo poznate podatke u tabelu:

	Bez kome	Sa komom	**Σ**
Kaciga		23	176
Bez kacige		56	
Σ			254

Dopunićemo računanje u redovima i kolonama i napravićemo tabelu kontigencije:

	Bez kome	Sa komom	**Σ**
Kaciga	153	23	176
Bez kacige	22	56	78
Σ	175	79	254

Iz ovako napravljene tabele još ne znamo odgovoriti na pitanje istraživača.

Zbir prvog reda govori koliko je bilo svih onih koji su nosili kacigu, a zbir drugog reda 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. Često je pitanje da li je pri konstrukciji tabele potrebno pridržavanje nekog redosleda, šta treba da bude u redu, a šta u koloni. Takvo pravilo ne postoji, jedino je potrebno obratiti pažnju redosleda prilikom interpretacije rezultata.

Vratimo se pitanju kako ćemo da saznamo da li kaciga stvarno čuva onog koje je nosi od kome prilikom udesa na motoru. Počinjemo sa bilansom kakva je verovatnoća da učesnik nesreće neće imati kacigu. Rezultat navedenog predstavlja odnosom onih koji su imali kacigu prema ukupnom broju. Iz tabele sa kompleksnim zadavanjem (Primer 1) saznaćemo da je to broj 176/254 = 0,69. Verovatnoća ne upasti u komu posle povrede je analogna broju svih koji nisu imali komu prema svima u uzorku, dakle 175/254 = 0,69 (razlika je tako mala, da se neće prikazati na dve decimale). Postavićemo pitanje kakva je verovatnoća nositi kacigu i ne pasti u komu. U srednjoj

školi se uči da je kombinacija dve verovatnoće jednaka njihovom proizvodu, dakle u našem slučaju to bi bilo 0,69 x 0,69 = 0,48. Koliko je to slučajeva iz celog uzorka saznaćemo tako što broj slučajeva pomnožimo verovatnoćom nositi kacigu i ne upasti u komu, odnosno 254 x 0,48 = 121,26. Dakle, ako bi važile izračunate verovatnoće, imali bismo u prvom kvadratu broj 121,26 slučajeva, a ne posmatranih 153 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. Ako zadržimo oznake iz tabele 1, onda možemo da napišemo formulu izračunavanja očekivane vrednosti za prvu ćeliju, označenu kao

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

Ovo pravilo uopštićemo za sve ćelije: očekivana vrednost ćelije jednaka je količniku proizvoda zbira reda i zbira kolone sa ukupnim zbirom.

Kako bi ustanovili da li postoji saglasnost između opaženog (Observed -O) i očekivanog (Expected - E), u praksi se koristi saznanje gospodina Pirsona, da hi-kvadrat raspodelu možemo da koristimo kao test slaganja između posmatranja i hipoteze. Izračunaćemo vrednost hi-kvadrat tako što ćemo izračunati koren razlika opažene i očekivane vrednosti u svakoj ćeliji u tabeli, pa ovaj podeliti vrednoššu očekivanom u datoj ćeliji, pa rezultate onda podelimo.

$$X^2 = \frac{\sum \left(O_i - E_i\right)^2}{E_i}$$

Broj stepena slobode se izračunava kao proizvod broja redova smanjen za jedan i broja kolona, isto tako smanjenih za 1.

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

(ako označimo broj redova *r,* broj kolona *k* i broj stepena slobode *df*).

Hi-kvadrat onda 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 verovatnoću, odbacujemo hipotezu o nezavisnosti.

Primer 2: Izračunavanje hi-kvadrata u tabeli 2x2

```
> kaciga <-c(153, 23, 22,56) # ulazni podaci svrstani po redovima
> x <- matrix(kaciga, nrow=2, byrow=T) # transformacija na matricu i njeno ispisivanje
> x
 [,1] [,2]
[1,] 153 23
[2,] 22 56
# dozivanje komande prop.test()
> prop.test(x, alternative = c("two.sided"), conf.level = 0.95, correct = FALSE)
```

2-sample test for equality of proportions without continuity correction

data: x
X-squared = 86.9855, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
0.4756761 0.6988577
sample estimates:
prop 1 prop 2
0.8693182 0.2820513

Rezultat testiranja hipoteze na nivou značajnosti 0,95 i 0,99 (p< 0,05 kao i < 0,01) nam omogućava odbacivanje hipoteze o nezavisnosti. Rezultat hi kvadrat testa je dovoljno visok (86,98), šta govori da posmatrana veza nije samo uticaj slučaja. Dakle, naša stručna interpretacija glasi da je očigledan uticaj kacige na pojavu kome pri teškim povredama mozga.

Prilikom određivanja testa nismo morali da koristimo Jejtsovu korekciju (imali smo dovoljno velike brojeve u tabeli kontigencije). Kad bismo ipak želeli da je koristimo (preporučuje se u slučaju kad je vrednost u polju manja od 5) morali bi da zadamo *correct = Yes.* Razlika u izračunavanju sa korišćenjem ove korekcije u našem slučaju neće biti dramatična.

m x n tabele

Situacija u kojoj tabela kontigencije ima više od dve kolone ili dva reda nije retka. Onda pojednostavimo postupak iz slučaja dve kolone i dva reda na proizvoljan broj kolona i redova. Postupak je sličan kao i u prethodnom slučaju, s tom razlikom da moramo da koristimo drugu proceduru – postupak za izračunavanje rezultata testa (Primer 3).

Primer 3: m x n kontingentna tabela i Pirsonov hi-kvadrat test

Zadavanje: Niska porođajna masa je definisana kao masa novorođenčeta manja od 2500g. Istraživač je ispitivao zavisnost porođajne mase od pušenja majke. Imao je tri grupe: nepušačice, pušačice i bivše pušače. Zapitao se da li pušačka navika utiče na porođajnu masu novorođenčeta.

	Nepušačica	Bivša pušačica	Pušačica
Porođajna masa < 2500g	140	153	27
Porođajna masa ≥ 2500 g	2197	1510	433

Rešenje

```
> masa <- c(140, 153, 27, 2197, 1510, 433)
> masa<- matrix(masa, ncol=2) # transformacija na matricu sa dve kolone
> masa
 [,1] [,2]
```

```
[1,] 140 2197
[2,] 153 1510
[3,] 27 433
> chisq.test(masa) # test

 Pearson's Chi-squared test

data: masa
X-squared = 16.3411, df = 2, p-value = 0.0002829
```

Interpretacija: 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).

Neparametrijski testovi

Kako smo već naveli, neparametrijski testovi su postupci koji ne pretpostavljanju da su podaci normalno raspodeljeni. Uopšteno gledano, ne pretpostavljaju poznavanje raspodele u populaciji. Zato se obično označavaju i nazivom statistički testovi bez raspodele (distribution-free statistics). Neparametrijski testovi imaju dve bitne karakteristike, kojima se razlikuju od parametrijskih: ne potvrđuju hipotezu o parametrima populacije, a ujedno ih možemo koristiti i u slučajevima kada je oblik populacije, iz koje potiče uzorak, poznat. Njihova zasigurna prednost je njihova dokazana sposobnost prilikom ispitivanja podataka koji govore o rezultatima merenja na nižim skalama – vrednuje se redosled. Obično se još spominje prednost jednostavnijeg računanja, šta danas, prilikom korišćenja kompjutera, nije tako bitno. Navedeni postupci imaju i negativnu stranu, a to je njihova niža snaga. Zato u slučaju kada je moguće da biramo između parametrijskog i neparametrijskog testa, a ujedno nam to omogućavaju uslovi, odlučićemo se za parametrijske.

Skale

Pre nego što pređemo na pojedine neparametrijske statističke testove, moramo bliže da popričamo o skalama, naročito zato što su neparametrijski testovi pogodan instrument za obradu podataka dobijenih na takozvanim nižim skalama merenja.

Skale merenja nastaju na taj način što određenom objektu ili dešavanju priključimo broj na osnovu skupa pravila. Tako nastaju skale koje je moguće podeliti u četiri osnovne vrste: nominalna, ordinalna, intervalna i odnosna (racio) skala.

Najjednostavnija je **nominalna skala**, koja nastaje priključivanjem imena posmatranja. Ime može da bude i broj, odnosno cifra. Primer je Međunarodna klasifikacija bolesti, koja svakoj dijagnozi priključuje jedinstvenu kombinaciju slova i brojeva. Svakodnevne su i dihotomne skale, na primer muškarac - žena, zdrav – bolestan, pušač – nepušač. Naravno, podela može da bude i više, na primer pušač – bivši pušač – nepušač. Pritom granice između kategorija ne moraju da budu ravnomerne i uopšte ne moraju kvantitativno da se upoređuju u smislu veći ili manji.

Ordinalna ili **redna skala** omogućava svrstavanje objekata prema tome koje imaju više, a koje manje kvaliteta. Za razliku od intervalskih promenljivih, one nam ne omogućavaju da kažemo za koliko je to više kvaliteta. Primer je rezultat lečenja: izlečeni bez posledica, izlečeni sa trajnim

posledicama, neizlečeni, umrli. Skale intenziteta mogu da budu određene kao nepušač – povremen pušač – redovan umeren pušač – redovan strastven pušač. Pritom granice između kategorija ne moraju da budu ravnomerne, ali ih već možemo kvantitativno upoređivati u smislu više ili manje (strastveni pušač je eksponiran većom dozom nikotina nego umereni ili povremeni pušač).

Ovu skalu koristimo prilikom 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 na taj način što je poznata udaljenost između dva merenja skale. Tipičan primer ovakve skale je skala merenja temperature. Celzijusova skala počinje sa nulom, šta je dogovorena vrednost, isto kao 100 stepeni. Tačke između njih su rastuće temperature, a njihova udaljenost je jednaka i jedinica je jedan stepen Celzijusa. Budući da postoje vrednosti manje od 0°C, znamo da i ova vrednost sadrži određen kvantitet i nije krajnja tačka na skali.

Odnosne skale su najkompleksnije skale. Sadrže sve karakteristike nominalne, ordinalne i intervalne skale. Sastoje se, dakle, 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 70-godišnji čovek tačno dva puta stariji 35-godišnjaka. Slično je i kod merenja telesne mase i visine. U ovako napravljenim skalama obično možemo da koristimo parametrijske testove. Neparametrijske testove smo prinuđeni da koristimo ako imamo na raspolaganju male uzorke (n je manje od 30, odnosno 20) i ne možemo da obezbedimo pretpostavku da podaci dolaze iz normalno distibuirane populacije.

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.

Znakovni test

Jedan od prvih statističkih testova koje smo prikazali bio je t-test za potvrđivanje hipoteze o aritmetičkoj sredini populacije. Potvrđivali smo da je aritmetička sredina populacije jednaka, manja ili veća od određene vrednosti i da je razlika dve aritmetičke sredine jednaka nuli, odnosno da su aritmetičke sredine jednake. Više puta smo naglasili da rezultati ovog testa mogu da budu tačni samo onda kada je bila ispoštovana pretpostavka normalne raspodele uzorka. Takođe smo naveli da, što je manji broj merenja u uzorku, to je pridržavanje ovog pravila sumnjivije. Šta onda raditi u slučaju kad je uzorak mali, a raspodela ne pokazuje da se radi o Gausovoj, odnosno merenja su pravljena na intervalnoj ili odnosnoj skali? U tom slučaju moramo da posegnemo za neparametrijskim testom, koji ne testira ni aritmetičku sredinu ni varijansu.

U ove testove spada i znakovni test, koji se ne bavi aritmetičkom sredinom, već medijanom kao testom centralne tendencije. Znamo da u slučaju normalne raspodele populacije (ili Studentove raspodele) vrednost medijane i aritmetičke sredine je jednaka. Pri drugačijoj raspodeli mogu se međusobno znatno razlikovati. Jedini uslov korišćenja ovog testa je uslov povezanosti potvrđivanja promenljive. Ovaj test se zove znakovni zato što se umesto brojeva koriste znakovi + i -. 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.

U prvom slučaju istraživač se pitao da li medijana skora jednaka određenoj vrednosti

(Primer 4).

Primer 4: Primer za izračunavanje znakovnog testa

Zadavanje: Na odeljenju intenzivne nege hospitalizovani su bolesnici sa različitim nivoom poremećaja svesti, merenim Glazgovskom skalom kome (GCS) sa rezultatima navedenim u tabeli. Istraživači su se pitali da li je moguće reći da je medijana skora 9.

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

Ovde je reč o skoru, kao i o malom broju poređenja na spojenoj skali. Zato nije opravdano da koristimo t-test, već moramo da tražimo znakovni test. Želimo da potvrdimo 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. Slučaj jednakosti označićemo sa 0 i dobićemo tabelu.

Tabela 3: Znakovi odstupanja GCS od potvrđivane vrednosti medijane 9

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

Vidimo da je broj znakova „+“ 5, a „–„ samo 3. Jednaka devetki su samo dva slučaja. Ako ne uzimamo u obzir merenja koja su jednaka potvrđivanoj vrednosti 9, onda imamo više merenja u plusu, nego u minusu. Pretpostavljamo, ako bi bilo tačno da je medijana jednaka 9, onda bi trebalo da bude isti broj odstupanja u smislu plus i minus. Prisetimo se načina izračunavanja medijane, kada svrstamo vrednosti prema veličini, a ona vrednost, koja se nakon podele nalazi tačno na polovini je medijana. Dakle, potvrđivana nulta hipoteza i alternativna hipoteza biće:

H_0 *median GCS = 9* H_A *median GCS* $\neq 9$

budući da pretpostavljamo da je broj pozitivnih i negativnih razlika od potvrđivane vrednosti jednak, možemo da hipoteze formulišemo ovako:

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

Nulta hipoteza govori da je verovatnoća pozitivnih i negativnih odstupanja jednaka, odnosno da je verovatnoća 0,5. Sledeći zaključak je zasnovan na potvrđivanju hipoteze na osnovu binometrijske podele. To će za nas napraviti program (Tabela 2), a nama ostane samo interpretacija

rezultata. Za izračunavanje ćemo koristiti komandu *sign.test()*, koja je deo paketa *BSDA*. Njega je neophodno najpre potražiti u *CRAN* i instalirati pre nego što postane moguće koristiti ovu komandu. Komanda zahteva više argumenata: *sign.test(x, y = NULL, md = 0, alternative = "two.sided", conf.level = 0.95).* Prilikom testa, gde utvrđujemo jednakost medijane merenja neke vrednosti, 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. Rezultat potvrđivanja našeg primera videćemo u sledećoj tabeli (Tabela 2).

Pozvali smo komandu *sign.test* sa podacima zamenjenim u promenljivu *GCS* i proveravali jednačinu medijana merenja prema vrednosti 9. Koristili smo dvostran test i interval poverenja 0.95. U rezultatu se pojavilo *s = 5*, što nam govori da smo imali pet posmatranja u smislu plusa. Verovatnoća odbacivanja alternativne hipoteze jeste *p-value = 0.7266*, čime moramo da prihvatimo nultu hipotezu o jednakosti medijana vrednosti 9. O tačnosti ove odluke svedoči i interval poverenja za medijanu, koji sadrži vrednost 9.

Tabela 2: Izračunavanje znakovnog testa za primer 4

```
>GCS <- c(12, 9, 6, 8, 7, 11, 9, 10, 12, 10)
> sign.test(GCS, y = NULL, md = 9, alternative = "two.sided", conf.level = 0.95)
$rval

 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 11.675556
sample estimates:
median of x
 9.5


$Confidence.Intervals
 Conf.Level L.E.pt U.E.pt
Lower Achieved CI 0.8906 8.0000 11.0000
Interpolated CI 0.9500 7.3244 11.6756
Upper Achieved CI 0.9785 7.0000 12.0000

Warning message:
In return(rval, Confidence.Intervals) :
 multi-argument returns are deprecated
```

Sada, kada znamo kako funkcioniše znakovni test, možemo ga koristiti i za potvrđivanje razlika između dve medijana pri parnom testu, u istraživanju sličnom kao i u slučaju parnog t-testa. Njegovo korišćenje ćemo pokazati na sledećem primeru (Primer 5).

Primer 5: Rezultati vrednovanja kvaliteta stručnih vodiča (guidelines) pomoću AGREE pre i posle sagledavanja sugestija

Zadavanje: Pri pripremi vodiča koristi se instrument AGREE za vrednovanje atributa, izraženog rečima kao *Stručno savetovanje je potkrepljeno sa praktičnim instrumentima.* Posle prvog vrednovanja autorski kolektiv je uradio savetovane promene i vrednovanje je bilo ponovljeno tim istim vrednovateljem. U tabeli su rezultati vrednovani pre i nakon realizacije preporuka pojedinih vrednovatelja na skali slaganja sa izjavom u rasponu 1 sve do 4. Prvi broj je izražavao potpuno neslaganje sa izjavom, a broj 4 potpuno slaganje.

	Vrednovatelj						
	1	2	3	4	5	6	7
Pre	2	3	3	3	4	3	2
Posle	3	4	2	4	3	2	4

Poručilac vodiča pita da li se postiglo poboljšanje vrednovanja posle izvršenih sugestija.

Radićemo isto kao i u prethodnom slučaju, dakle 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 formulacije vodiča.

H_0 *medijana Pred = medijana Po* H_A *medijana Po < medijana Pred*

ili je možemo formulisati i ovako:

H_0 *medijana razlike je jednaka 0* H_A *medijana razlike je veća od 0.*

Za potvrđivanje hipoteza koristićemo istu komandu, samo u ovom slučaju imaćemo dve promenljive. U pogledu jednostanog karaktera testa, koristićemo nivo poverenja 0,95, gde će biti granica prihvatanja/odbacivanja hipoteze u slučaju da je veća od 0. Rezultat potvrđivanja je u tabeli 3. Potvrđivali smo da je medijana vrednovanja sadržana u promenljivoj *posle* je veća od medijane u promenljivoj *pre*. Rezultat provere dat je u tabeli 3. Iz tog razloga smo kao prvu promenljivu komande *sign.test()* uveli promenljivu *posle,* a posle nje je sledila promenljiva *pre*. Potvrđivali smo situaciju alternativne hipoteze u smislu jednostranosti testa „veći“ pri nivou poverenja alfa = 0,05.

Tabela 5: Rezultat znakovnog testa za primer 5

```
> pred <- c(2, 3, 3, 3, 4, 3, 2)
> po <- c(3, 4, 2, 4, 3, 2, 4)
> sign.test(po, pred, alternative = "greater", conf.level = 0.05)
$rval

 Dependent-samples Sign-Test

data: po and pred
S = 4, p-value = 0.5
alternative hypothesis: true median difference is greater than 0
```

```
5 percent confidence interval:
 1.228571 Inf
sample estimates:
median of x-y
 1

$Confidence.Intervals
 Conf.Level L.E.pt U.E.pt
Lower Achieved CI 0.0078 2.0000 Inf
Interpolated CI 0.0500 1.2286 Inf
Upper Achieved CI 0.0625 1.0000 Inf

Warning message:
In return(rval, Confidence.Intervals) :
 multi-argument returns are deprecated
```

Rezultat je pokazao da postoje četiri merenja gde je razlika pozitivna. Dakle, rezultat potvrđivanja nulte hipoteze je pokazao da je moramo prihvatiti, odnosno odbaciti alternativnu hipotezu da postoji razlika između medijana.

Test medijana

Znakovni test je zasnovan na pretpostavci da su dva uzorka u uzajamnom odnosu i predstavlja neparametrijsku varijantu parnog t-testa. Često se, međutim, pojavljuju situacije kada dva uzorka nisu uzajamno povezana, a broj merenja nije uparen i može da bude različit. Kod normalno raspodeljenih populacija to nije problem i onda koristimo t-test, a varijansu populacije otkrićemo na osnovu varijanse uzorka. U slučaju da ne možemo da očekujemo normalnu raspodelu populacije iz kojih biramo uzorak, onda moramo da radimo sa medijanama. Za potvrđivanje da li postoji ili ne razlika između medijana dve populacije, koristimo test medijana, poznat i po njegovim autorima kao **Vestenberg-Mudov test medijana**. Statističko okruženje R nudi u osnovnom paketu stats komandu *mood.test(x, y, alternative = c("two.sided", "less", "greater"), ...).* 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 2x2. Pogledajte tabelu 6.

Tabela 6: Tabela 2x2 za izračunavanje testa medijane

	Promenljiva 1	Promenljiva 2	Zajedno
Broj slučajeva većih od zajedničke medijane	a	b	a+b
Broj slučajeva manjih od zajedničke medijane	c	d	c+d
Zajedno	a+c	b+d	a+b+c+d

Dalje računanje je, zapravo, jednostavan hi-kvadrat test, isti kao na početku poglavlja.

Izračunavanje za korišćenje komande *mood.test()* ilustrovaćemo na sledećem primeru 6. Korišćenje komande je veoma jednostavno, navešćemo obe promenljive i dok god želimo da testiramo obostranim testom, to ne moramo posebno ni da specifikujemo.

Primer 6: Upoređivanje medijane težine povrede hospitalizovanih na dva odeljenja

Zadavanje: Primarijus odeljenja traumatologije u većem gradu tvrdi da su povrede, koje su hospitalizovane na njegovom odeljenju, teže od onih koje 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 (tabela).

Bolnica	**ISS skor**
gradska	45 52 35 37 62 75 35 25 45 49 25 75 9
prigradska	36 42 75 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)
> mood.test(grad, selo)

 Mood two-sample test of scale

data: grad and selo
Z = -0.1285, p-value = 0.8977
alternative hypothesis: two.sided
```

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

Danas se ovaj test koristi samo retko, zbog njegove male snage. Za rešenje navedenog problema prednost dajemo Vilkoksonovom testu za dva uzorka. Komanda koja ga obavlja je napisana tako da će rešiti i test jednog uzorka, kao i parni test. Mi ćemo navesti samo prvi od njih.

Vilkoksonov test za dva uzorka

Vilkokson i Man Vitni su opisali testove ranga koji su se pokazali sličnim. Pod pojmom test ranga podrazumeva se neparametrijska procedura upoređivanja razlika, odnosno povezanost zasnovana na posmatranju poretka testirajućih podataka. Ovaj test se ponaša isto kao t-test. Komanda koja ga realizuje zove se veoma jednostavno *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, ...)*[26]. Lako je razumeti da se iza *x* i *y* zameni naziv promenljive sa uzorcima. I izbor hipoteze koju želimo da potvrđujemo isti je kao što je bio u prethodnim

[26] Obratite pažnju da korišćenje testa parnog upoređivanja potražuje zadavanje *paired = T*

slučajevima. Ako želimo da utvrđujemo 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 uvedena argumenta *conf.int = TRUE, conf.level = 0.99.* Za demonstraciju korišćenja testa iskoristićemo situaciju iz primera 6 i nazvaćemo ga primer 7.

Primer 7: Korišćenje Vilkoksonovog testa ranga za rešenje primera 6.

```
> 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 vselo
W = 62.5, p-value = 0.62
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 je analogan prethodnom primeru - nema razlike u medijanama obe populacije iz kojih potiču uzorci.

Kruskal-Volis analiza varijanse

Dešava se da ne možemo da ispunimo uslove parametarske ANOVA-e, najčešće usled malog broja podataka pojedinih uzoraka, ili u slučajevima kada ne možemo da govorimo o normalnoj raspodeli podataka, zato što su to, na primer, niže skale, ili podaci u obliku skora. U ovakvom slučaju moramo da posegnemo za neparametrijskom 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. Njen rezultat je zaključak da su najmanje dve populacije različite i onda je moguće koristiti Vilkoksonov test za dva uzorka kako bi saznali koje su to dve populacije različite. Za razliku od ANOVA-e, koja radi sa varijansama populacija, Kruskal-Vollis upoređuje medijane i saznaje da li su ili nisu jednake. Postupak je zasnovan na principu upoređivanja redosleda, isto kao Vilkoksonov test. Postupak sledi ukoliko su redosledi slični, odnosno različiti. Njegovo korišćenje pokazaćemo na primeru u kojem je student saznavao kako su pacijenti zadovoljni sa radom tri lekara, koje smo nazvali Jan, Fero i Mihal. Svakog od njih vrednovali smo njihovim pacijentima, tako da se nije desilo da pacijent jednog vrednuje drugog lekara, već se uvek izjašnjavao samo prema tome koji se lekar za njega direktno brinuo. Vrednovali su na skali od 1 do 4, gde je 1 izražavalo veliko nezadovoljstvo, a 4 veliko zadovoljstvo (Primer 8).

Primer 8: Upoređivanje tri uzorka pomoću Kruskal Volisovog neparametrijskog testa

Zadavanje: Pacijenti su vrednovali rad trojice lekara; svakog lekara, u pogledu zadovoljstva njegovim radom, vrednovalo je njegovih 10 pacijenata. Na skali od 1 do 4 izrazili su zadovoljstvo (4) ili nezadovoljstvo (1). Studenta je zanimalo postoji li razlika među ispitivanim lekarima.

Lekari		
Jan	Fero	Mihal
4	2	2
3	3	3
4	3	3
4	3	2
3	4	2
3	4	1
2	3	3
3	4	2
2	4	1
4	3	2

Postupak: Istraživač je koristio komandu kruskal.test() sledećim postupkom:

```
> jan <-c(4, 3, 4, 4, 3, 3, 2, 3, 2, 4)
> fero <-c(2, 3, 3, 3, 4, 4, 3, 4, 4, 3)
> michal <-c(2, 3, 3, 2, 2, 1, 3, 2, 1, 2) # zadao je podatke
> lekari <- list(jan, fero, michal) # komanda traži zadavanje podataka u oblike liste
> kruskal.test(lekari) # pozivanje komande

 Kruskal-Wallis rank sum test

data: lekari
Kruskal-Wallis chi-squared = 10.6238, df = 2, p-value = 0.004933
```

Rezultat je jasan, prisutne su razlike kod najmanje dva uzorka na nivou poverenja $p < 0,01$. Koji se od uzoraka međusobno razlikuju - ne znamo. Pokušajmo da nacrtamo boksplot dijagram, verovatno će nam pomoći da se orijentišemo. Nacrtaćemo ga pozivanjem komande koju smo već koristili, dakle *boxplot()* sa parametrom *lekari*, koji sadrži rezultate tri lekara (Slika 2). Iz njega jasno vidimo da je razlika između trećeg lekara, odnosno Mihala i njegove dvojice kolega Jana i Fere. Da bismo ovu činjenicu proverili i računanjem, uradićemo Vilkoksonov test za svaki par lekara, odnosno napravićemo sledeće parove: Jan-Fero, Jan-Mihal, Fero-Mihal (Tabela 7).

Slika 4: Prikazivanje podataka iz zadatka 8

```
> boxplot(lekari)
```

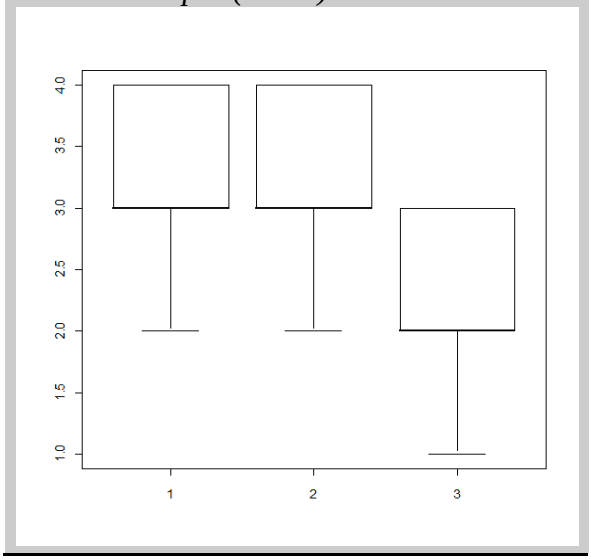


Tabela 7: Utvrđivanje razlike medijana između dvojice lekara iz primera 8

```
> wilcox.test(jan, fero)

 Wilcoxon rank sum test with continuity correction

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

Warning message:
In wilcox.test.default(jan, fero) : cannot compute exact p-value with ties
> wilcox.test(jan, michal)

 Wilcoxon rank sum test with continuity correction

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

Warning message:
In wilcox.test.default(jan, michal) :
 cannot compute exact p-value with ties
> wilcox.test(fero, michal)

 Wilcoxon rank sum test with continuity correction
```

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

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

Utvrđivanje je donelo očekivani rezultat. Između Jana i Fere nije bilo statistički značajne razlike, ali između Jana i Mihala i Fere i Mihala ova razlika je bila na nivou poverenja $p < 0,01$. Dakle potvrdili smo sumnju koju nam je donela slika 2.

Sažetak

Neparametrijki statistički testovi su u nekim oblastima neopravdano zanemareni. U slučajevima kad se radi sa skalama, sa skorovima, ili su opravdane sumnje o raspodeli populacije, da je malo merenja, naročito je dobro korišćenje ovih postupaka. Testova je više nego što smo mogli da navedemo u ovom poglavlju. Većinom su jako jednostavni i intuitivni. Odabrali smo samo one koje ilustruju korišćenje posmatrane i očekivane frekvencije, testiranje srednje vrednosti uzoraka i upoređivanje srednjih vrednosti dva ili više uzoraka, kako bismo ukazali na to kako su ovi testovi konstruisani. Ako čitalac želi da zna više, lako će ih pronaći u odgovarajućim publikacijama ili na internetu.

Vežbe

- U bolnici *Odense University Hospital* dr K.T. Drževjecki se interesovao za preživele bolesnike posle operacije melanoma. Zabeležio je podatke o 207 bolesnika. Interesovalo ga je da li postoji razlika između polova (promenljiva *sex*) u pojavljivanju melanoma u odnosu pojavljivanja čira (promenljiva *ulc*), (uzorak *melanom* iz datoteke *ISwR*).

- Studija energetskog unosa (kJ) napravljena je kod 11 žena u periodu pre (promenljiva *pre*) i posle (promenljiva *post*) menstruacije (uzorak *intake* iz datoteke *ISwR*). Da li je moguće zaključiti, da se energetski unos razlikuje u ovakvim periodima?

- U zadatku 1 iz poglavlja 7 primenite neparametrijsku metodu analize varijanse.

DODATAK

Rešenje zadataka

Sadržaj

Poglavlje 3
PREZENTACIJA I PRVOBITNA OBRADA PODATAKA

Primer 1.

```
> library(ISwR)
> data(ISwR)
> attach(cystfibr)
> summary(cystfibr)
> length(height)
> barplot(height) # stubičasti dijagram
> hist(height) # histogram
```

Primer 2.

```
> library(ISwR)
> data(ISwR)
> attach(eba1977)
> summary(pop)
> summary(cases)
> length(pop)
> length(cases)
> barplot(pop) # stubičasti dijagram
> barplot(cases)
```

Poglavlje 4
MERE CENTRALNE TENDENCIJE I DISPERZIJE

Primer 1.

```
> summary(cystfibr)
> range(age)
> range(pemax)
> IQR(age)
> IQR(pemax)
> var(age)
> var(pemax)
> sd(cystfibr)
> (sd(age)/mean(age))*100 # koeficijent varijanse
```

Primer 2.

```
> library(e1071)
> kurtosis(age)
> skewness(age)
> summary(cystfibr) # na osnovu blizine aritmetičke sredine i medijane u svim promenljivim možemo da zaključimo da su sve promenljive približno normalno raspodeljene
```

Primer 3.

```
> IQR(pop)
[1] 326.75
```

> IQR(cases)
[1] 4
> summary(pop)
Min. 1st Qu. Median Mean 3rd Qu. Max.
509.0 628.0 791.0 1100.0 954.8 3142.0
> summary(cases)
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.000 7.000 10.000 9.333 11.000 15.000
Interpretacija: relativno velika disperzija podataka, naročito u promenljivoj *pop.*

Poglavlje 5
OCENE UZORKA

Primer 1.
Se 67.48 - 162.22
Zn 14.45 - 18.25
Cu 9.64 - 13.1
Mg 0.85 – 0.93

Primer 2.
MB 128.09 - 134.71
BH 131.13 - 137.00
BL 94.44 - 101.03
NH 48.69 – 51.84

Primer 3.
fitness centar najmanje jednom u nedelji
90% CI: 0,14 0,25
95% CI: 0,13 0,26
99% CI: 0,12 0,28

fitness centar nikad
90% CI: 0.52 0.66
95% CI: 0.51 0.67
99% CI: 0,48 0,69

Poglavlje 6
POTVRĐIVANJE HIPOTEZA

Primer 1.
> t.test(expend~stature,data=energy)
t = -3.8555, df = 15.919, p-value = 0.001411
Interpretacija: Prisutna je razlika među dve aritmetičke sredine sa nivoom poverenja p<0,01

Primer 2.

> t.test(intake$pre, intake$post, paired=T)
t = 11.9414, df = 10, p-value = 3.059e-07
Interpretacija: Prisutna je razlika između dve aritmetičke sredine sa nivoom poverenja p<0,0001
> t.test(intake$pre, intake$post, alternative = c("g"), paired=T)
t = 11.9414, df = 10, p-value = 1.530e-07
Interpretacija: Prva aritmetička sredina je veća od druge, sa nivoom poverenja p<0,0001

Primer 3.

> t.test(vital.capacity~group, data=vitcap, alternative = c("l"))
t = -2.9228, df = 19.019, p-value = 0.004362
Interpretacija: Prva aritmetička sredina je manja od druge sa nivoom poverenja p<0,01

Poglavlje 7
ANALIZA VARIJANSE

Primer 1.

> data(red.cell.folate)
> attach(red.cell.folate)
[1] "red.cell.folate"
> anova(lm(folate~ventilation))
Analysis of Variance Table
Response: folate
Df Sum Sq Mean Sq F value Pr(>F)
*ventilation 2 15516 7757.9 3.7113 0.04359 **
Residuals 19 39716 2090.3
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Interpretacija: Postoje najmanje dve aritmetičke sredine koje se razlikuju na nivou poverenja p<0,05

Primer 2.

> attach(heart.rate)
> anova(lm(hr~subj+time))
Analysis of Variance Table

Response: hr
Df Sum Sq Mean Sq F value Pr(>F)
*subj 8 8966.6 1120.82 90.6391 4.863e-16 ****
*time 3 151.0 50.32 4.0696 0.01802 **
Residuals 24 296.8 12.37
---
*Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Interpretacija: Postoje najmanje dve aritmetičke sredine, koje se razlikuju na nivou poverenja p<0,0001

Zadatak 3.

```
> attach(Indometh)
> anova(lm(conc~Subject+time))
Analysis of Variance Table

Response: conc
          Df Sum Sq Mean Sq F value    Pr(>F)
Subject    5  0.729  0.1458  0.7078    0.6199
time       1 13.129 13.1294 63.7440 5.824e-11 ***
Residuals 59 12.152  0.2060
---
Signif. codes:  0 ‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1
```

Interpretacija: Postoje najmanje dve aritmetičke sredie koje se razlikuju na nivou poverenja p<0,0001

Poglavlje 8
REGRESIJA I KORELACIJA

Zadatak 1

```
> muskarci<-lm(bp~obese, subset=(sex==0))
> zene<-lm(bp~obese, subset=(sex==1))
> summary(muškarci)

Call:
lm(formula = bp ~ obese, subset = (sex == 0))

Residuals:
   Min     1Q  Median     3Q    Max
-25.645  -9.533  -1.598   7.060  59.315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   102.11      17.49   5.839 6.76e-07 ***
obese          21.65      14.50   1.493    0.143
---
Signif. codes:  0 ‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1

Residual standard error: 16.36 on 42 degrees of freedom
Multiple R-squared: 0.05038,    Adjusted R-squared: 0.02777
F-statistic: 2.228 on 1 and 42 DF,  p-value: 0.1430

> summary(zene)
```

Call:
lm(formula = bp ~ obese, subset = (sex == 1))

Residuals:
Min 1Q Median 3Q Max
-23.522 -12.348 -2.203 3.907 71.500

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 82.517 12.048 6.849 6.14e-09 ****
*obese 31.204 8.426 3.703 0.000488 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 17.56 on 56 degrees of freedom
Multiple R-squared: 0.1967, Adjusted R-squared: 0.1824
F-statistic: 13.72 on 1 and 56 DF, p-value: 0.000488

Interpretacija: U grupi muškaraca ovaj odnos nije bio linearan, u grupi žena je bio, na nivou poverenja p<0,001

Grafički prikaz linija regresije
> muskarci.koef=coef(muskarci)
> zene.koef<-coef(zene)
> plot(bp[sex==1]~obese[sex==1], type="p", data=bp.obese)
> abline(zene.koef)
> plot(bp[sex==0]~obese[sex==0], type="p")
> abline(muskarci.koef)

Primer 2.

> cor.test(bp[sex==1], obese[sex==1], use="complete.obs")

Pearson's product-moment correlation

data: bp[sex == 1] and obese[sex == 1]
t = 3.7035, df = 56, p-value = 0.000488
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2092256 0.6297033
sample estimates:
cor
0.443551

Poglavlje 9
LOGISTIČKA REGRESIJA

Primer 1

 # Nabrojmo potrebne pakete komandi
library(Epi)
library(Design)
data(bdendo)#učitajmo datoteku bdendo, koju ćemo analizirati
list(bdendo)#ispišimo sadržaj datoteke
ishod:
set d gall hyp ob est dur non duration age cest agegrp age3
1 1 1 No No Yes Yes 4 Yes 96 74 3 70-74 65-74
2 1 0 No No <NA> No 0 No 0 75 0 70-74 65-74
3 1 0 No No <NA> No 0 No 0 74 0 70-74 65-74
4 1 0 No No <NA> No 0 No 0 74 0 70-74 65-74
5 1 0 No No Yes Yes 3 Yes 48 75 1 70-74 65-74
6 2 1 No No No Yes 4 Yes 96 67 3 65-69 65-74
7 2 0 No No No Yes 1 No 5 67 3 65-69 65-74
...
 # sprovodimo analizu logističke regresije
regresia = glm (d ~ gall + age + hyp + ob + duration, family=binomial, data=bdendo)
summary (regresia)
ishod:
Call:
glm(formula = d ~ gall + age + hyp + ob + duration, family = binomial,
data = bdendo)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.6151 -0.6210 -0.5032 -0.3819 2.3037
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.786590 2.110149 -1.794 0.0727 .
*gallYes 0.948717 0.424708 2.234 0.0255 **

age 0.017384 0.029468 0.590 0.5552

hypYes 0.155425 0.351137 0.443 0.6580

obYes 0.591857 0.368249 1.607 0.1080

*duration 0.017941 0.004328 4.146 3.39e-05 ****

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 255.65 on 249 degrees of freedom

Residual deviance: 225.95 on 244 degrees of freedom

(65 observations deleted due to missingness)

AIC: 237.95

Number of Fisher Scoring iterations: 4

exp(regresia$coefficients)#Izračunajmo unakrsni odnos šansi (OR)

confint(regresia)#Izračunajmo intervale poverenja za koeficijente regresije

exp(confint(regresia))# Izračunajmo intervale poverenja za OR

Ishod:

> exp(regresia$coefficients)

(Intercept) gallYes age hypYes obYes duration

0.02267278 2.58239544 1.01753618 1.16815423 1.80734072 1.01810324

> confint(regresija)

Waiting for profiling to be done...

2.5 % 97.5 %

(Intercept) -8.021432122 0.29136951

gallYes 0.099872975 1.77538514

age -0.040501445 0.07559717

hypYes -0.540143340 0.84263645

obYes -0.111409248 1.34114768

duration 0.009530787 0.02656875

> exp(confint(regresia))

Waiting for profiling to be done...

2.5 % 97.5 %

(Intercept) 0.0003283494 1.338259

gallYes 1.1050305429 5.902554

age 0.9603077765 1.078528

hypYes 0.5826647270 2.322482

obYes 0.8945725721 3.823429

duration 1.0095763495 1.026925

#Iskoristićemo sledeće komande na stvaranje zbirne tabele sa koeficijentima regresije, unakrsnim odnosima šansi i intervalima poverenja

sum.coef<-summary(regresija)$coef

est<-exp(sum.coef[,1])

*upper.ci<-exp(sum.coef[,1]+1.96*sum.coef[,2])*

*lower.ci<-exp(sum.coef[,1]-1.96*sum.coef[,2])*

cbind(regresia$coef,est,lower.ci,upper.ci)

Ishod:

est lower.ci upper.ci

(Intercept) -3.78659011 0.02267278 0.0003625014 1.418077

gallYes 0.94871744 2.58239544 1.1233213219 5.936651

age 0.01738419 1.01753618 0.9604311989 1.078036

hypYes 0.15542492 1.16815423 0.5869576117 2.324843

obYes 0.59185655 1.80734072 0.8781733449 3.719631

duration 0.01794133 1.01810324 1.0095038341 1.026776

Interpretacija:

Prisustvo oboljenja mokraćne bešike kod ispitanika iz analizirane populacije povećava šansu nastanka raka endometriju,a za 2,58 puta. Produženje estrogenskog lečenja povećaće šansu pojave raka endometrijuma kod ispitanika iz analizirane populacije za 1,01 puta. Uticaj ostalih promenljivuh u okviru datog modela nije statistički značajan.

Primer 2

 # Na osnovu rezultata primera 1 izdvojićemo iz odela promenljive sa nesignifikantnim uticajem i uradićemo analizu logističke regresije pomoću sledeće komande:

lrm (d ~ gall + duration, data=bdendo)

Ishof:

Logistic Regression Model

lrm(formula = d ~ gall + duration, data = bdendo)

Frequencies of Responses

0 1

241 57

Frequencies of Missing Values Due to Each Variable

d gall duration

0 0 17

Obs Max Deriv Model L.R. d.f. P C Dxy Gamma Tau-a R2 Brier

298 1e-08 32.67 2 0 0.762 0.525 0.578 0.163 0.167 0.137

Coef S.E. Wald Z P

Intercept -2.17803 0.223378 -9.75 0.0000

gall=Yes 0.99616 0.399998 2.49 0.0128

duration 0.01884 0.003944 4.78 0.0000

Interpretacija:

Uticaj obe promenljive u modelu je ostao statistički signifikantan. Na osnovu vrednosti C (odnosno AUC) = 0,762 i vrednosti R2=0,167 možemo da zaključimo da model ima dobru diskriminativnu sposobnost. Niska vrednost R2 može biti prouzrokovana niskim brojem prediktora i modela.

Poglavlje 10
HI-KVADRAT I NEPARAMETRIJSKI TESTOVI

Primer 1.

```
> data(melanom)
> table(ulc, sex)
   sex
ulc  1  2
  1 47 43
  2 79 36
chisq.test(ulc, sex)

        Pearson's Chi-squared test with Yates' continuity correction

data:  ulc and sex
X-squared = 5.1099, df = 1, p-value = 0.02379
```

Interpretacija: na nivou poverenja p<0,05 možemo da zaključimo da postoji razlika između polova u pogledu pojavljivanja čira.

Primer 2.

```
> data(intake)
> attach(intake)
> wilcox.test(pre, post, paired=T)

        Wilcoxon signed rank test with continuity correction
```

data: pre and post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(pre, post, paired = T) :
cannot compute exact p-value with ties

Interpretacija: Na nivou poverenja p<0,01 možemo da zaključimo da postoji razlika između enegetskog unosa pre i posle menstruacije.

Primer 3.

> data(red.cell.folate)
> attach(red.cell.folate)
> kruskal.test(folate~ventilation)

Kruskal-Wallis rank sum test

data: folate by ventilation
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234

Interpretacija: Nisu se pronašli značajene razlike između grupa.