Univerzitet u Novom Sadu
Centar za primenjenu statistiku

Vodič kroz metodologiju nastave primenjene statistike

Odabrana poglavlja iz metodologije nastave primenjene statistike

Ova publikacija objavljena je u okviru Tempus projekta „Master program iz primenjene statistike“ 511140-Tempus-1-2010-1-RS-Tempus-JPCR

The publishing of this booklet is a part of the Tempus project “Master programme in applied statistics” MAS 511140-Tempus-1-2010-1-RS-Tempus-JPCR

Urednici
Zorana Lužanin
Andreja Tepavčević
Petar Milin
Branimir Šešelja

Tehnička obrada
Davorka Radaković

Izdavač:
Univerzitetski centar za primenjenu statistiku
Univerzitet u Novom Sadu

Tiraž: 200

„Ovaj projekat je finansiran od strane Evropske komisije. Ova publikacija odražava samo stavove autora i Komisija ne može biti odgovorna za bilo kakvu upotrebu informacija koje se u publikaciji sadrže“.

“This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein."

SADRŽAJ

UČENJE STATISTIČKIH POJMOVA UZ POMOĆ SIMULIRANIH PODATAKA

Andrej Blejec
Univerzitet u Ljubljani, Slovenija

U nastavi statistike koriste se različite vrste podataka. U kursevima primenjene statistike obično se koriste podaci iz stvarnog života, koji su u vezi sa materijom predmeta koju izučavaju studenti. Takvi podaci su zanimljivi za studente i motivišu finalnu interpretaciju statističkih rezultata. Za demonstraciju statističkih pojmova mogu biti korišćeni podaci dobijeni računarskim simulacijama, koji poseduju poznata statistička svojstva. Prednost ovakvih podataka jeste to što je rezultate statističke analize moguće uporediti sa poznatim i unapred definisanim svojstvima podataka. Primena računarskih simulacija i dinamičke grafike omogućava očiglednan prikaz brojnih važnih statističkih pojmova i postupaka. Takve simulacije ponekad mogu biti ubedljivije od dokaza i nailaze na simpatije kod studenata.

1. UVOD

Jedan od ciljeva nastave statistike jeste da pokaže studentima kako da primene statističke metode. Njihovu pažnju pokušavamo da privučemo primenom statistike, kako na realne životne probleme, tako na probleme iz domena njihovog specifičnog interesovanja. Takvi primeri, uz koje obično ide i priča koja je u pozadini problema, motivišu studente da tumače rezultate statističke analize u kontekstu samog problema (Fillebrown, 1994). Premda takav način rada omogućava studentima da sagledaju „statistički način razmišljanja” i upoznaju se s njim, ponekad je teško uočiti analitički potencijal i ograničenja primenjenog statističkog metoda.

U cilju razumevanja i tumačenja statističkih rezultata pojedinac mora da usvoji brojne statističke pojmove. Neki su jednostavni, poput mera centriranosti ili varijacije prirodnih i društvenih pojava. Neki nisu tako očigledni i često se opisuju i predstavljaju na način koji zahteva matematičko sagledavanje problema ili apstraktno razmišljanje. U takvim slučajevima, nematematički orijentisani studenti osećaju nelagodu i nisu u stanju da razumeju značenje takvih pojmova. Generacije studenata imaju problem da razumeju značajne pojmove, poput, na primer, intervala poverenja, standardne greške ili pravog značenja p-vrednosti.

Primera radi, pravilno tumačenje intervala poverenja zahteva da se razume njihovo značenje. Bez toga, izveštaj o intervalu poverenja za srednju vrednost predstavlja samo puku računsku vežbu ili još jedan klik mišem u okviru statističkog paketa, bez obzira na zanimljivost samog projekta. Ponekad su podaci iz realnog životnog problema isuviše kompleksni i teško je oceniti jesu li opisani stvarni međusobni odnosi (Mackisack, 1994). U cilju boljeg razumevanja, značenja nekih pojmova i metoda mogu biti demonstrirana i predstavljena primenom simuliranih podataka sa unapred poznatim statističkim svojstvima. U takvom slučaju, moguće je oceniti da li ispitivani metod zaista otkriva stvarno svojstvo ili međusobni odnos poadataka.

2. STVARNI, IZMIŠLJENI I SIMULIRANI PODACI

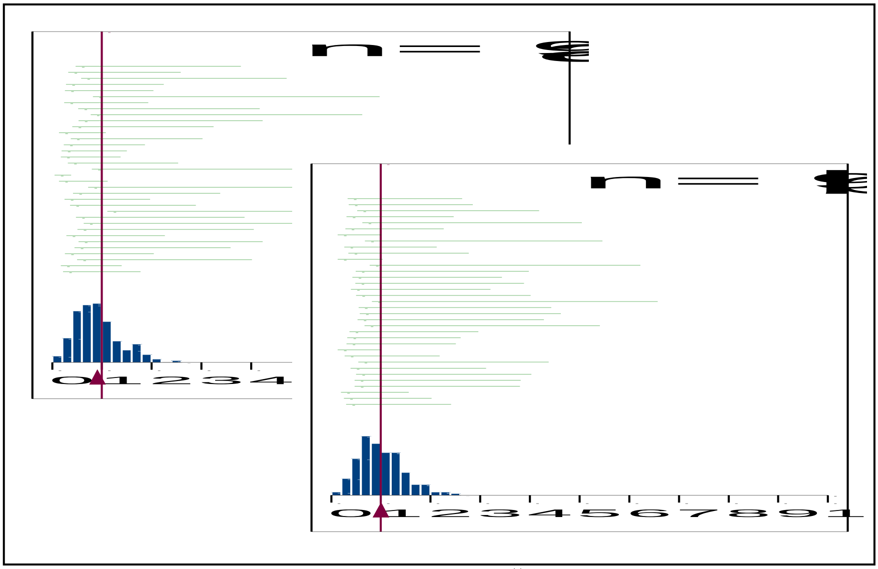
Premda je krajnji cilj statističke analize donošenje odluka u kontekstu izučavanog problema, nije neophodno da proces učenja bude u celosti sproveden isključivo na problemima koji su vezani za konkretno područje. Iako su *stvarni podaci* privlačni i motivišu studente, struktura problema je neretko suviše složena, pri čemu smisleno tumačenje ponekad zahteva dublje poznavanje problematike. Budući da struktura stvarnih podataka nije poznata, ne možemo biti sigurni da ju je primenjeni statistički metod zaista razotkrio. U cilju zaobilaženja problema kompleksnosti, problemi se, ponekad i preterano, uprošćavaju. Time, na neki način, postaju *izmišljeni podaci*, skupovi sirovih brojki koji se koriste isključivo u svrhu uvežbavanja statističkih izračunavanja. Iza njih ne stoji kontekst, a rezultati predstavljaju puke brojeve čija se ispravnost može proveriti na strani sa tačnim odgovorima u udžbeniku.

Radi demonstracije statističkih svojstava i pojmova, mogu biti korišćeni *simulirani podaci* sa poznatim statističkim svojstvima. To su uzorci iz raspodela poznatog tipa i parametara, kao što je to normalna raspodela, sa poznatom srednjom vrednosti, μ, i varijansom, σ^2. Na ovakvim podacima moguće je sagleati da li primenjeni statistički metod (npr. određivanje aritmetičke sredine uzorka) može da razotkrije traženo svojstvo (npr. pravu srednju vrednost populacije, μ). Isto tako, za demonstraciju svojstava i moći linearne regresije, možemo konstruisati promenljive sa poznatim lineranim svojstvima: uzorci promenljive sa normalnom raspodelom, $X \sim N(\mu, \sigma_X{}^2)$ i članom greške, $\varepsilon \sim N(0, \sigma^2)$, sa poznatom varijabilnosti X i ε, koji su kombinovani u promenljivu Y, *linearnom relacijom* $Y = \beta_0 + \beta_1 X + \varepsilon$, gde su β_0 i β_1 poznate modelske konstante. Ovakvi podaci su često generisani na računaru, korišćenjem generatora slučajnih brojeva kojima raspolaže svaki analitički paket ili programski jezik. Računari su od suštinske važnosti za demonstraciju statističkih pojmova, primenom ponovljenog uzorkovanja, tj. generisanjem velikog broja uzoraka koji poseduju ista, unapred definisana, statistička svojstva, gde se potom porede statistički rezultati dobijeni analizom takvih skupova podataka.

3. PONOVLJENO UZORKOVANJE

Jedan od osnovnih metoda u računarom podržanoj demonstraciji statističkih podataka jeste ponovljeno uzorkovanje (Good, 2001). Na osnovu poznatih parametara, uzorci se, jedan za drugim, uzimaju iz populacije. Razmatrani statistički postupak se primenjuje na svaki od uzoraka i posmatra se raspodela rezutata. Tipičan primer primene postupka ponovljenog uzorkovanja jeste demonstracija raspodele srednje vrednosti, standardne devijacije i intervala poverenja. Rezultati mogu biti predstavljeni uz pomoć dinamičke računarske grafike, na atraktivan i očigledan način. Pošto važi pravilo „verueš u ono što vidiš", student može da stekne osećaj za pojmove kao što su centralna granična teorema, uticaj veličine uzorka na oblik uzoračke distribucije i varijabilnost ocene. Zahvaljujući tome što je poznata stvarna srednja vrednost, studenti mogu da provere koliko intervala poverenja zaista obuhvata stvarnu srednju vrednost, te da tako steknu uvid u stvarno značenje intervala poverenja i stepena pouzdanosti.

Primenom istog postupka za ocenu varijanse, mogu biti demonstrirana svojstva pravila "podeli sa *n*-1", koje zbunjuje mnoge studente na osnovnim kursevima statistike. Dijagram približnih vrednosti pristrasne ocene (delilac *n*) i njihova uzoračka distribucija za male uzorke (Sl.1a) prikazuje nagnutu raspodelu sa očekivanom vrednosti, koja je manja od stvarne. Pristrasnost nestaje ukoliko se koristi nepristrasna ocena (delilac *n*-1) (Sl.1b). Uzoračka distribucija je takođe nagnuta ka desnoj strani, što znači da, za varijansu σ^2, ne treba konstruisati simetrične intervale poverenja (koji su zasnovani na svojstvima normalne raspodele) već treba pribeći asimetričnim intervalima poverenja, kakve ima χ^2 raspodela.



SLIKA 1. - EMPIRIJSKA UZORAČKA DISTRIBUCIJA

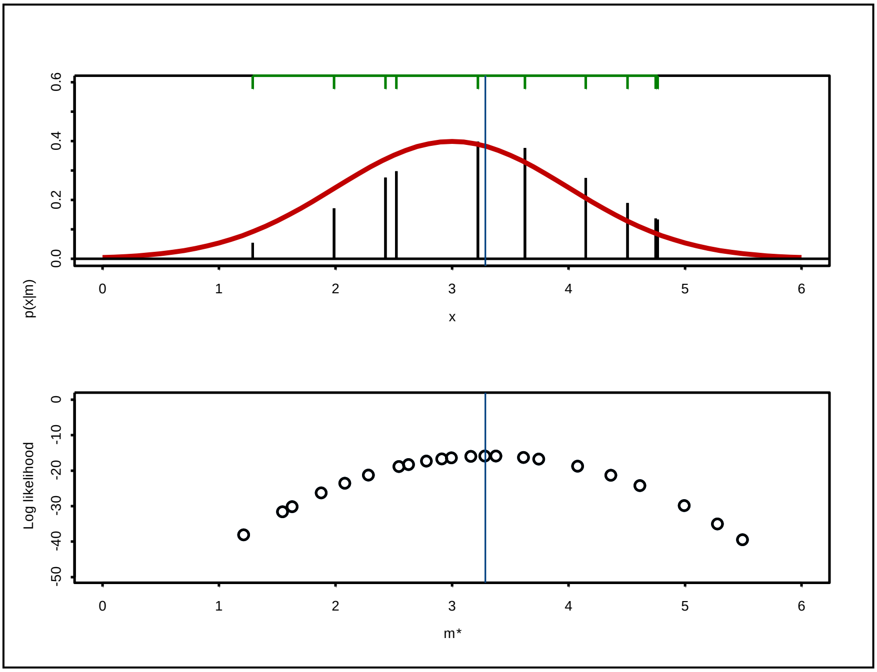
Slika 1. Empirijska uzoračka distribucija (histogrami, 200 uzoraka) za pristrasnu (a) i nepristrasnu (b) ocenu varijanse. Tačkice: ocene varijanse (veličina uzorka n=9), horizontalne linije: intervali poverenja, vertikalna linija: stvarna varijansa, σ^2, trougao: stvarna srednja vrednost uzoračke distribucije (histogram). Uočljivo je da je srednja vrednost (trougao) na slici (a) manja od stvarne vrednosti.

4. OCENA MAKSIMALNE VERODOSTOJNOSTI

Mnogi studenti imaju problem da razumeju ocenu maksimalne verodostojnosti. Primenom interaktivne, dinamičke računarske grafike, može se pokazati da se tu u stvari radi o postupku pametnog pogađanja. U tu svrhu, najpre se uzme uzorak iz poznate populacije i iscrta se dijagram uz pomoć crtica, kakav vidimo u gornjem prozoru slike 2. Studentima može biti postavljeno pitanje da pogode šta bi bila srednja vrednost. Zatim se, uz pomoć miša ili drugog pokazivačkog uređaja, odabere predloženi parametar (npr. srednja vrednost). Na osnovu predložene raspodele, pojedinačna verodostojnost podatka se nanesu na dijagram kao vertikalni segmenti (Sl. 2, gornji prozor) uz prikaz crteža predložene raspodele. U donjem prozoru slike 2, iscrta se *log verodostojnosti* za predloženu vrednost. Pošto se odabere nekoliko vrednosti i ispitaju različite situacije, oblik funkcije *log verodostojnosti* ukazuje na to da je najbolje odabrati maksimum *log verodostojnosti* funkcije. Zatim se iscrta situacija u kojoj su prikazani najbolja ocena za dati uzorak i kriva stvarne raspodele, gde se potom može izvršiti poređenje i uočiti nepodudaranje (eng. lack of fit).

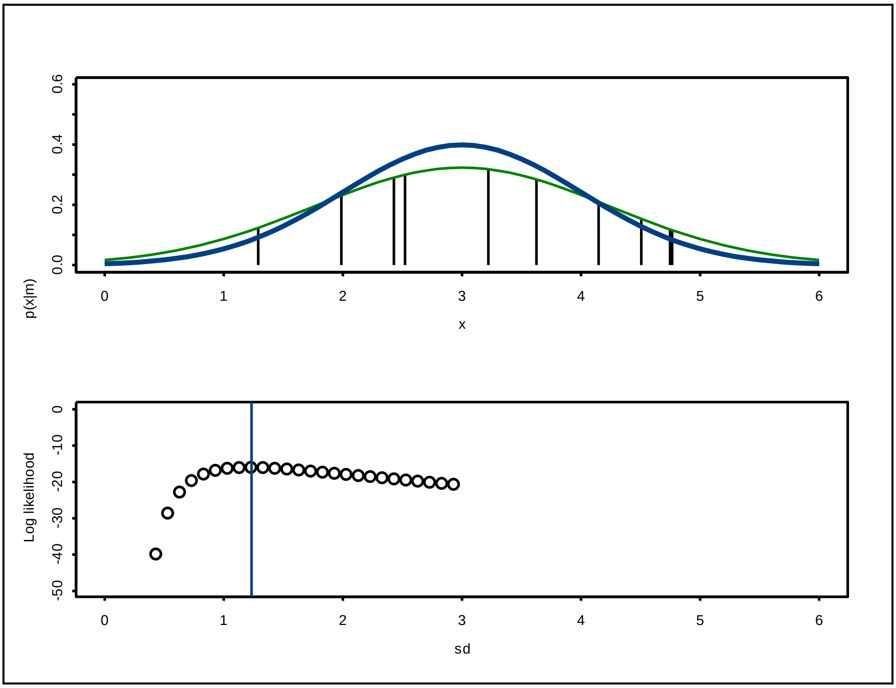
Na sličan način se može ilustrovati i ocena standardnog odstupanja (Sl. 3, podaci su isti kao na Slici 2). Ovoga puta, vrednosti predloženog parametra menjaju oblik raspodele. Sično prethodnom primeru, moguće je uočiti da neki pokušaji, u susedstvu maksimuma *log verodostojnosti* funkcije u donjem prozoru slike 3, imaju više smisla od onih koji su udaljeniji.

Ocena metodom najmanjih kvadrata može biti ilustrovana na sličan način, pojašnjavanjem pojmova odstupanja od srednje vrednosti i principa minimalne sume kvadrata odstupanja.



SLIKA 2.- OCENA MAKSIMALNE VERODOSTOJNOSTI SREDNJE VREDNOSTI

Slika 2. Ocena maksimalne verodostojnosti srednje vrednosti. *log verodostojnosti* za neke od predloženih vrednosti μ (kružići) sa najboljom ocenom za dati uzorak (vertikalna linija), iscrtane su u donjem prozoru. Pojedinačni podaci su predstavljeni crticama (gornji prozor). Vertikalni segmenti predstavljaju pojedinačne verovatnoće za najbolju ocenu. Kriva predstavlja matičnu raspodelu (μ = 3) iz koje je uzet uzorak (n=10).



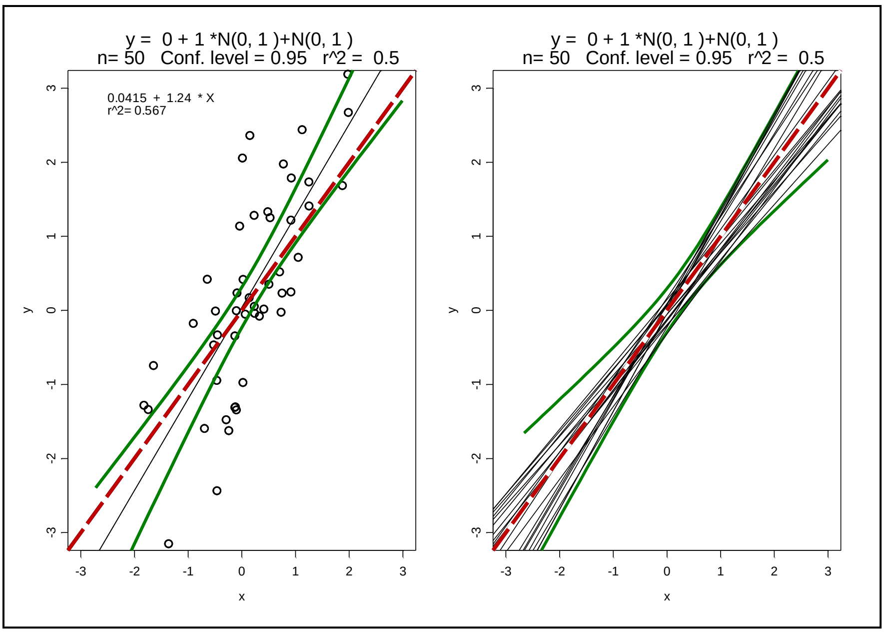
SLIKA 3. - OCENA MAKSIMALNE VERODOSTOJNOSTI STANDARDNOG ODSTUPANJA

Slika 3. Ocena maksimalne verodostojnosti standardnog odstupanja (σ=1). Gornji prozor: modelirana kriva raspodele (deblja linija), kriva raspodele dobijena ocenom (tanja linija) sa vertikalnim segmentima na mestima pojedinačnih podataka. Donji prozor: *log verodostojnosti* za neke predložene vrednosti (kružići) i najbolja ocena (vertikalna linija) za dati uzorak.

5. LINEARNA REGRESIJA

Radi demonstracije svojstava linearne regresije, simulirani su podaci na osnovu linearnog modela $Y = \beta_0 + \beta_1 X + \varepsilon$. Poznati su svi parametri linearnog modela: β_0 i β_1 su odabrane konstante modela, X i ε su normalne raspodele

$X \sim N(\mu, \sigma_X{}^2)$ i $\varepsilon \sim N(0, \sigma^2)$ sa odabranim parametrima. Uz pomoć uzorkovanja može biti pokazano da je raspodela Y normalna, pri čemu je srednja vrednost $\mu_Y = \beta_0 + \beta_1\mu$, a varijansa je $\sigma^2_Y = \beta_1{}^2\sigma_X{}^2 + \sigma^2$. Da bi se stekao osećaj za bliskost između slike koju daje uzorak i slike koju daje model, možemo generisati serije uzoraka i iscrtati podatke sa regresionim pravama. Studenti mogu da steknu utisak o uticaju koji ima varijacija člana koji predstavlja grešku, kao i na koji način je to povezano sa koeficijentom determinacije $r^2 = 1 - \sigma^2/\sigma^2_Y$. Na slici 4 su iscrtane serije regresionih pravih i prava koja opisuje model. Regresione prave se razlikuju od modelske prave za iznos varijacije člana koji predstavlja grešku. Studenti mogu lako da uoče da su regresione prave umetnute u pojas predikcije, u okolini modelske prave (Sl. 4, desni prozor), što se može transformisati u širinu intervala poverenja za određenu regresiju (Sl. 4, levi prozor). Posmatrajući rezultate regresije za veći broj simuliranih uzoraka, te poredeći ocene modelskih parametara i koeficijenta determinacije (na primeru sa Sl. 4: 0.0415, 1.2400, 0.567) sa modelskim parametrima (0, 1, 0.5), studenti mogu da saznaju u kojoj meri metod može da prikaže stvarnu strukturu podataka. Upoznajući se sa snagom metoda na primeru jednostavnih i simuliranih podataka, studenti se pripremaju za ispitivanje realnih podataka iz stvarnog života, gde će biti u stanju da tumače nepodudaranje ili da razumeju značenje širine intervala poverenja.



SLIKA 4. - ŠIRINA INTERVALA POVERENJA ZA LINEARNU REGRESIJU

Slika 4. Širina intervala poverenja za linearnu regresiju. Levi prozor: regresiona prava (tanka linija), na osnovu uzorka (kružići) veličine n=50 sa modelskom pravom (crtice) za modelske podatke (iznad) koji su sadržani u 95% intervalu poverenja (deblje krive). Desni prozor: regresione prave za 15 različitih uzoraka (tanke linije) sa 95% intervalom poverenja (deblje krive) – koji u stvari obuhvata skup regresionih pravih.

6. DISKUSIJA

Računarski simulirani podaci imaju brojne prednosti u savladavanju statističkih pojmova. Njihova statistička svojstva su poznata i moguće je sagledati povezanost između svojstava podataka i rezultata analize. Moguće je jednostavno promeniti svojstva podataka i potom posmatrati uticaj koji te promene imaju na analizu. Mnogi korisnici statistike imaju problem sa izborom odgovarajućeg statističkog metoda, jer nisu sigurni da li njihovi podaci zadovoljavaju neophodne pretpostavke za primenu nekog metoda. S tim u vezi, mogu se jednostavno simulirati podaci koji ne zadovoljavaju date pretpostavke (npr. šum čija varijansa nije konstantna) i na taj način pokazati moguću neistinitost rezultata.

Grafički podržane simulacije mogu u izvesnoj meri da nadomeste i dokaze, koji obično nisu razumljivi nematematičarima. Možda ovakve simulacije mogu da daju odgovor na Murovo pitanje: "*Ako dokaz nije ubedio publiku, čemu dokaz?"* (Moore, 1996).

Simulirani podaci moraju biti kombinovani sa realnim podacima i projektima iz stvarnog života (Mooney, 1995). Oni imaju ulogu čistih i jednostavnih podataka na kojima možemo obučiti svoju statističku percepciju i naučiti koje se šeme i svojstva mogu razotkriti primenjenim metodom. Nakon takve pripreme, studenti će biti u stanju da tumače podatke iz stvarnog života i izučavane oblasti, u svoj njihovoj kompleksnosti.

7. LITERATURA

1. Fillebraun, S. (1994). Using projects in an elementary statistics course for non-science majors, *Journal of Statistics Education*, v.2, n.2.
2. Good P. (2001). *Resampling methods*, 2nd ed Berlin: Birkhauser.
3. Mackisack M (1994). What is the use of experiments conducted by statistics students?. *Journal of Statistics Education*, v.2, n.1
4. Mooney C. (1995). Conveying truth with the artificial: using simulated data to teach statistics in the social sciences, *SocInfo Journal 1.*

5. Moore S.D. (1996). New Pedagogy and New Content: The Case of Statistics. In: Phillips B (Ed.) *Papers on Statistical Education. ICME-8*, Ed. B. Phillips, pp 1-4. Swinburne: Swinburne University of Technology.

NEKOLIKO TESTOVA ZA PROVERU ELEMENTARNIH ZNANJA IZ STATISTIKE

Vesna Jevremović
Matematički fakultet Beograd

U okviru predmeta Verovatnoća i statistika za studente III godine modula Informatika jedna predispitna obaveza bila je i sastavljanje testova za proveru znanja iz gradiva samog predmeta.

Zamisao je bila da kroz detaljnu pripremu jednog takvog testa studenti nauče odlično bar tu „svoju“ lekciju. Stoga je zadatak glasio: formulisati 20 zadataka ili pitanja u vezi odabrane lekcije i za svako ponuditi po tri odgovora, ali da netačni ne budu očigledno netačni (npr. verovatnoća koja bi bila negativna i sl.). Pri svakom startovanju programa na slučajan način se izabere 10 od raspoloživih 20 pitanja, i ponuđeni odgovori se na slučajan način permutuju. Kad korisnik odgovori na sva pitanja, dobija rezultat (na koliko je pitanja tačno odgovorio) i priliku da sazna i nauči tačan odgovor na pitanja na koja nije tačno odgovorio. Posle toga može ponovo startovati test i dobiti nekih drugih 10 pitanja. Na taj način se i nastavniku pruža mogućnost da na istom testu ispita sve studente, s obzirom da ima na raspolaganju $\binom{20}{10} = 184756$ različitih izbora pitanja, pa kad uračunamo i permutacije odgovora, imamo zaista više nego dovoljno testova za individualno testiranje!

U okviru svakog rada studenti su davali i otvoren kod programa, tako da se testovi mogu po potrebi menjati i/ili dopunjavati.

Ukupni rezultati su bili više nego zadovoljavajući – testovi su ispunili svoju funkciju, u smislu da su njihovi autori veoma dobro znali „svoje“ lekcije, što je bilo podsticajno, pa su dobro pripremili i ispit u celini. Takođe su grafička rešenja bila interesantna, studenti su koristili svoje znanje u radu sa kompjuterima i pravili elegantna i/ili duhovita grafička rešenja.

Nekoliko primera testova možete pogledati na niže navedenim adresama:

1. Rad studenta Martina Hofera se odnosi na rešavanje zadataka u vezi sa određivanjem tačkastih ocena parametara http://www.alas.matf.bg.ac.rs/~mi08005/test/
2. Rad studentkinje Bojane Kovačević se odnosi na osnovne osobine nekih poznatih raspodela – binomne, uniformne, normalne... http://alas.matf.bg.ac.rs/~mi08131/vis/
3. Rad studentkinje Katarine Radović se odnosi na osnovne pojmove – populaciju, obeležje, uzorak... http://alas.matf.bg.ac.rs/~mi08222/vis/vis.php
4. Rad studenta Vladana Stankovića se odnosi na testiranje parametarskih hipoteza kod obeležja sa normalnom raspodelom. http://vaxter.neospindle.com/vis/

POPIS ILI UZORKOVANJE?

Sanja Rapajić
Departman za matematiku i informatiku
Prirodno-matematički fakultet
Univerzitet u Novom Sadu

Populacija je celokupna kolekcija objekata u kojoj se može vršiti ispitivanje neke karakteristike tj. nekog obeležja. Ukoliko je populacija malobrojna, ona se može izučavati u celini. Međutim, ako ona sadrži veliki broj elemenata, ispitivanje cele populacije je skupo, dugotrajno, u nekim slučajevima može biti destruktivno, a ponekad je čak i principijelno nemoguće. Iz tog razloga se obično bira podskup populacije koji se naziva uzorak, na kome se vrši ispitivanje. Ideja je da pokušamo da izvedemo zaključak o celoj populaciji na osnovu analize izabranog uzorka.

Oblast statistike koja se bavi proučavanjem izbora uzorka i ocenjivanjem odgovarajućih parametara populacije naziva se teorija uzoraka. Postoje mnogobrojne tehnike i načini odabira uzorka, a samim tim i razne vrste uzoraka.

Treba napomenuti da čak i kada postoji mogućnost ispitivanja cele populacije, obično se istraživač opredeljuje za uzorak, jer je proučavanje uzorka jeftinije nego ispitivanje cele populacije, jer je kontrola tačnosti kolekcije podataka jednostavnija i lakša na uzorku nego na populaciji i jer se informacije mnogo brže dobijaju iz uzorka nego iz čitave populacije.

Da bi statističko zaključivanje o celoj populaciji dobijeno na osnovu uzorka bilo relevantno, neophodno je da uzorak bude reprezentativan. Idealno bi bilo da je uzorak procentualno umanjena tj. skalirana verzija cele populacije koja oslikava sve karakteristike date populacije. Nažalost, u praksi je gotovo nemoguće izabrati savršen uzorak kada su u pitanju složene populacije, a čak i da takav uzorak postoji, mi ne znamo da je on idealan ukoliko ne izvršimo merenje na celoj populaciji. Iz tog razloga je potrebno odabrati uzorak koji će što bolje prikazati karakteristike cele populacije, tj. koji će biti reprezentativan. Reprezentativnost podrazumeva da svaka jedinica iz uzorka predstavlja karakteristike poznatog broja jedinica populacije.

Prilikom svakog istraživanja javljaju se greške koje se dele na uzoračke i neuzoračke. Razumevanje razlika između ovih grešaka je od velike važnosti pri opredeljivanju za popis ili za uzorkovanje, kao i pri odabiru odgovarajućih procedura u slučaju da se odlučimo za uzorkovanje. Da bi rezultati istraživanja bili što pouzdaniji potrebno je minimizirati sve vrste grešaka.

Uzoračka greška je greška koja nastaje kao posledica korišćenja uzorka umesto ispitivanja cele populacije. Ona predstavlja razliku između ocene nekog obeležja dobijene na osnovu uzorka i prave vrednosti obeležja dobijene na celoj populaciji. Uzoračka greška se razlikuje od uzorka do uzorka. Ako je u pitanju verovatnosno uzorkovanje, uzoračka greška je granica greške ocene nekog obeležja dobijene na osnovu uzorka. Uzoračka greška se može smanjiti povećanjem obima uzorka i primenom stratifikovanog uzorka.

U većini istraživanja uzoračke greške su zanemarljive u odnosu na neuzoračke. Uzoračka greška odražava preciznost ocene, a neuzoračka greška validnost ocene.

Neuzoračke greške su greške koje se ne pripisuju variranju od uzorka do uzorka. To su greške koje potiču od načina uzorkovanja kojim se dobijaju ocene koje se sistematski razlikuju od pravih vrednosti obeležja populacije. Neuzoračke greške su razne vrste pristrasnosti koje nastaju u procesu istraživanja. Njihova klasifikacija je izvršena prema fazi istraživanja u kojoj se javljaju. Postoje tri osnovne kategorije ovih grešaka i to su: pristrasnost pri izboru elemenata, pristrasnost pri sakupljanju podataka i pristrasnost pri analizi prikupljenih podataka.

Pristrasnost prilikom izbora elemenata obuhvata pristrasnost specificiranja populacije, pristrasnost pokrivenosti i pristrasnost odabira.

Pristrasnost specificiranja populacije se javlja u slučaju lošeg izbora ciljne populacije ili neprecizno definisanog cilja istraživanja. Ona se može ublažiti specificiranjem ciljne populacije i pravilnim dizajniranjem upitnika koje podrazumeva razumljivost pitanja.

Najčešći tipovi pristrasnosti pokrivenosti su: nepokrivenost, prepokrivenost i višestruka pokrivenost. Nepokrivenost podrazumeva nemogućnost uključivanja cele ciljne populacije u uzorački okvir. Prepokrivenost je uključivanje u uzorački okvir onih elemenata koji ne pripadaju ciljnoj populaciji, a višestruko uključivanje elemenata populacije u uzorački okvir predstavlja višestruku pokrivenost. Ove vrste pristrasnosti se mogu minimizirati pažljivim kontrolisanjem uzoračkog okvira i ciljne populacije.

Pristrasnost prilikom odabira se javlja u slučaju kada se izabrane jedinice u uzorku sistematski razlikuju od ostalih članova populacije. Primer ove vrste pristrasnosti je namerno ili svrsishodno biranje uzorka, kao i korišćenje procedure uzorkovanja koja zavisi od nekih osobina vezanih za obeležje koje se ispituje. Kompletnim popisom eliminiše se pristrasnost prilikom odabira. Primenom uzorkovanja sa jednakim verovatnoćama, obučavanjem istraživačkog osoblja i sveobuhvatnom kontrolom kvaliteta može se smanjiti ova vrsta pristrasnosti.

Pristrasnost prilikom sakupljanja podataka se obično javlja u vidu neodziva i netačnih odgovora. Često se dešava da neki elementi populacije koji bi trebalo da čine uzorak nisu dostupni istraživaču, nisu sposobni ili ne žele da učestvuju u istraživanju. Osim toga, sakuljeni podaci dobijeni na osnovu uzorka mogu biti netačni, nepotpuni ili neodgovarajući iz mnogobrojnih razloga kao što su loše dizajniran upitnik, subjektivnost istraživača, nezainteresovanost ispitanika i dr. Dakle, ova vrsta pristrasnosti se javlja zbog neuspeha pri sakupljanju podataka ili nedovoljne količine i kvaliteta dobijenih podataka na osnovu uzorka. Greške pri merenju takođe spadaju u ovu grupu pristrasnosti. One se javljaju u slučaju kad merni instrument ima tendenciju da se razlikuje od tačne vrednosti u jednom smeru, tj. kad uvek meri manje ili više za neku konstantnu vrednost. Greške pri merenju se teško otkrivaju i mogu biti neprimetne i podmukle. Pristrasnost pri sakupljanju podataka može se minimizirati statističkim prilagođavanjima, dobro dizajniranim upitnicima, procedurama i instrumentima za prikupljanje podataka, kao i treningom istraživača.

Pristrasnost prilikom analize prikupljenih podataka predstavlja razliku između prave vrednosti populacionih parametara i ocena dobijenih na osnovu uzorka, korišćenjem određenih procedura za analizu podataka. Ona obuhvata pristrasnost nastalu usled grešaka obrade podataka i usled grešaka pri analizi podataka. Obe vrste grešaka se mogu umanjiti dobrom obukom istraživača, implementiranjem procedura za kontrolu kvaliteta i pravilnim korišćenjem statističkih alata.

Uzimajući u obzir cilj istraživanja, uzoračke i neuzoračke greške, troškove istraživanja i druge faktore, mogu se uporediti prednosti i nedostaci kompletnog popisa i uzorkovanja. Iako se na prvi pogled čini da su greške manje prilikom popisa, to ne mora biti tačno. Popis eliminiše uzoračke greške, ali neuzoračke greške mogu biti velike prilikom popisa.

Ukoliko je populacija heterogena, malobrojna, ako postoji potreba za podacima visokog kredibiliteta kao i potreba za detaljnom analizom podataka, ako je neophodno minimizirati uzoračku grešku i ako postoje određeni etički razlozi, preporučuje se ispitivanje cele populacije tj. kompletan popis. Ukoliko je pak populacija brojna i razbacana ili se lako može uništiti, ako postoji potreba za brzim donošenjem odluke kao i potreba za najsvežijim i pouzdanim informacijama, ako je neophodno minimizirati netačnost odgovora i koristiti jednostavne procedure i ukoliko su raspoloživi resursi kao što su novac, vreme i osoblje ograničeni, tada je bolje opredeliti se za uzorkovanje. Prilikom donošenja odluke o tome da li treba uzorkovati ili vršiti ispitivanje na celoj populaciji tj. raditi popis, potrebno je voditi računa o ciljevima, značaju i dizajnu istraživanja, željenom publicitetu, prirodi i osobinama populacije, raspoloživim resursima, procedurama za obradu i analizu podataka, kao i o etičkim i zakonskim zahtevima.

U današnje vreme postoji veliki broj loše odrađenih istraživanja, pa su mnogi ljudi prilično skeptični prema većini istraživanja. Neki smatraju da je uzorkovanje loše i da treba istraživanje vršiti isključivo na celoj populaciji, tj. treba raditi kompletan popis. Kao što je rečeno, za male populacije to nije problem, i kompletnim popisom se eliminiše uzoračka greška. Međutim, popis ne može isključiti neuzoračke greške. Najveći uzroci grešaka u istraživanjima su nepokrivenost, neodziv i nemarnost pri sakupljanju podataka. U opštem slučaju kompletan popis populacije iziskuje previše vremena i novca, a ponekad može biti destruktivan, a ne elimiše sve greške. Dakle, čak i kad smo u prilici da ispitujemo celu populaciju, često se odlučujemo da selektujemo samo njen manji deo (uzorak) i da na osnovu njega donosimo odluke vezane za celu populaciju. Postoje mnogobrojni razlozi za to. Uzorkovanje može obezbediti pouzdane informacije za mnogo manje novca od popisa. Podaci dobijeni na osnovu uzorka se brže prikupljaju, pa se i ocene veličina objavljuju blagovremeno. Osim toga, ocene bazirane na uzorku su često tačnije od onih dobijenih na osnovu popisa, jer se veća pažnja posvećuje kvalitetu podataka i obučavanju osoblja koje sprovodi istraživanje, što bitno smanjuje greške prilikom istraživanja.

Mnogo je bolje imati dobra merenja na reprezentativnom uzorku, nego pristrasna ili nepouzdana merenja na celoj populaciji.

METODOLOGIJA RAZVIJANJA UPITNIKA UZ POMOĆ DELFI METODA NA KURSEVIMA IZ STATISTIKE

Mirko Savić
Ekonomski fakultet u Subotici
savicmirko@ef.uns.ac.rs

Izrada upitnika i prikupljanje podataka uz pomoć upitnika je segmet koji se zanemaruje na velikom broju osnovnih kurseva iz statistike. Takvim pristupom studenti bivaju obučeni da koriste statistički instrumentarijum za matematičko-statističku analizu podataka a da pri tome ne znaju da naprave osnovni merni instrument u statistici i da prikupe podatke uz pomoć njega. Delfi metod je metod za postizanje konsenzusa unutar grupe eksperata u pogledu određenog problema i kao takav može se veoma lako i efikasno iskoristiti za definisanje i selekciju pitanja koja ulaze u upitnik. Iz tog razloga je moguće iskoristiti ovaj metod da bi se studentima demonstrirao proces izgradnje statističkog upitnika.

Ključne reči: *Delfi metod, upitnik, kurs*

1. UVOD

Kada je u pitanju nastava iz Statistike i ostalih kvantitativnih predmeta na fakultetima, gotovo po pravilu i uz svega nekoliko pozitivnih izuzetaka akcenat je na matematičko-statističkoj analizi koja se detaljno izučava tokom celog kursa. Ono što se stavlja u drugi plan a često se i u potpunosti zanemaruje jeste planiranje statističkog istraživanja i u okviru toga statističko posmatranje odnosno prikupljanje sirovih podataka. Dolazi se do apsurdne situacije u kojoj su studenti obučeni da primene statističke metode za analizu ali nemaju ni najosnovnije znanje o tome kako izvršiti prikupljanje podataka nad kojima analiza treba da se izvrši.

Posebno osetljivu tačku predstavlja izrada upitnika jer se metodologija za njihovu izradu uopšte ne predaje na većini kurseva. Ovaj problem nije prisutan samo na fakultetima u Srbiji nego i u mnogim drugim zemljama, što se lako može zaključiti ako se pogledaju udžbenici stranih autora. Takođe, ovaj problem je posebno prisutan na fakultetima iz oblasti ekonomije i srodnih disciplina.

Kao posledica dolazi do toga da u konkretnim situacijama studenti (ili diplomirani studenti) koji treba da obave istraživanje se ustručavaju da prikupe podatke odnosno ne znaju da razviju merni instrument (upitnik) za posmatranje pojave od interesa. Takođe se dešava da se napravi upitnik koji nije prošao ni najmanji test u pogledu validnosti pojedinih pitanja i upitnika kao celine nego su pitanja prosto nabacana, bez ikakve konsultacije sa stručnom literaturom, ekspertima iz date oblasti i slično. Podaci prikupljeni na taj način nikako se ne mogu smatrati relevantnim za posmatrani predmet istraživanja. Kao krajnja posledica pojavljuju se rezultati istraživanja koji daju lažnu i neobjektivnu sliku o posmatranim varijablama a na osnovu takvih rezultata izvlače se pogrešne informacije i zaključci.

U velikom broju naučnih oblasti, a pogotovo kad su u pitanju društveno-humanističke nauke, često se javlja potreba da se prikupe odgovarajući podaci o jedinicama posmatranja neposredno sa terena, odnosno od primarnih izvora podataka preko upitnika ili intervjua. Kada je u pitanju ekonomija, preduzeća imaju potrebu da istražuju tržište i da prikupljaju podatke o karakteristikama i ponašanju potencijalnih kupaca, poslovnih partnera, konkurencije, zaposlenih i slično. Takođe, javna i državna preduzeća imaju potrebu da prate veliki broj varijabli vezanih za stanovništvo, urbanizam, privredu, saobraćaj itd.

Iz navedenih razloga javila se potreba za razvijanje metodologije koja bi bila implementirana u nastavi statistike na nivou osnovnih studija i koja bi se bavila problemom razvijanja upitnika. Postoji veliki spektar različitih metoda za njihovo razvijanje a ovde je akcenat na izgradnji upitnika uz pomoć Delfi metoda.

Jasno je da metodologija za izradu upitnika podrazumeva jedan sistematičan, obiman i do detalja razrađen postupak za koji nema dovoljno vremena u toku nastave. U pitanju je čitav niz testova, panel studija i pilot istraživanja kojima se gradi i testira merni instrument. Takođe, u različitim disciplinama se preferiraju različite metodologije za razvijanje upitnika. Zbog dužine i obima materije iz te oblasti potrebno je odabrati i primeniti metodologiju za razvoj upitnika koja nije previše zahtevna u pogledu vremena a koja će dati kao rezultat dovoljno robustan upitnik sa relevantnim pitanjima kojima se zaista meri posmatrani fenomen. Naravno, studentima treba predočiti koliko je izrada upitnika ozbiljan i kompleksan postupak i da ne treba da misle da ako savladaju primenu Delfi metoda da su naučili sve što je potrebno da bi se definisao visoko kvalitetan upitnik i da je za to potrebno primeniti još testova i provera.

Dodatna korist koju će imati studenti ako nauče Delfi metod jeste u tome što ovaj metod ima veoma široku primenu i može da se koristi u bilo kojoj situaciji kada je cilj postizanje konsenzusa oko određenog pitanja.

Predmet ovog rada nije detaljna prezentacija Delfi metoda nego način na koji se Delfi metod može iskoristiti za izradu kvalitetnog upitnika odnosno kako je moguće iskoristiti ovaj veoma popularni metod za postizanje konsenzusa u proceduri za izradu upitnika. O samom Delfi metodu postoji bogata literatura koja potiče iz različitih naučnih disciplina i u kojoj se ističu sve mane i prednosti ovog metoda. O tome u će ovom radu biti reči samo u sažetom obliku i za nekoga ko je više zainteresovan za tu temu autor upućuje čitaoca na više izvora koji se nalaze u spisku referenci.

Takođe, predmet ovog rada nije prikaz kompletne metodologije za izvođenje statističkog istraživanja nego samo jednog od njegovih najosetljivijih komponenti, a to je izrada mernog instrumenta odnosno upitnika. Predavač na kursu iz Statistike može sam da odluči da li će ovu metodologiju predstaviti izdvojeno ili će je uklopiti u jednu širu prezentaciju celokupnog statističkog istraživanja sa svim fazama, počevši od planiranja istraživanja, preko prikupljanja podataka, pa do statističke analize i evaluacije istraživanja. Mnogo toga zavisi od raspoloživog vremena kao i obima ostale materije koja je predviđena da se obuhvati kursom.

2. DELFI METOD

Delfi metod je metod za postizanje konsenzusa oko određenog pitanja i kao takav može da se koristi u velikom broju konkretnih situacija. Prilikom kreiranja upitnika, odnosno izbora pitanja koja treba da sačinjavaju upitnik često se javlja dilema oko toga koja pitanja treba uvrstiti u upitnik a koja ne. Potrebno je, dakle, biti siguran da se definisanim upitnikom zaista meri pojava koja je predmet istraživanja odnosno da su odabrana pitanja relevantna. Osoba koja dizajnira upitnik, pa čak i tim istraživača, ne može biti nikada u potpunosti siguran da je izabrao prava pitanja i potrebna je određena vrsta eksterne evaluacije. Ono što bi bilo od velike pomoći dizajnerima upitnika jeste određeni nivo konsenzusa oko izbora pitanja od strane osoba koje su stručne u oblasti koja se istražuje. Delfi metod upravo nudi mogućnost da se izmeri i postigne zadovoljavajući nivo konsenzusa oko izbora pitanja. Ovde će biti reči o Delfi metodu koji koristi rangiranje. Procedura ima nekoliko koraka:

a) Izbor grupe eksperata
b) Pozivanje eksperata na saradnju i dobijanje njihovog pristanka
c) Slanje inicijalne liste pitanja ekspertima
d) Evaluacija pitanja od strane eksperata
e) Sumiranje i analiza rangiranja
f) Slanje povratne informacije ekspertima sa modifikovanom listom
g) Reevaluacija pitanja od strane eksperata
h) Sumiranje, analiza reevaluacije i izračunavanje koeficijenta konkordacije
i) Kraj iteracije

Koraci od d) do h) se ponavljaju najviše u tri iteracije a sa iteracijama može da se prestane onog trenutka kada je postignut koeficijent konkordacije veći od 0,7 odnosno tada se smatra da je postignut konsenzus među ekspertima o listi pitanja.

a) Izbor eksperata. Ova faza predstavlja najosetljiviji deo metoda i ona se sastoji u tome da se izaberu stručnjaci iz oblasti iz koje je i predmet istraživanja. Pod ekspertima se ne podrazumevaju samo naučni radnici odsnosno istraživači iz date oblasti nego to mogu da budu i drugi stejkholderi. Na primer, ako je u pitanju anketa koja je vezana za tržište rada, onda kao eksperti mogu da se pojave i zaposleni u nacionalnoj službi za zapošljavanje, predstavnici sindikata, predstavnici lokalne samouprave, predstavnici poslodavaca i slično. Ključno je da to budu osobe koje zaista poseduju visok nivo znanja o predmetu istraživanja. Pored toga, preporučuje se da kao eksperti budu uključeni stručnjaci koji poznaju posmatranu oblast iz različitih uglova.

Preporuka je da grupa eksperata bude sačinjena od 10 do 18 osoba. Za potrebe prezentacije Delfi metoda studentima nije potrebno angažovati tako veliki broj eksperata. Za demonstraciju celog postupka dovoljno je da se angažuje 3 do 5 stručnjaka da bi studenti prošli postupak evaluacije od početka do kraja.

Takođe je veoma bitno da niko od eksperata ne zna ko su ostali članovi ekspertske grupe. Jedino administrator (dizajner) celog postupka treba da zna koji su eksperti izabani u grupu. Na taj način se obezbeđuje nepristrasnost celog postupka jer eksperti ne mogu da utiču jedan na drugoga.

b) Pozivanje eksperata na saradnju i dobijanje njihovog pristanka. Kada je formirana grupa eksperata, administrator stupa u kontakt sa njima i poziva ih na saradnju odnosno učešće u procesu evaluacije pitanja. Poziv mora da sadrži kratko i razumljivo obrazloženje zašto je potrebno njihovo učešće kao i planirano opterećenje eksperata u pogledu vremena koje oni treba da odvoje za rangiranje pitanja. U nekim slučajevima se traži i pismena potrvrda o prihvatanju učešća u evaluaciji. Ono što je potvrdila praksa jeste da ukoliko se neko prihvati da bude ekspert velika je verovatnoća će to ostati do kraja postupka. Drugim rečima, stopa odustajanja eksperata usred procedure je veoma mala što doprinosi održivosti celog postupka (Okoli & Pawlowski, str. 20).

c) Slanje inicijalne liste pitanja ekspertima. Administrator šalje inicijalnu listu pitanja ekspertima sa jasnim obrazloženjem kako da rangiraju pitanja i kako da daju svoje sugestije. Ukoliko nije u pitanju prva iteracija nego neka naredna, administrator šalje listu pitanja sa naznakama i zapažanjima vezanim za prethodnu evaluaciju.

d) Evaluacija pitanja od strane eksperata. Prema uputstvu koje su dobili, eksperti rangiraju pitanja prema značaju za posmatrani problem. Rang 1 se dodeljuje najbitnijem pitanju, rang 2 sledećem po značaju i tako dalje. Moguće je dodeliti isti rang za nekoliko pitanja ako ih ekspert smatra podjednako važnim. Postoji i opširnija varijanta koja pruža mogućnost ekspertima da predlože modifikaciju ponuđenih pitanja kao i dodavanje novih pitanja u upitnik.

Ukoliko nije u pitanju prva iteracija nego neka naredna, ekspert je dobio korigovanu listu pitanja sa naznakama i zapažanjima o tome kako su drugi eksperti izvršili rangiranje, kao i eventualna nova pitanja koja su predložena. Na osnovu toga ekspert iznosi nova zapažanja i eventualno koriguje svoje stavove.

U ovoj fazi može da se pojavi problem u slučaju da neko od eksperata nije dovoljno ažuran pa kasni sa svojim delom zadatka. Administrator u tom slučaju mora da uloži dodatni trud i da podstakne datog eksperta da završi evaluaciju i pošalje svoje rezultate na vreme.

e) Sumiranje i analiza rangiranja. Ova faza je izuzetno osetljiva i mnogo toga zavisi od iskustva i sposobnosti administratora da na pravi način sagleda rezultate koje dobija od eksperata. Potrebno je dodati nova pitanja u listu, izmeniti stara (ako su to sugerisali eksperti), uporediti rangiranja eksperata i razmotriti sve njihove sugestije. Samim tim jasno je da administrator mora da bude neko ko dobro poznaje problematiku istraživanja da bi mogao da obavi taj zadatak.

f) Slanje povratne informacije ekspertima sa modifikovanom listom. Kada je sve sumirao, administrator šalje korigovanu listu pitanja ponovo ekspertima da bi oni mogli još jednom da obave rangiranje. Uz ovu korigovanu listu administrator mora da pošalje i pažljivo formulisana mišljenja drugi eksperata iz grupe sa ciljem da se približe međusobni stavovi unutar grupe.

g) Reevaluacija pitanja od strane eksperata. Na osnovu informacija koje su dobili od administratora eksperti rangiraju korigovanu listu pitanja i ponovo daju mišljenje o trenutnoj listi, takođe komentarišući stavove drugih eksperata, ako za tim imaju potrebu.

h) Sumiranje, analiza reevaluacije i izračunavanje koeficijenta konkordacije. Administrator još jednom rezimira i usklađuje stavove eksperata i izračunava koeficijent konkordacije. Postoji više načina za izračunavanje ovog koeficijenta ali preovladava mišljenje da je Kendall-ov koeficijent korelacije ranga najpodesniji. Pomenuti koeficijent se izračunava na sledeći način:

Formula ukoliko ne postoje zajednički rangovi:

$$r''_{12} = \frac{12}{m^2 \cdot n} \cdot \frac{n\sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i\right)^2}{n^3 - n}$$

gde je:

m – broj numeričkih serija

n – broj podataka u svakoj seriji

S_i, $i = 1, 2, ..., n$ – zbir rangova po redovima.

Formula ukoliko postoje zajednički rangovi:

$$r''_{12} = \frac{12}{m^2 \cdot n} \cdot \frac{n\sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i\right)^2}{\left(n^3 - n\right) - \sum_{i=1}^{k} T}$$

Korektivni faktor se izračunava na osnovu sledeće formule:

$$\sum_{i=1}^{k} T = \frac{1}{12}\sum_{i=1}^{k}\left(k^3 - k\right)$$

gde je:

k – broj obeležja koja imaju zajedničke rangove.

Ukoliko je koeficijent konkordacije veći od 0,7 smatra se da je postignut konsenzus između eksperata u pogledu liste pitanja. U suprotnom konsenzus ne postoji i potrebno je nastaviti postupak evaluacije.

i) Kraj iteracije. Delfi metod zapravo može da ima tri ishoda:

- Konsenzus je postignut odnosno koficijent konkordacije je veći od 0,7. To znači da je lista pitanja koja treba da uđu u upitnik definisana i time se Delfi metod završava, a u upitnik ulazi unapred određeni broj pitanja počevši od onoga koji ima najveći prosečni rang.
- Konsenzus nije postignut i administrator nastavlja sa iteracijama dok se ne postigne željeni nivo konsenzusa. Postupak se nakon treće iteracije može nastaviti jedino ako su se eksperti složili da nastave sa evaluacijom pitanja.

- Konsenzus nije postignut ni posle treće iteracije i administrator zaključuje da nema svrhe nastaviti sa iteracijama. Definitivna lista pitanja za upitnik se ne može izraditi.

3. IMPLEMENTACIJA METODOLOGIJE NA KURSU IZ STATISTIKE

Polazi se od pretpostavke da su studenti upoznati sa pojmom statističkog istraživanja i sa njegovim fazama. Pored toga, studenti treba da poseduju osnovna znanja o načinima prikupljanja podataka. Tek nakon što su studenti stekli pomenuto predznanje, moguće je primeniti ovu metodologiju.

Metodologija za implementaciju razvijanja upitnika preko Delfi metoda u nastavi se sastoji iz sledećih koraka:

1. Prezentacija Delfi metoda
2. Definisanje predmeta i problema istraživanja
3. Formiranje tima za istraživanje
4. Izrada inicijalne liste pitanja
5. Implementacija Delfi metoda
6. Evaluacija postupka razvijanja upitnika

Ceo postupak zahteva izvesno vreme i nije moguće završiti implementaciju na samo jednom predavanju. Pravi pristup bi bio da se u nekoliko uzastopnih predavanja odvoji po jedan deo časa za rad na kreiranju upitnika.

3.1. Prezentacija Delfi metoda

Prvi korak u implementaciji metodologije jeste da se Delfi metod predstavi na jednostavan i razumljiv način studentima da bi oni mogli da shvate njegov značaj u celoj proceduri.

3.2. Definisanje predmeta i problema istraživanja

Da bi se studenti na adekvatan način upoznali sa metodologijom za razvoj upitnika najbolji način je da se osmisli konkretno istraživanje u kojem bi bio upotrebljen upitnik za prikupljanje podataka. Potrebno je da predavač odredi predmet istraživanja koji je vezan za neku aktuelnu problematiku u okviru naučnog polja u kojem se predaje statistika. Na primer, ako je to ekonomija, onda predmet istraživanja može da bude vezan za tržište, cene, potrošnju, robu i slične kategorije. Bitno je da tema bude aktuelna, zanimljiva i vezana za buduću struku studenata.

U ovoj fazi predavač može da donese odluku da li da se napravi dizajn kompletnog statističkog istraživanja ili da se fokusira samo na postupak izrade upitnika. U ovom radu je prezentovana druga opcija.

3.3. Formiranje tima za istraživanje

Od prisutnih studenata formira se tim za istraživanje koji će usko sarađivati sa predavačem. Tim ne treba da broji više od tri do pet studenata. Bilo bi korisno da sve što članovi tima budu radili da se radi pred ostalim studentima, na samom času, da bi svi studenti mogli da prate postupak. Druga varijanta koja bi bila manje zahtevna po pitanju vremena jeste da studentski tim nakon svake faze izrade upitnika referiše na času dokle se stiglo sa izradom upitnika i kako je protekla poslednja urađena faza.

U okviru tima je potrebno odabrati jednog studenta koji bi bio administrator u okviru Delfi metoda dok bi ostali studenti bili zaduženi za direktnu komunikaciju sa ekspertima.

3.4. Izrada inicijalne liste pitanja

Organizuje se mali brainstorming gde studenti u komunikaciji sa predavačem predlažu pitanja koja bi bila pogodna za upitnik. Takođe se raspravlja o formulaciji pitanja, na koji način respondenti treba da odgovore na njih (vrsta obeležja) i koji će biti njihov redosled. U ovoj fazi može se dati vremena studentima da konsultuju odgovarajuću literaturu, pronađu ranija istraživanja na istu temu i da na naredni čas donesu svoje predloge.

Kao rezultat ove faze dobija se inicijalna lista pitanja koja predstavlja input za Delfi metod. To znači da će ova lista pitanja biti obrađena kroz Delfi metod i da evaluacija eksperata kreće sa njom. Broj pitanja u inicijalnoj listi nije ograničen, ali za lakše savladavanje materije od strane studenata nije potrebno da bude mnogo pitanja. Dovoljno je da lista sadrži 10 do 15 pitanja od kojih će u finalnoj formi upitnika ostati 7 do 8 pitanja. Naravno, studentima treba naglasiti da je ovo jedan školski primer i da u praksi broj pitanja u inicijalnoj listi može da bude i po nekoliko puta veći.

Lista pitanja je smeštena u tabelu koja ima sledeći izgled:

Pitanje	**Rang pitanja**	**Napomena**
1. Pitanje		
2. Pitanje		
3. Pitanje		
...		
Predlog novih pitanja:		

U koloni sa napomenom ekspert može da iznese svoje mišljenje o tome kako je moguće preformulisati dato pitanje kao i dodatna zapažanja u vezi datog pitanja. U poslednjem redu ekspert može da predloži dodatna pitanja koja mogu da budu uvrštena u upitnik.

3.5. Implementacija Delfi metoda

Implementacija Delfi metoda treba da ide po istoj proceduri koja je prezentovana studentima na početku predavanja.

Izbor grupe eksperata. Studenti biraju članove grupe eksperata, a najbolje i najjednostavnije je da to bude iz redova profesora i asistenata na svom fakultetu. U toj grupi sasvim je dovoljno da se nađe oko 3 člana. Na taj način ceo postupak evaluacije će biti brže obavljen. Određuje se student (ili studenti) koji će biti zadužen za direktnu komunikaciju sa ekspertima. Prikupljaju se njihovi kontakt telefoni i mail-ovi.

Pozivanje eksperata na saradnju i dobijanje njihovog pristanka. Stupa se u kontakt sa ekspertima i zamoljavaju se za saradnju. Iskustvo je pokazalo da je većina nastavnika i saradnika veoma voljna za učešće u procesu evaluacije. Ono što je ovde ključno jeste da se utvrdi da li će svaki ekspert biti dostupan u predviđenom periodu od desetak dana da bi se postupak izveo do kraja. Takođe je bitno objasniti svakom ocenjivaču koliko vremena približno treba da izdvoji da bi obavio evaluaciju.

Slanje inicijalne liste pitanja ekspertima. Inicijalna lista se dostavlja ocenjivačima u odgovarajućoj formi (na papiru ili mail-om). Student koji je zadužen za kontakt daje uputstva ekspertu kako da izvrši evaluaciju. Preporučuje se da se lista pitanja dostavi lično i da se ekspert zamoli da evaluaciju obavi odmah jer će se tako ceo proces dodatno ubrzati.

Evaluacija pitanja od strane eksperata. Ukoliko ekspert dobije listu pitanja lično postoji šansa da će on odmah izvršiti evaluaciju a to znaći da će administrator istoga dana imati rangove i sugestije eksperata na raspolaganju.

Sumiranje i analiza rangiranja. Najosetljiviji deo procedure za studente kod kojeg je neophodna značajna pomoć predavača. Administrator može već ovde da pokuša da izračuna koeficijent konkordacije i da izvede zaključak koliko je blizu željenoj vrednosti od 0,7. Potrebno je preko deskriptivne analize za svako pitanje otkriti gde se javljaju najveća odstupanja i ekstremne vrednosti rangova. Tamo gde se otkriju ekstremne vrednosti sastavlja se odgovarajući komentar koji će biti poslat onom ekspertu koji je naveo tu vrednost ranga da mu se ukaže da njegovo mišljenje značajno odstupa od ostalih.

U ovoj fazi se vrši modifikovanje postojećih pitanja na osnovu sugestija eksperata kao i dodavanje novih pitanja koje su oni predložili. Ovaj deo sa novim pitanjima bi trebalo da se javlja samo u prvoj i evenutalno u drugoj iteraciji Delfi metoda.

Ova faza može da bude podeljena na više manjih operacija gde bi svaku operaciju obavljao po jedan student. Na taj način bi više njih bilo uključeno pa bi jedan student izračunavao koeficijent konkordacije, drugi bi radio deskriptivnu analizu a treći bi davao odgovarajuće komentare.

Slanje povratne informacije ekspertima sa modifikovanom listom. Studenti zaduženi za kontakt sa ekspertima šalju im korigovanu listu sa komentarima. Pomenuti studenti moraju da budu upoznati sa sadržajem korigovane liste da bi mogli da daju dodatne informacije ekspertima ako je to potrebno.

Reevaluacija pitanja od strane eksperata. Eksperti na osnovu dobijenih komentara ponavljaju postupak evaluacije i eventualno daju dodatna objašnjenja.

Sumiranje, analiza reevaluacije i izračunavanje koeficijenta konkordacije. Nakon reevaluacije administrator izračunava koeficijent konkordacije i radi deskriptivnu statističku analizu. Ukoliko je koeficijent iznad 0,7 postupak je završen a ukoliko nije započinje se nova iteracija.

Kraj iteracije. U idealnoj situaciji slaganje između eksperata se postiže sa najviše tri iteracije i dobija se spisak relevantnih pitanja za upitnik. U suprotnom treba napraviti još nekoliko iteracija ili zavšiti Delfi metod uz konstataciju da konsenzus nije postignut. Čak i ako se to dogodi studenti će upoznati metodologiju i shvatiti koliko definisanje upitnika može da bude ozbiljan problem u praksi.

3.6. Evaluacija

Nakon što je Delfi metod završen i bez obzira na to da li je upitnik definisan ili nije, vrši se rekapitulacija implementirane metodologije. Svaka faza od početka do kraja treba da se objasni. Prilikom ove prezentacije potrebno je da studenti vide kako je izgledao spisak pitanja nakon svake iteracije a takođe i kako su izgledali komentari i sugestije kako eksperata tako i administratora.

Najbolja varijanta u ovoj fazi bi bila kada bi sami studenti koji su radili na kreiranju upitnika objašnjavali svojim kolegama šta je urađeno i na kakve su konkretne probleme nailazili.

Izuzetno je važno skrenuti pažnju da je pored Delfi metoda potrebno izvršiti testiranje upitnika sa aspekta kako interne tako i eksterne validnosti. U kranjoj liniji, studenti moraju znati da ako su dobili pitanja koja su relevantna sa aspekta eksperata, da to još uvek ne znači da pitanja čine homogenu i funkcionalnu celinu u potpunosti. Ukoliko postoji prostor na kursu, ovaj deo razvijanja upitnika bi takođe trebalo obraditi odnosno demonstrirati kako se izvodi testiranje upitnika, kako se sprovode pilot istraživanja i drugi metodi.

4. ZAKLJUČNA RAZMATRANJA

Veoma je bitno da studenti na kursu iz statistike nauče kako se u praksi prikupljaju podaci na različite načine. Zbog velikog obima materije koja treba da se savlada iz celog predmeta nije moguće posvetiti dovoljno pažnje ovom segmentu statističkog istraživanja. Ipak, da bi studenti znali da prikupe kvalitetne podatke moraju da bar u osnovnim crtama dobiju informacije o tome kako se definiše upitnik kao osnovni merni instrument u statistici.

U ovom radu je predstavljena metodologija za definisanje upitnika uz pomoć Delfi metoda. Pored svih svojih nedostataka ovaj metod može da posluži tome da studenti nauče metodologiju za izbor relevantnih pitanja koja treba da se uvrste u upitnik.

5. LITERATURA

1. Gordon, J. (2009). *The Delphi Method*. The Millennium Project.
2. Holey, E., Feeley, J., Dixon, J., & Whittaker, V. (2007). *An exploration of the use of simple statistics to measure consensus and stability in Delphi studies*. BMC Medical Research Methodology , 1-10.
3. Horvat, N., & Kos, M. (2010, May 10). *Development and Initial Validation of a Patient Satisfaction With Pharmacy Performance Questionnaire (PSPP-Q)*. Retrieved September 19, 2011, from Evaluation & the Health Professions: http://ehp.sagepub.com/content/33/2/197
4. Legendre, P. (2004). Species Associations: *The Kendall Coefficient of Concordance* Revisited. Retrieved January 9, 2012, from http://www.bio.umontreal.ca/legendre/reprints/Kendall_W_paper.pdf
5. Okoli, C., & Pawlowski, S. (2009). *The Delphi method as a research tool: an example, design considerations and applications*. Information & Management (42), 15-29.
6. Rayens, M., & Hahn, E. (2000). *Building Consensus Using the Policy Delphi Method*. Policy, Politics, & Nursing Practice , 1 (4), 308-315.
7. Skulmoski, G., Hartman, F., & Krahn, J. (2007). *The Delphi Method for Graduate Research*. Journal of Information Technology Education , 6, 1-21.

KAKO PREDAVATI MATEMATIKU STUDENTIMA PRIMENJENE STATISTIKE

Branimir Šešelja
Departman za matematiku i informatiku
Prirodno-matematički fakultet,
Univerzitet u Novom Sadu, Srbija

U ovom tekstu analiziraju se neki aspekti univerzitetske nastave osnova matematike za ne-matematičare. Posebno se bavimo programom master studija iz primenjene statistike, preciznije predmetom linearna algebra i kalkulus. Analiziramo matematičke pojmove potrebne za ovaj program i raspravljamo o didaktičkim aspektima njihove prezentacije. Nakon pregleda nastavnih jedinica ovog programa i njihove povezanosti sa drugim poljima relevantnim za ovu naučnu oblast, analiziramo didaktičke aspekte. Ti aspekti povezani su sklonostima i usmerenjima studenata, njihovom motivacijom da razumeju i primenjuju matematiku, kao i sa nekim drugim uslovima. U ovoj analizi dajemo i neke opšte zaključke o ulozi matematike u studijama primenjene statistike.

1. UVOD

Nakon Bolonjskih promena u programima visokog obrazovanja, sve se češće događa da bečelor sa jednog studijskog programa upiše master studije iz neke druge oblasti. Razlozi su različiti, ponekad povezani sa stanjem na tržištu rada, ili sa uverenjem da je u novom području lakše steći kvalifikaciju, ili najzad zato što je student shvatio da prethodne studije nisu bile njegov optimalni izbor. Osim ovih razloga, na nivou master studija postoje i interdisciplinarni studijski programi, dakle oni koji povezuju ili se odnose na nekoliko područja. Upis na takve master studije je logičan izbor za bečelore u svim oblastima koje su na neki način vezane uz temu ovih master studija.

Master studijski program iz primenjene statistike upućen je bečelorima iz društvenih nauka, ekonomije, tehnologije, medicine itd. Jasno je da je ovim studentima potrebno veliko znanje matematike. Budući da upisuju program nakon raznih osnovnih (bečelor) studija, njihova predznanja iz matematike u opštem slučaju veoma se razlikuju.

To otvara problem ne samo u izboru tema koje će na kursu biti izlagane, već i u načinu na koji ove teme treba razraditi I prezentovati. Osim toga, primeri i zadaci za vežbu trebalo bi da budu neposredno povezani sa primenama statistike i mogućnostima njene upotrebe.

Navedeni aspekti nastave matematike za (primenjene) statističare obrađeni su u ovom tekstu.

2. NASTAVNE JEDINICE I KAKO IH PREZENTOVATI

Uvidom u planove i programe za primenjenu statistiku na svim univerzitetima u Evropi koji imaju takve ili slične studijske programe, lako se uočava da se teme na osnovnom kursu iz matematike više ili manje poklapaju:

- neke uvodne teme iz logike, skupova, relacija i funkcija;
- osnovi kombinatorike ;
- osnove linearne algebre (sistemi linearnih jednačina, determinante, matrice);
- kalkulus (struktura realnih brojeva, realne funkcije jedne promenljive, diferencijalni i integralni račun), funkcije s više promenljivih, parcijalni izvodi, problem optimizacije.

Logički pristup treba da obnovi neka opšte pojmove u vezi sa izvođenjem matematičkih tvrdnji i nekoliko osnovnih pravila, kao što su modus ponens, zakoni kontrapozicije i svođenja na protivrečnost i još neka, ali samo na neformalan način i kroz konkretna izvođenja. Prilikom podsećanja na (Bulovu) algebru skupova i na svojstva skupovnih operacija, treba imati na umu algebru događaja u teoriji verovatnoće. To je isto tako dobro mesto da se objasne osnovni pojmovi kombinatorike. Binarne relacije bi treble biti povezane sa njihovim karakterističnim funkcijama, i zato prikazivane tablicama; odgovarajuće primene su baze podataka. Osim toga, tablični prikaz omogućuje vizualnu analizu relacija (refleksivnost, simetričnost, anti-simetričnost, a takođe relacije ekvivalencije i njihovi blokovi mogu biti identifikovani). Među posebnim binarnim relacijama, ekvivalencija i poredak treba da budu posebno naglašeni. Dovoljno je da se poredak poveže sa brojevima i skupovima. Relacije ekvivalencije moraju biti povezane sa količničkim skupovima, odnosno particijama. Pogodni su primeri npr. razlomaka kao klasa uređenih parova, ili vektora u ravni; u oba slučaja jednakost i operacije mogu se objasniti u preko predstavnika klasa.

Korespondencija treba da se uvede analogno binarnim relacijama, uz pomoć reprezentacije putem dijagrama i tablica (karakteristična funkcija). Jasno je da funkciju treba predstaviti kao posebnu korespondenciju, ali je objasniti ravnopravno i kao "proceduru" odnosno "pravilo". Posebno treba istaći operacije kao specijalne funkcije i ukazati na razliku između njih i relacija. Pogodni su i poznati primeri binarnih operacija s brojevima (sabiranje, množenje).

Osnovna srednjoškolska znanja o brojevima treba obnoviti i sistematizovati. Nije potrebno eksplicitno uvoditi algebarske strukture kao što su grupoid, polugrupa, grupa, prsten, integralni domen ili polje; ali je svakako potrebno da se jasno istaknu i obrade njihova svojstva i to kroz izlaganje poznatih svojstava operacija sa brojevima. Metodološki, postoje dve glavne opcije za uvođenje i razvoj pojma broja: Ili početi s prirodnim brojevima a zatim analizirati i opisati strukturu celih brojeva i postepeno doći do realnih i kompleksnih brojeva; ili analizirati strukturu (polje) realnih (kompleksnih) brojeva i identifikovati druge skupove brojeva kao odgovarajuće podstrukture. Za ovaj kurs predlažemo prvu opciju. Postoji nekoliko razloga. Prvo, ove teme nisu nove, ali ih treba obnoviti i zatim sistematizovati imajući u vidu njihova svojstva i upotrebu. U tu svrhu, npr., motivisati uvođenje brojeva polazeći od potrebe dobijanja rešenja jednačine $a+x = b$ za sve prirodne brojeve a,b bolje je nego identifikovanje strukture celih brojeva u polju realnih, na sličan način i u drugim slučajevima. Eksplicitnu skupovnu izgradnju brojeva (npr. celi brojevi kao klase ekvivalencije uređenih parova prirodnih brojeva u odnosu na jednakost zbira odgovarajućih komponenata) treba izbjegavati, ili je svesti na opis i komentare u vezi sa razlomcima, jer je ta konstrukcija dobro poznata. Drugi razlog za uvođenje brojeva preko skupovne inkluzije odnosi se na algebarska svojstva navedenih struktura (grupe, prsteni, itd.). Polazeći od prirodnih bojeva (kao strukture polu-prstena), lako je uvesti svojstva prstena bez izričitog definisanja same strukture, slično je i sa (poljima), racionalnih i realnih brojeva. Konačno, uzimajući u obzir cilj i svrhu ovog studijskog programa, neophodno je uvesti i uvežbavati aritmetičke tehnike kroz konkretne probleme i primere.

Osnove linearne algebre ovde su predstavljene sistemima linearnih jednačina, determinantama i matricama. Vektorski prostor R^n nad R nije neophodno eksplicitno definisati, iako se vektori, skalari i odgovarajuće operacije pojavljuju u tom kontekstu i njihova svojstva treba objasniti. Kao i obično, determinante se uvode zajedno sa sistemom od dve i tri linearne jednačine. U tom kontekstu pogodno je i predstaviti njihova svojstva. Definicija determinante reda n može biti uvedena pomoću rekurzivne formule, a pravilna definicija u vidu odgovarajućeg zbira proizvoda (korišćenjem permutacija i broja inverzija) može se dodatno dati kao informacija. U vezi sa matričnim računom, važno je naglasiti svojstava zbira i proizvoda, a time i (implicitno) predstaviti osobine prstena kvadratnih matrica: nekomutativnost množenja, postojanje jedinične matrice i delitelja nule. U tom kontekstu, inverznu matricu treba uvesti kao inverzni elemenat same strukture, a samu konstrukciju opisati kroz primere.

Kao i obično, nizovi brojeva uvode se pomoću poznatih primera, a njihove granične vrednosti detaljnim geometrijskim opisom i objašnjenjem na realnoj pravi; zatim dolazi pravilno definisanje tog pojma i opis tehnika za određivanje i analizu graničnih vrednosti nizova. Opšti pristup realnim funkcija je standard na svim kursevima matematike za ne-matematičare. Ono što je specifično u ovom kursu je da svojstva treba motivisati i predstavljati kroz konkretne primere koji proizlaze iz rada sa raznim bazama podataka, eksperimenata, događaja. Kao uvod u grafike funkcija, pogodno je korištenje poligona, histograma frekvencija, raznih tablica iz prakse. Nakon standardne prezentacije elementarnih funkcija, diferencijalni račun treba izložiti i putem odgovarajuće definicije ali i predstavljanjem izvoda kao operatora. Ovaj drugi pristup je veoma važan, budući da se koristi kao alat za sve primene. U sveukupnom predstavljanju realnih funkcija koristeći diferencijalni račun, potrebno je analizirati one koje su povezane s raspodelom verovatnoće. Isto važi i za integralni račun i parcijalne izvode. Konkretno, koristiti geometrijsku interpretaciju pri objašnjavanju osnovnih svojstava integrala uvek imajući na umu svojstva raspodela. Za parcijalne izvode važi isto što je već spomenuto za diferencijalni račun: naglasiti opis putem operatora.

3. KAKO PREDAVATI

Kao što je već navedeno, studenti koji upisuju ove master studije imaju različita znanja iz matematike. Neki od njih nikada nisu ni imali matematiku na nivou osnovnih studija, a kod onih koji jesu, taj kurs imao je različit obim i težinu. Stoga nastavniku nije lako održavati odgovarajući tempo na predavanjima, niti je jednostavno odrediti koliko informacija i detalja odgovarajućeg gradiva prezentovati. Jedan način da se prevaziđe ovaj problem je da se svaka tema uvede paralelno sa problemom do koga se dolazi u njenim primenama i da se oblast, pojam i sl., analizira pre navođenja stroge definicije, odnosno kada se uvodi nova matematička tema, oblast, pojam i slično:

- Krenuti sa primerom u kome se novi pojam (implicitno) javlja.
- Zatim objasniti neformalno, opisno novi pojam, bez (previše) formula.
- Ponovo prezentovati primer.
- Najzad dati preciznu matematičku formulaciju (definiciju, tvrđenje,...).
- Ako treba dokazati tvrđenje, prvo ilustrovati dokaz na primeru.
- Zatim navesti korektan dokaz.
- Ponovo se vratiti polaznom primeru, sada u novom kontekstu.

Za vežbe, treba pronaći probleme u kojima bi studenti morali otkriti nužnost korišćenja nekog matematičkog postupka. Predstaviti probleme s nekoliko mogućih pristupa i motivisati studente da istražuju i pronalaze pogodne algoritme. Ohrabriti studente na razumevanje tema, formulacija i teorema.

Treba održavati stalan kontakt sa kolegama koji predaju stručne predmete (vezane za statistiku i njene primene).

Isto tako, važno je korišćenje nekog opšteg matematičkog softvera (to može biti npr. Mathematica), naročito za rešavanje praktičnih problema i uopšte na vežbama. Tako se studenti uvežbavaju u radu sa odgovarajućim programima, kako bi kasnije lakše prihvatili i koristili neophodne statističke pakete.

4. LITERATURA

1. V. Jevremović, *Verovatnoća I statistika*, Matematički fakultet, Beograd, 2009.
2. J. Rosenblatt, *Basic Statistical Methods and Models for the Sciences*, Chapman&Hall/CRC, 2002.
3. Đ. Takači, Ar. Takači, Al. TakačI, *Elementi više matematike*, Symbol, Novi Sad, 2008.
4. A. Tepavčević, Z. Lužanin, *Matematičke metode u taksonomiji*, Prirodno-matematički fakultet Novi Sad, 2006.

STATISTIKA NA OSNOVNIM AKADEMSKIM STUDIJAMA SOCIOLOGIJE

Valentina Sokolovska
Univerzitet u Novom Sadu
Fillozofski fakultet

Nakon ukazivanja na značaj statističkog znanja za razvoj sociologije kao nauke, rad se bavi prikazom odabranih, najznačajnijih problema koji se javljaju u nastavi statistike na fakultetima društvenih nauka. Takođe se nakon izdvajanja problema nude i moguća rešenja do kojih se došlo u dosadašnjim istraživanjima ove problematike. Autor izdvaja dve velike grupe poteškoća kada su u pitanju statistički kursevi na studijama sociologije: prvi su metodičke i sadržinske prirode, a drugi organizacione.

Ključne reči: *sociologija, statistika, statistički kursevi.*

1. UVOD

Sociologija i statistika su oduvek bile tesno povezane. Međutim, njihova saradnja je tokom istorije često bila uslovljena kako načinom prikupljanja podataka u sociologiji, tako i razvojem statističkih tehnika koje su mogle da zadovolje složenost istraživanja društvenih pojava. Tako Clogg (1992) smatra da je razvoj sociološke metodologije i kvantitativne sociologije oduvek bio tesno povezan sa razvojem statističke teorije i metodologije. On primenu kvantitativnih metoda u sociologiji deli na dva istorijska perioda, pre i posle Drugog svetskog rata. Zbog fragmentarnog prikupljanja podataka, prvi period odlikuje primena deskriptivne statistike i jednostavnijih metoda, dok su se u drugom periodu, sa povećanjem obima podataka, počele primenjivati i složenije statističke tehnike.

U skladu sa ovakvom podelom, Raftery (2001) izdvaja tri posleratna perioda u primeni statističkih metoda u sociologiji i naziva ih: unakrsno tabeliranje, istraživanje podataka na nivou jedinica i novije forme podataka. Prvi period nastaje odmah nakon Drugog svetskog rata, drugi počinje da se razvija u ranim 60-im godinama prošlog veka, dok treći kreće od kasnih 80-ih. Ono što je, po ovom autoru, karakteristično za ovako izdvojene periode primene statistike u sociologije je to što su svi oni i danas aktuelni u sociološkim istraživanjima.

Očito je da sociologija koristi statističke metode u proveravanju svojih teorija, ali je problem koji se uočava koliko često i ispravno to čini. Stoga se i mnoge kritike upućene na račun sociologije od strane samih sociologa odnose upravo i na primenu statističkih metoda u sociologiji. Jedna od ovih kritika pojavila se devedesetih godina XX veka među američkim sociolozima. Jednu stranu ovog problema objašnjava Cole (1994). On smatra da je sociologija toga vremena 1) patila od nedostatka teorije koja se može operacionalizovati u istraživanjima, 2) da se njen teorijski razvoj nije mogao meriti sa razvojem teorija u prirodnim naukama, 3) da ne postoji kognitivni konsenzus, 4) a i da je metodologija u velikom neskladu sa istraživačkim nalazima. Na sličan način reaguje i Collins (1994). Po njegovom mišljenju, napredak statističke metodologije nije doveo do visokog stepena uniformnosti u način na koji sociolozi sprovode svoja istraživanja i nije proizveo konsenzus ili pak brzo otkrivanje suštinskih pitanja.

Pored proveravanja svojih teorija, razvijanja metodologije, preciznosti obrade i analize podataka, sociologija prepoznaje se još jednu korist koju može imati od statistike. Ona se ogleda u sve prisutnijoj saradnji sa drugim naučnicima prilikom objavljivanja naučnih radova. Baveći se analizom koautorskih mreža u sociologiji, Moody (2004) zaključuje da je koautorstvo sve češće u sociologiji, ali da nije ravnomerno distribuirano među sociološkim disciplinama. Ono je češće prisutno u oblastima u kojima postoji kvantitativna analiza. Ovaj nalaz upućuje na to da teorijski naklonjeni sociolozi sve češće prepoznaju značaj proveravanja svojih teorijskih stanovišta statističkim metodama, ali i na to da im je u tu svrhu potrebna pomoć drugih. Iako se na nastavnike statistike, pa sam tim i na statistiku, gleda kao na "servisere" koji opslužuju druge studijske programe (Snelgar and Maquire, 2010), Ray (1974) veruje da statističar u kombinaciji sa nestatističarem nikada neće biti toliko efikasan kao jedan naučnik koji kombinuje obe veštine. Samo dobro poznavanje socioloških problema i statističkih tehnika može da da optimalnu kombinaciju karakteristike podataka i analitičkih metoda.

2. MOGUĆI PROBLEMI U ORGANIZOVANJU NASTAVE STATISTIKE

Na osnovu nalaza gore iznesenih radova, neminovno se nameće zaključak da je za savremeni poziv sociologa, ali i budućnosti sociologije kao nauke, neophodno adekvatno obrazovanje obogaćeno kursevima iz statistike. Ali, kakav sadržaj ponuditi na ovim kursevima i kako približiti statistički način razmišljanja i statističke metode studentima društvenih nauka, pitanja su kojima se odavno poklanja velika pažnja.

Iako studenti sociologije imaju predstavu o tome šta su istraživanja masovnih pojava, oni imaju vrlo malo ili čak nimalo statističkog predznanja kada upisuju fakultete. Stoga Boynton (2004) upozorava da je nastava iz statistike često neefikasna jer je previše napredna za mnoge studente. Pogrešno planiranje statističkih kurseva od strane nastavnika dešava se jer se znanje studenata ne procenjuje pre početka kursa. Posledica neprilagođenog nivoa znanja koje nudimo studentima je da oni ne mogu da koriste primere koje su radili na času, nego se fokusiraju na memorisanje testova a da pri tome nemaju razvijenu svest o tome pod kojim okolnostima treba da ih koriste.

Druga česta greška u nastavi statistike je što mnogi nastavnici kombinuju klasična predavanja sa softverima za statističku analizu. Iako je korišćenje ovih programa od vitalnog značaja, Boynton (2004) smatra da njihovo prerano korišćenje može dovesti do toga da studenti na kraju kursa odlično vladaju programom za statističku analizu, a da ne znaju ništa o testu koji su koristili.

Kada se statistika prikazuje kao grupa testova, a ne povezuje se sa sociološkim problemima, studenti mogu da se upoznaju sa formulama i podacima, ali za njih ostaje nejasno kako da tumače i razumeju rezultate. Sasvim drugačiji efekat se dobija kada se podaci povezuju sa stvarnim primerima i studenti mnogo lakše razumeju šta se dešava.

Boynton (2004) predlaže da se predavanja iz metodologije istraživanja spoje sa statističkim kursevima. Na taj način se studenti podstiču da povezuju statistiku sa istraživačkim metodama i da razumeju kako njihov izbor metoda i istraživački dizajn može uticati na kasniju analizu podataka.

U odnosu na ove probleme sadržajne i metodičke prirode, u obrazovanju sociologa pojavljuju se i problemi organizacione prirode.

Oni se ogledaju u nejednakom tretmanu koji se pridaje mestu statistike u studijskim programima sociologije. Naša ranija istraživanja studijskih programa osnovnih akademskih studija sociologije u Srbiji (Sokolovska, 2011) pokazuju veliku neujednačenost u pridavanju pažnje statističkim kursevima. Tako postoje programi u kojima ne postoji nijedan kurs posvećen statističkom obrazovanju, kao i oni u kojima statistički kursevi zauzimaju dva obavezna i izborne kurseve. Slična situacija se ogleda i u zemljama u okruženju. Ovakva neusklađenost odraz je još uvek prisutnog mišljenja kreatora socioloških programa da je sociologija isključivo teorijska nauka kojoj statističko obrazovanje nije potrebno. Rezultat ovakvog pristupa je mnoštvo jednostavnih, deskriptivnih pokazatelja u naučnim radovima i veoma malo rezultata ozbiljno planiranih, sprovedenih i analiziranih socioloških istraživanja. Nažalost, ovakvo stanje u sociologiji će potrajati sve dotle dok se ne promeni mišljenje o ulozi statističkog znanja kao “servisa” koji može da se unajmi kada vam je potrebno, a potom se sa olakšanjem zaboravi. Jedan od izlaza iz krize u koju se sociologija kao nauka našla je i prihvatanje statistike kao nezaobilazne u obrazovanju sociologa.

3. ZAKLJUČAK

I pored dugotrajne saradnje između sociologije i statistike, kao i prepoznavanja koristi koje sociologija kao nauka može imati od statističkog znanja, ono još uvek nije sastavni i opšteprihvaćeni deo u visokoškolskom obrazovanju sociologa.

Problemi koji se pri tom javljaju mogu se podeliti u dve grupe. Prva se odnosi na sadržinsku i metodičku stranu ponuđenih statističkih kurseva u smislu da su pogrešno isplanirani i neprilagođeni studentima, da se često prave greške u načinu prezentovanja statističkih metoda stavljujući prerano akcenat na upotrebu statističkih softvera što za posledicu ima nerazumevanje uloge metoda kao i njihove povezanosti sa metodologijom istraživanja, ali i mnogi drugi koji u ovom radu nisu navedeni.

Druga grupa problema se može odrediti kao organizaciona. Naši prethodni radovi pokazuju da su statistički kursevi nejednako zastupljeni u studijskim programima sociologije. Ova neujednačenost je uočena kako na univerzitetima u Srbiji, tako i u pojedinim zemljama koje je okružuju.

Ono što je i pored svih poteškoća koje su uočene značajno jeste to što je prepoznata višestruka korist od primene statistike u sociologiji. Savremena sociologija se ne može razvijati bez provere njenih teorijskih stanovišta, razvijene metodologije istraživanja i primene adekvatnih statističkih metoda. Skladna kombinacija teorijskog, metodološkog i statističkog znanja takođe je dovela do intenziviranja saradnje između sociologa i naučnika iz drugih oblasti.

4. LITERATURA

1. Boynton, Petra (2004). Teaching statistics – the missing ingredients. *Radical Statistics*, 87: 19-30.
2. Clogg, C. Clifford (1992). The Impact of Sociological Methodology on Statistical Methodology. *Statistical Science*, 7(2): 183-207.
3. Cole, Stephen (1994). Introduction: What`s Wrong with Sociology? *Sociological Forum* 9(2): 129-131.
4. Collins, Rendall (1994). Why the Social Sciences Won`t Become High-Consensus. *Sociological Forum* 9(2): 157-177.
5. Moody, James (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohezion from 1963 to 1999. *American Sociological Review* 69(2): 213: 238.
6. Raftery, E. Adrian (2001). Statistics in Sociology, 1950-2000: A Selective Review. *Sociological Methodology*, 31, 1-45.
7. Ray, John (1974). Should Sociology Require Statistics? *The Pacific Sociological Review*, 17(3), 370-376.
8. Snelgar, Rosemary and Moira Maquire (2010). Assessing for success: An evidence-based approach that promotes learning in diverse, non-specialist student groups. In: P. Bidgoog at all (ed), *Assessment Methods in Statistical Education. An international perspective*. John Wiley & Sons Ltd.
9. Sokolovska, Valentina (2011). Position and Perspective of Statistics in Sociology. *Chinese Business Review* 10(10): 924-929.

ODABIR PROMENLJIVIH U LOGISTIČKOJ REGRESIJI

Antonio Lucadamo[1], Biagio Simonetti
Department of Economical, Juridical and Social Studies
University of Sannio
Via delle Puglie, 82, 82100, Benevento (Italy)
[1] alucadam@unisannio.it

U mnogim oblastima, kao što su transportni sistemi, medicina, hemiometrija ili procena zadovoljstva klijenata, srećemo se sa zavisnim, dihotomnim karakterom, uz mnoštvo opisnih promenljivih. U takvim situacijama nije moguće koristiti model linearne regresije, već jedno od mogućih rešenja predstavlja model logističke regresije. U svakom slučaju, postoje brojne situacije u kojima se problem koji prati primenu ovog modela ogleda u činjenici da se raspolaže malim brojem opservacija uz prisustvo brojnih opisnih promenljivih. U tom slučaju je ocena parametara vrlo skupa, a postupak često daje pogrešne rezultate. U cilju prevazilaženja ovih problema, poslednjih godina je predloženo nekoliko mogućih rešenja; u ovom radu je predložen DISCO (eng. discrimination coefficient) koeficijent koji izdvaja one promenljive koje mogu biti od značaja za primenu logističke regresije.

Ključne reči: *logistička regresija, DISCO koeficijent*

1. UVOD

U brojnim aplikacijama vezanim za medicinske, društvene ili inženjerske nauke, od velikog je značaja izučavanje odnosa između zavisne promenljive i nekoliko nezavisnih promenljivih. U tom kontekstu, višestruka regresija spada u najčešće korišćene statističke alate. Međutim, kada je zavisna promenljiva dihotomnog tipa, model logističke regresije omogućava proučavanje zavisnosti. U ranoj fazi istraživanja, broj nezavisnih promenljivih koje utiču na zavisnu promenljivu može biti vrlo velik, posebno u kontekstu broja raspoloživih opservacija. Da bi se prevazišao ovaj problem, u modelu se može koristiti neproizvoljan izbor broja nezavisnih promenljivih. Cilj ovog rada jeste primena kriterijuma za izbor broja nezavisnih promenljivih koje će biti korišćene u logističkoj regresiji. Rad je organizovan na sledeći način:

tačka 2 se odnosi teoriju *binarnog logit modela* (*Binarni Logit Model)*; u tački 3 je predstavljen indeks koji je predložen u literaturi za discriminantnu analizu, *DISCO coefficient*; u tački 4 je prikazana logistička regresija, izvedena na pogodnom podskupu promenljivih koje su izabrane na osnovu *DISCO coefficient*; konačno, tačka 5 prikazuje studiju simulacije, u kojoj je pokazano da predloženi postupak omogućava prednost u smislu ocene i troškova izračunavanja.

2. BINARNI LOGIT MODEL

Za potrebe analize dihotomnih ishoda predložene su brojne funkcije raspodele. Logistički model je jedan od najviše korišćenih, a svoju popularnost duguje činjenici da je formula za verovatnoću logit izbora jednostavna za tumačenje, posebno u poređenju sa drugim modelima za kvalitativni izbor (Train, 2003; Ben-Akiva & Lerman, 1985).

Da bismo razumeli kako funkcioniše ovaj model, moramo najpre uočiti da je odziv binarni, uz pretpostavku samo dve vrednosti. Pri tom su, radi jednostavnosti, te vrednosti kodirane kao 1 i 0. Tako, na primer, imamo:

$$y_i = \begin{cases} 1 \\ 0 \end{cases}$$

Ovde je y_i realizacija slučajne promenljive Y_i, koja može uzimati vrednosti 1 i 0, sa verovatnoćama π_i i 1- π_i , respektivno. Slučajna promenljiva Y_i ima Bernulijevu raspodelu sa parametrom π_i. Očekivana vrednost i varijansa Y_i su tada:

$$E(Y_i) = \mu_i = \pi_i$$

$$\text{var}(Y_i) = \sigma_i^2 = \pi_i(1-\pi_i)$$

Lako je uočiti da srednja vrednost i varijansa zavise od verovatnoće π_i.

Dalje, želeli bismo da verovatnoće π_i zavise od vektora realizovanih kovarijansi, x_i. Najjednostavnija ideja jeste da π_i bude linearna funkcija kovarijansi:

$$\pi_i = x_i^{'}\beta$$

gde je β vektor sa koeficijentima regresije. Verovatnoća na levoj strani mora biti između 0 i 1, ali linearni prediktor na desnoj strani može uzimati bilo koju realnu vrednost, te ako ocenjujemo model primenom standardnog metoda najmanjih kvadrata, ne postoji garancija da će predviđene vrednosti biti u očekivanom opsegu. Jednostavno rešenje bi bilo da se transformiše verovatnoća, eliminišu ograničenja opsega i da se transformacija modelira kao linearna funkcija kovarijansi.

Najpre je bitno preći sa verovatnoće na odnos šansi (eng. the odds), na sledeći način:

$$odds_i = \frac{\pi_i}{1-\pi_i}$$

Prethodna vrednost se definiše kao količnik između verovatnoće i njenog komplementa. U drugom koraku, logaritmovanjem se dolazi do *logit* vrednosti:

$$\eta_i = \log it(\pi_i) = \log\frac{\pi_i}{1-\pi_i}$$

Rešavanjem po π_i, dobijamo:

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Na ovaj način smo sigurni da će verovatnoća uzimati vrednosti između 0 i 1. Nadalje, možemo pretpostaviti da je logit linearna funkcija prediktora:

$$\eta_i = \log it(\pi_i) = x_i^{'}\beta$$

te tako:

$$\pi_i = \frac{e^{x_i^{'}\beta}}{1+e^{x_i^{'}\beta}}$$

gde je x_i vektor kovarijansi, a β je vektor koeficijenata regresije.

Parametri logit modela mogu biti ocenjeni primenom ocene maksimalne verodostojnosti. Razmatramo funkciju log-verodostojnosti:

$$\log L(\beta) = \sum_i \left[y_i \log(\pi_i) + (\eta_i - y_i)\log(1-\pi_i) \right]$$

izračunavajući prvi i drugi izvod i razvijajući Fišerovu bodovnu (scoring) moguće je dobiti ocene parametara.

Ocena parametara modela može se realizovati primenom različitih statističkih softverskih paketa, ali, pod određenim okolnostima, mogu se pojaviti problemi. U stvari, u nekim oblastima, kao npr. u transportnim sistemima ili hemiometriji, broj nezavisnih promenljivih je zatvoren ili je veći od broja opservacija. U tom slučaju, algoritmi za ocenu parametara rade jako sporo i, u mnogim situacijama, može postojati jaka korelacija između prediktora (multikolinearnost), čime ocena modelskih parametara postaje netačna usled potrebe da se invertuju blizu singularne i loše uslovljene informaciona matrice.

Da bi se omogućila tačna aproksimacija modelskih parametara, Aguilera et al. (2006) su predložili *Principal Component Logistic Regression*, a Camminatiello & Lucadamo (2010), *Principal Component Multinomial Regression*. Ove dve tehnike omogućavaju prevazilaženje uobičajenih zamerki koje se stavljaju pred *Principal Component Regression*. U stvari, one iz analize izostavljaju komponente koje bolje objašnjavaju varijabilnost podataka, već samo one komponente koje su značajno asocirane sa zavisnom promenljivom (u opštem slučaju, koristi se postupna regresija (eng. *stepwise regression*). U svakom slučaju, broj komponenti je vrlo visok i, takođe u ovom slučaju, postupna regresija može biti skupa za izračunavanje. Uvođenje DISCO indeksa, u analizu diskriminante, može poslužiti kao dobar alat za otkrivanje promenljivih sa najvećom moći diskriminacije, a ove se, u drugom koraku analize, mogu iskoristiti u klasičnom logističkom modelu kao regresori.

3. DISCO KOEFICIJENT

Analiza linearne diskriminante (LDA - Fisher, 1936) je jedna od najčešćih metoda za klasifikaciju. Ona registruje odnose između kategoričkih zavisnih promenljivih i većeg broja nezavisnih promenljivih. U klasičnom problemu dikriminante postoje dve klase i dobija se jedna funkcija diskriminante. Fišerov prilaz je zasnovan na izboru linearnih kombinacija promenljivih koje daju najveći količnik između međugrupnih i unutargrupnih varijansi. Linearne kombinacije originalnih opservacija su date kao:

$$Z_i^{(1)} = \sum_{l=1}^{p} W_l X_{il}^{(1)}, \quad i = 1, \ldots, n_1$$

$$Z_j^{(2)} = \sum_{l=1}^{p} W_l X_{jl}^{(2)}, \quad j = 1, \ldots, n_2$$

Gde je $X_{il}^{(1)}$ l-ta promenljiva i-te opservacije iz Grupe 1, $X_{jl}^{(2)}$ je l-ta promenljiva j-te opservacije iz Grupe 2, n_1 i n_2 su veličine ove dve grupe, p je broj nezavisnih opisnih promenljivih i W_l su težinski koeficijenti koji definišu funkciju diskriminante. Fišerob postupak omogućava maksimiziranje $\frac{S_B^2}{S_w^2}$, gde je S_B^2 varijansa linearnih kombinacija između grupa, dok je S_w^2 objedinjena varijansa unutar grupe. Ovaj prilaz je isto što i maksimiziranje Pirsonovog faktora korelacije, $\eta^2 = \frac{S_B^2}{S_B^2 + S_w^2}$. Ova veličina se takođe može napisati kao:

$$\eta^2 = \frac{N\left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)^2}{N\left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)^2 + S^2}$$

Gde su $\bar{Z}^{(1)}$ i $\bar{Z}^{(2)}$ dve aritmetičke sredine suma $Z_i^{(1)}$ i $Z_j^{(2)}$, $N = n_1 \cdot n_2 /(n_1 + n_2)$, $S^2 = \left[(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2\right]/(n_1 + n_2 - 2)$, a S_1^2 i S_2^2 su varijanse raspodela ovih dveju suma.

Godine 1983, Raveh je predložio nemetričku analizu diskriminante (NDA), koja je zasnovana na pravilu razdvajanja koje se razlikuje od onog koje se koristi u linearnoj analizi diskriminante.

U NDA, mera koju treba maksimizovati zasnovana je na skupu sledećih nejednakosti:

$$Z_i^{(1)} \geq Z_j^{(2)}, \quad i = 1, \ldots, n_1, \quad j = 1, \ldots, n_2$$

Ove nejednakosti su ekvivalentne sledećem:

$$\left(Z_i^{(1)} - Z_j^{(2)}\right) = \left|Z_i^{(1)} - Z_j^{(2)}\right| \quad za \quad svako \quad i \quad \text{i} \quad j$$

Indeks razdvajanja grupa je dat kao:

$$IS = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(Z_i^{(1)} - Z_j^{(2)}\right)}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left|Z_i^{(1)} - Z_j^{(2)}\right|} = \frac{n_1 n_2 \left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left|Z_i^{(1)} - Z_j^{(2)}\right|}$$

Raveh je takođe tvrdio da će dati i uopšteni koeficijent diskriminacije za više grupa, ali ovo je eksplicitno učinio Gutman (Guttman) (1988), a potom i Raveh (1989). Gutman ga je nazvao *DISCO* indeks:

$$disco = \frac{\sum_{g=1}^{G} \sum_{h=1}^{G} n_g n_h \left(\bar{Z}^{(g)} - \bar{Z}^{(h)}\right)}{\sum_{g=1}^{G} \sum_{h=1}^{G} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \left|Z_i^{(g)} - Z_j^{(h)}\right|}$$

Lako se pokazuje da:

$$0 \leq disco \leq 1$$

$disco = 0$ ako i samo ako svi uzorci imaju istu srednju vrednost; $disco = 1$ kada nema preklapanja između sume bilo koje dve grupa, osim, možda, u jednoj tački.

DISCO koeficijent je robusniji na autlajere od Fišerovog indeksa (L2 norma), zato što se zasniva na apsolutnoj vrednosti (L1 norma). Štaviše DISCO koeficijent je više informativan i ne zahteva pretpostavke o distribuciji za primenu.

Zbog ovih razloga, velik broj primena DISCO koeficijenta u diskriminacionoj analizi je razvijen poslednjih godina (Allard J. et al., 2000; Choulakian & Almhana, 2001; Simonetti & Choulakian, 2003)

4. LOGISTIČKA REGRESIJA PRIMENOM DISCO INDEKSA

Imajući u vidu značajne karakteristike DISCO indeksa, naša ideja je da ga iskoristimo kao alat u modelu logističke regresije.

Kao što smo rekli u prethodnom delu rada, kada postoji mnogo nezavisnih promenljivih, postupna logistička regresija zahteva mnogo vremena za ocenu parametara.

Mi predlažemo sledeću proceduru:

- Korak 1: izračunavanje DISCO indeksa za sve nezavisne promenljive;
- Korak 2: izbor samo onih promenljivih kod kojih je DISCO koeficijent veći od definisanih vrednosti;
- Korak 3: korišćenje ovih promenljivih u klasičnoj logističkoj regresiji i ocena parametara;
- Korak 4: provera dobijenih rezultata.

Očigledno je da je jedan od važnih koraka definisanje vrednosti praga DISCO koeficijenta, kako bi promenljive bile zadržane u analizi.

Radi definisanja ispravne vrednosti, kao i za proveru valjanost rezultata dobijenih na ovaj način, razmotrićemo postupak za simulaciju koji je opisan u narednoj tački rada.

5. ANALIZA I REZULTATI

Prvi korak našeg postupka, implementiran u R-softveru (R Development Core Team, 2010), sastoji se u dobijanju skupa p nezavisnih promenljivih koje smo generisali iz normalne raspodele. U drugom koraku se fiksira vektor realnih parametara , β , i izračunavaju se realne verovatnoće. Naposletku se svaka vrednost koja ulazi u sastav vektora odziva, *y*, simulira iz binomne raspodele. Na ovaj način imamo vektor odziva i matricu nezavisnih promenljivih, pa možemo da fitujemo binarni logit model.

Kako je već rečeno, cilj ovog rada jeste da se razmotri korišćenje DISCO indeksa koji može da skrati vreme potrebno za izračunavanje i, ukoliko je to moguće, da definiše vrednost praga za izbor promenljivih koje će biti zadržane u regresionom modelu. U cilju postizanja navedenog, ponovili smo simulaciju uzimajući u obzir isti broj opservacija, ali menjajući broj nezavisnih promenljivih. Odlučili smo da fiksiramo broj opservacija na 1000, dok je broj promenljivih varirao između 5 i 115, sa korakom 5.

U prvoj fazi analize, za svaku od simulacija razmatrali smo sledeće tačke:

- Ocenu parametara primenom binarne logit postupne regresije;
- Vreme koje je neophodno za ocenu prethodnog modela;
- DISCO vrednosti za sve promenljive iz analize.

Ovde se problem odnosio na definisanje vrednosti DISCO indeksa koji će obezbediti da u analizi budu sačuvane sve značajne promenljive.

U cilju odlučivanja o tome koja bi vrednost mogla biti smatrana ispravnom, uporedili smo rezultate dobijene na osnovu svih skupova podataka i uvideli smo da je prava vrednost 0.1. Ova vrednost je očigledno empirijska, imajući u vidu da je izvršeno samo nekoliko simulacija. Neophodno je izvršiti još studija u cilju otkrivanja vrednosti koja može biti korišćena kao prag za DISCO indeks, ali u našem slučaju je i prethodni postupak je poslužio svrsi.

U stvari, u skoro svim skupovima podataka, promenljive sa vrednošću iznad 0.1 su ujedno signifikantne u postupnoj regresiji. Samo u nekim situacijama, nekoliko signifikantnih promenljivih ima vrednost manju od 0.1. Važno je, međutim, istaći da u svim simulacijama nesignifikantne promenljive uvek imaju vrednost koja je bliska nuli.

Na ovaj način možemo realizovati binarni logit model (bez primene opisanog postupka), uzimajući u obzir samo one nezavisne promenljive koje su selektovane na osnovu DISCO indeksa, te tako možemo izračunati vreme koje je potrebno za ocenu ovog modela.

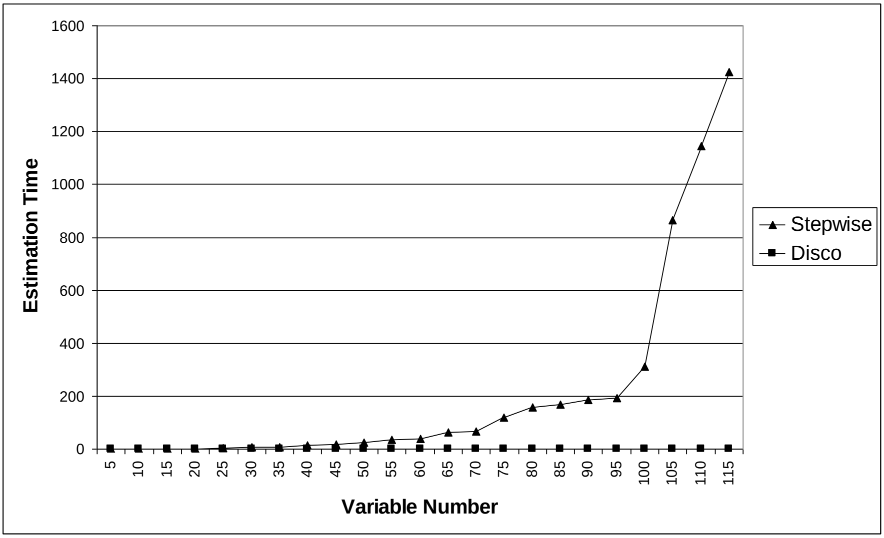
U poslednjem koraku analize, neophodno je izvršiti poređenje vremena koje je neophodno za ocenu postupnog modela i vremena koje je potrebno uz primenu drugog kriterijuma. Na taj način se mogu pokazati prednosti korišćenja DISCO indeksa.

Lako je primetiti da je kod DISCO postupka ocenjivanja (Tabela 1.) vreme uvek ispod sekunda, dok pri porastu broja promenljivih, postupna regresija beleži eksponencijalni porast potrebnog vremena. Situacija postaje jasnije kada razmotrimo Sliku 1.

Rezultati u vezi sa vremenima su pokazani u sledećoj tabeli:

TABELA 1 – VREMENA OCENJIVANJA ZA DVA RAZMATRANA POSTUPKA

Broj promenljivih	***Vreme potrebno za postupni postupak***	***Vreme potrebno za DISCO postupak***
5	0.16	0.02
10	0.17	0.03
15	1.14	0.04
20	1.35	0.04
25	3.77	0.05
30	5.56	0.05
35	7.00	0.06
40	13.76	0.06
45	18.18	0.08
50	23.33	0.09
55	36.69	0.09
60	38.33	0.12
65	63.25	0.12
70	67.36	0.14
75	120.49	0.18
80	156.78	0.18
85	168.45	0.42
90	186.23	0.43
95	191.95	0.50
100	310.64	0.51
105	866.44	0.56
110	1144.28	0.58
115	1424.42	0.62



SLIKA 1 – POREĐENJE VREMENA OCENJIVANJA DVAJU POSTUPAKA

U ovim simulacijama smo smatrali da je maksimalni broj promenljivih 115, ali prednosti DISCO postupka su očiglednije pri porastu broja nezavisnih promenljivih. Takva situacija nije nemoguća, pogotovo u poljima primene koja smo već navodili.

6. ZAKLJUČCI

U radu je stavljen naglasak na problem ocenjivanja koji se mogu javiti u logističkoj regresiji sa velikim brojem nezavisnih promenljivih. Pokazali smo da DISCO indeks, koji je do sada korišćen u analizi diskriminante, može poslužiti i kao koristan alat u logističkoj regresiji. Pokazali smo da se vreme ocenjivanja parametara logističkog modela značajno skraćuje ukoliko se koristi DISCO indeks umesto klasične postupne regresije. Takođe, parametri ocenjeni primenom novog postupka imaju vrednosti koje su bliske jedinici i ne postoji suštinska razlika između njih i vrednosti dobijenih primenom postupne logističke regresije.

Očigledno je da su neophodna dalja istraživanja, posebno u pogledu definisanja praga DISCO indeksa. Ta vrednost je ovde bila definisana emprijski, ali su neophodne dublje analize i simulacije u cilju definisanja uopštenog kriterijuma.

Konačno, naredna istraživanja će obuhvatiti i analize koje su neophodne da se ispita da li parametri ocenjeni primenom DISCO potupka omogućuju dobijanje zadovoljavajućih rezultata i za potrebe predikcije.

7. LITERATURA

5. Aguilera A.M., Escabias M., Valderrama M.J. (2006) *Using principal components for estimating logistic regression with high-dimensional multicollinear data,* Computational Statistics & Data Analysis, 50: 1905-1924.
6. Allard J., Choulakian V., LeBlanc R., MacNeill S., Mahdi S. (2000) *Discriminant Anlysis of Seal Data,* The Canadian Journal of Statistics, 28 (1), 205-212.
7. Ben-Akiva M., Lerman S. (1985) *Discrete Choice Analysis*: Theory and Application to Travel Demand, The MIT Press, Cambridge, Ma.
8. Camminatiello I, Lucadamo A. (2010) *Estimating multinomial logit model with multicollinear data,* Asian Journal of Mathematics and Statistics, vol 3(2), 93-101, 2010, ISSN: 1994-5418
9. Choulakian V., Almhana J. (2001) *An algorithm for nonmetric discriminant analysis*, Computational Statistics & Data Analysis, 35, 253-264. V. Choulakian and J. Almhana, “An Algorithm for Nonmetric Discriminant Analysis,” Computational Statistics and Data Analysis Journal, pp. 253-264, Jan. 2001.
10. Fisher R.A.(1936) *The use of multiple measurements in Taxonomic Problems*, Annals of Eugenics, 7, 179-188
11. Guttman L. (1988) Eta, DISCO, oDISCO and F. Psychometrika 53, 393-405.
12. R Development Core Team (2010). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

13. Raveh A. (1983) *Preference structure analysis: A nonmetric approach*. Pattern Recognition 16 (2), 253-259.
14. Raveh A. (1989) *A nonmetric approach to lineat discriminant analysis*. J. Amer. Statist. Assoc. 84, 176-183.
15. Simonetti B., Choulakian V. (2003) *Discriminat analysis for spectroscopic data*, Relazione invitata, Atti del Convegno intermedio SiS 2003 “Analisi Statistica Multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia”, Napoli 2003. ISBN: 88-8399-053-6.
16. Train K. (2003) *Discrete choice methods with simulation*. Cambridge University Press.

UZORAK - SA VEROVATNOĆOM ILI BEZ?

Sanja Konjik
Departman za matematiku i informatiku
PMF, Univerzitet u Novom Sadu
Trg D. Obradovića 4, 21000 Novi Sad, Srbija
sanja.konjik@dmi.uns.ac.rs

1. UZORAK UMESTO POPISA

Pretpostavimo da je potrebno prikupiti podatke o ukupnom broju računara koji se trenutno nalaze u svim državnim osnovnim i srednjim školama, iznosu troškova koji se mesečno potroše na životne namirnice po domaćinstvu nekog grada ili ukupnom broju noćenja ostvarenih u svim hotelima i pansionima u nekoj turističkoj oblasti tokom određenog perioda. Uvid u sve državne škole, ispitivanje svih domaćinstava u gradu ili provera lista gostiju svih hotela i pansiona u nekoj turističkoj oblasti bili bi previše skupi i dugotrajni. Otuda se čini prirodno izabrati određen broj škola, domaćinstava ili hotela i pansiona i za njih prikupiti sve relevantne podatke, na osnovu kojih će se dobiti slika o ukupnim vrednostima, ekvialentna rezultatima koji bi se dobili ispitivanjem svih škola, domaćinstava ili hotela i pansiona.

Svedoci smo da različite statističke agencije svakodnevno putem medija saopštavaju kvantitativne podatke o nacionalnom dohotku, stopi nezaposlenosti, političkom opredeljenju, stepenu obrazovanja, ostvarenoj dobiti iz uvoza i izvoza i sl. Neke statistike mogu se dobiti iz popisa, ali se većina zasniva na reprezentativnom uzorku. Tako se podaci o 100-milionskom narodu dobijaju ispitivanjem uzorka od svega nekoliko hiljada stanovnika.

U istraživanjima u kojima je potrebno prikupiti podatke o nekoj populaciji (velikom broju ljudi, kompanija, poljoprivrednih dobara i dr.), najčešće korišćena metodologija je ispitivanje uzorka umesto cele populacije. Uzorak je podskup populacije na kojem će se vršiti istraživanje. Ukoliko se napravi dobar odbir uzorka, obezbeđuju se precizni, pouzdani i korisni podaci, uz uštedu vremena, troškova i napora. S druge strane, loš uzorak može da dovede u pitanje validnost istraživanja i relevantnost izvedenih zaključaka.

Prema tome, dobar odbir uzorka, koji će što vernije oslikati one karakteristike cele populacije koje su od interesa za dato ispitivanje, jedan je od najvažnijih koraka u osmišljavanju i realizaciji naučnog istraživanja. U literaturi se navodi sledećih šest koraka u procesu odbira uzorka:

1. priprema istraživanja
2. izbor između popisa i uzorka
3. odabiranje između metoda uzorkovanja - verovatnosnog, neverovatnosnog ili kombinovanog
4. odabiranje oblika verovatnosnog, neverovatnosnog ili kombinovanog uzorka
5. određivanje obima (veličine) uzorka
6. odbir uzorka.

Veliki broj naučnika ističe da je ključ uspeha istraživanja priprema. Isto važi i za proces odabira i ispitivanja uzorka. Priprema obuhvata postavljanje ciljeva istraživanja, preciziranje značaja očekivanih rezultata istraživanja, određivanje željene preciznosti, odnosno greške koja će biti tolerisana, utvrđivanje raspoloživih sredstava za istraživanje, definisanje ciljne populacije, odabir istraživačkih i uzoračkih metoda. U drugom koraku je potrebno doneti odluku o tome da li ispitivati celu populaciju (popis) ili neki njen deo (uzorak). Ova je odluka najčešće uslovljena sredstvima i vremenom kojima istraživač raspolaže. Kada je doneta odluka o ispitivanju uzorka, na red dolazi odabir između dve osnovne metode uzorkovanja - verovatnosnog i neverovatnosnog. U verovatnosnom uzorku svaki elemenat ciljne populacije ima određenu pozitivnu verovatnoću sa kojom može biti izabran u uzorak. Time se omogućava razvoj teorije za ispitivanje osobina uzoračkih ocenjivača. Ako ovaj uslov nije ispunjen reč je o neverovatnosnom uzorku. U praksi se često kombinuju ove dve metode. Nakon što je odabrana metoda uzorkovanja, prelazi se na odredjivanje obima potrebnog uzorka i odabiranje konkretnog oblika uzorkovanja.

Kod verovatnosnog uzorka to mogu biti prost slučajni, stratifikovani, klaster ili sistematski uzorak, dok su neke od vrsta neverovatnosnog uzorka namerni, prigodni, kvota uzorak ili uzorak grudve snega.

U nastavku rada pažnja će se posvetiti trećem koraku u šemi odabira uzorka koja je gore navedena, tj. biranju između verovatnosnog i neverovatnosnog uzorka. Ukazaće se na prednosti i slabosti oba metoda, ali i na njihovu povezanost i isprepletanost.

2. TREĆI KORAK

Nakon što je u drugom koraku doneta odluka o ispitivanju uzorka umesto cele populacije, sledeći, treći, korak zahteva od istraživača da se opredeli da li će koristiti verovatnosno uzorkovanje, neverovatnosno uzorkovanje ili će kombinovati ova dva metoda. Podsetićemo da verovatnosna uzoračka procedura podrazumeva situaciju da se u uzorku sa određenom pozitivnom verovatnoćom može naći svaki elemenat populacije, pri čemu je osnovna karakteristika procesa odbira uzorka slučajnost. U suprotnom, ako se pojedini elementi populacije ne mogu naći u uzorku, verovatnoće odabira elemenata u uzorak se ne mogu precizno odrediti ili odabir elemenata u uzorak nije slučajan, reč je o neverovatnosnom uzorkovanju. Glavna prednost verovatnosnog uzorkovanja se ogleda u tome što poznavanje verovatnoća izbora svih elemenata populacije formira matematički aparat koji omogućava ispitivanje osobina ocenjivača, kao i procene uzoračkih grešaka. S druge strane, nepoznavanje verovatnoća odabira elemenata u uzorak dovodi do nužnih subjektivnih ocena, što predstavlja najveću slabost neverovatnosnog uzorkovanja. Pa čak i kada se neverovatnosno uzorkovanje pokazalo izuzetno efikasnim (u nekim pređašnjim istraživanjima), ne postoji nikakva garancija da će se isto ponoviti i u sličnim budućim studijama. Otuda se sasvim prirodno postavlja pitanje zašto bi se uopšte izučavali i koristili neverovatnosni uzorci. Pokušaćemo u nastavku da damo odgovor na ovo pitanje.

Osvrnimo se prvo na najvažnije oblike verovatnosnih i neverovatnosnih uzoraka u cilju što boljeg razumevanja njihovih prednosti i nedostataka. Pretpostavimo da je populacija konačna, tj. da sadrži N elemenata, i da je potrebno izabrati uzorak obima n iz populacije.

Najvažniji oblici verovatnosnih uzoraka su:

- Prost slučajni uzorak – osnovni oblik verovatnosnog uzorka kod kojeg svaki n-točlani podskup populacije ima istu verovatnoću sa kojom može biti izabran u uzorak. Najčešće je potrebno imati listu svih elemenata populacije kako bi se slučajnim izborom moglo odabrati n jedinica. Na primer, pretpostavimo da se želi čuti mišljenje studenata Novosadskog uzniverziteta o primeni Bolonjskog procesa u nastavi i da je određen obim uzorka od 300 studenata. Prost slučajni uzorak dobio bi se slučajnim odabirom 300 imena iz liste svih studenata Novosadskog univerziteta.
- Stratifikovani uzorak – vrsta verovatnosnog uzorka u kome je populacija podeljena u stratume (slojeve) iz kojih se zatim uzimaju prosti slučajni uzorci. Stratumi se najčešće formiraju od elemenata koji su slični u odnosu na obeležje koje se ispituje, pri čemu se svaki elemenat populacije nalazi u tačno jednom stratumu. Važno je napomenuti da se prost slučajni uzorak uzima iz svakog stratuma. Prema tome, stratifikacija, u opštem slučaju, povećava preciznost. U prethodnom primeru, stratumi mogu biti fakulteti Novosadskog univerziteta. Odgovarajući stratifikovani uzorak dobio bi se uzimanjem prostog slučajnog uzorka studenata sa svakog fakulteta.
- Klaster uzorak – dobija se grupisanjem elemenata populacije u klastere (grupe), koji postaju nove uzoračke jedinice, uzimanjem prostog slučajnog uzorka klastera, a zatim ispitivanjem svih elemenata odabranih klastera, ili ponovnim uzimanjem prostog slučajnog uzorka elemenata iz odabranih klastera. Klasteri su često prirodno određeni, na primer, razredi u školi. S obzirom da se prost slučajni uzorak bira na nivou klastera, a zatim se ispituju svi ili neki od elemenata izabranih klastera, ovaj oblik uzorkovanaj u opštem slučaju smanjuje preciznost. U našem primeru, prirodni klasteri bi bili fakulteti ili departmani. Klaster uzorak bi se dobio uzimanjem prostog slučajnog uzorka fakulteta ili departmana, a potom ispitivanjem svih studenata izabranih fakulteta ili departmana, ili biranjem novog prostog slučajnog uzorka studenata izabranih fakulteta ili departmana.

Važno je istaći da je u sva tri navedena slučaja proces odabira elemenata iz populacije u uzorak slučajan: kod prostog slučajnog uzorka elementi se biraju slučajno iz populacije, kod stratifikovanog je izbor elemenata iz stratuma slučajan, a kod klaster uzorka slučajno se biraju klasteri.

Navodimo sada i neke od najčešće korišćenih oblika neverovatnosnih uzoraka.

- Kvota uzorak je tip neverovatnosnog uzorka u kome se populacija deli na disjunktne podskupove prema obeležju koje se ispituje (kao kod stratifikovanog uzorka), određuje se njihova veličina, a potom se određuje potreban obim uzorka i broj elemenata svakog podskupa (kvota) koji treba uključiti u uzorak. Suštinska razlika u odnosu na stratifikovani uzorak se sastoji u tome što se u poslednjem koraku elementi iz svakog podskupa ne biraju korišćenjem tehnika verovatnosnog uzorkovanja, već je taj izbor prepušten slobodnom prosuđivanju istraživača. Na primeru stratifikovanog uzorka koji smo gore dali, odredila bi se kvota studenata sa svakog fakulteta koju treba ispitati, recimo 10% studenata sa svakog fakulteta, a kako će ti studenti biti izabrani, odluka je istraživača.
- Prigodni uzorak je uzorak koji istraživaču stoji na raspolaganju, koji je pogodan, zgodan. Na primer, profesori koji vode istraživanje zamole svoje studente da popune upitnik o primeni Bolonjskog procesa u nastavi. Očigledno je da ovakav uzorak nije, u opštem slučaju, reprezentativan, da većina jedinica populacije nema nikakvu šansu da bude uključena u uzorak i da ne postoje teorijske osnove da se izračuna greška koja proizilazi iz zaključivanja na osnovu ovakvog uzorka. S druge strane, prigodni uzorak je najjednostavniji za ispitivanje.
- Namerni uzorak je takođe oblik neverovatnosnog uzorka u kojem istraživač iz populacije bira one elemente u uzorak koji prema njegovoj proceni najviše doprinose ciljevima istraživanja. Za razliku od prigodnog uzorka, jedinice iz populacije nisu odabrane u uzorak zbog svoje raspoloživosti ili pogodnosti, već istraživač namerno odabira uzorak od elemenata koji ispunjavaju kriterijume koji su od značaja za istraživanje. U primeru anketiranja studenata o primeni Bolonjskog procesa u nastavi na Novosadskom univerzitetu, namerni uzorak bi mogao biti izabran od studenata koji redovno pohađaju nastavu.

- Uzorak grudve snega pripada neverovatnosnim uzoračkim procedurama koje se koriste samo u humanističkim istraživanjima. Prvo se odabira određen broj ispitanika koji će istraživača uputiti na nove ispitanike koje bi trebalo uključiti u uzorak. Pogodan je za ispitivanje retkih karakteristika populacije kao što su narkomanija, bolesti zavisnosti, AIDS, kriminal, seksualna orijentacija, prostitucija i sl.

Možemo zaključiti da je u verovatnosnom uzorkovanju pristrasnost istraživača isključena, a odabir elemenata koji će se naći u uzorku nezavisan od potreba ili želja istraživača. Međutim, kod neverovatnosnog uzorkovanja, odabir uzorka za ispitivanje je u većoj ili manjoj meri uslovljen subjektivnošću istraživača, čime se isključuje mogućnost koričćenja teorije verovatnoće i statistike u obradi prikupljenih podataka, izvođenju zaključaka i ocenjivanju greške nastale usled uzimanja uzorka umesto popisa. Navedimo osnovne slabosti neverovatnosnog uzorkovanja u poređenju sa verovatnosnim:

- nije pogodno u deskriptivnim i eksplikativnim istraživanjima
- ne omogućava određivanje preciznosti dobijenih ocena
- ne omogućava obradu podataka matematičkim aparatima
- pristrasnost prilikom izbora
- gubi se reprezentativnost
- izražena subjektivnost istraživača
- nije pogodno u heterogenim populacijama.

Pa opet, neverovatnosni uzorci su pogodniji za korišćenje od verovatnosnih uzorcaka u sledećim slučajevima:

- u eksplorativnim istraživanjima
- kada su raspoloživa sredstva za realizaciju istraživanja ograničena ili mala
- kada je raspoloživo vreme za realizaciju istraživanja kratko
- kada je potrebno ispitati specifične jedinice populacije
- kada nije poznata cela populacija
- kada je populacija homogena
- kada nije neophodan reprezentativan uzorak

- kada nije potrebno minimizovati pristrasnost prilikom izbora
- kada je istraživački tim mali ili nedovoljno obučen
- kada je populacija široko rasprostranjena
- kada su u pitanju komercijalna ispitivanja, ispitivanja javnog mnjenja ili ispitivanja retkih karakteristika populacije.

Sumirajući prethodno rečeno, možemo dati i tabelarnu šemu prednosti i nedostataka verovatnosnog i neverovatnosnog uzorkovanja:

TABELA 1 – ŠEMA PREDNOSTI I NEDOSTATAKA VEROVATNOSNOG I NEVEROVATNOSNOG UZORKOVANJA

Karakteristika istraživanja	**Verovatnosni uzorak**	**Neverovatnosni uzorak**
Eksplorativno istraživanje	nije pogodan	pogodan
Deskriptivno ili eksplikativno istraživanje	pogodan	nije pogodan
Kratko raspoloživo vreme za realizaciju istraživanja	manje pogodan	pogodan
Reprezentativnost uzorka	pogodan	nije pogodan
Ocena preciznosti uzorka	pogodan	nije pogodan
Minimalna pristrasnost prilikom izbora	pogodan	nije pogodan
Homogena populacija	manje pogodan	pogodan
Heterogena populacija	pogodan	nije pogodan
Mala raspoloživa sredstva za realizaciju istraživanja	manje pogodan	pogodan
Ispitivanje javnog mnjenja	manje pogodan	pogodan
Ispitivanje retkih karakteristika populacije (ljudi)	nije pogodan	pogodan
Komercijalna istraživanja	manje pogodan	pogodan

Na sličan način daćemo i tabelarni prikaz prednosti i nedostataka za pojedine tipove verovatnosnih i neverovatnosnih uzoraka:

TABELA 2 – PREDNOSTI I NEDOSTATCI ZA POJEDINE TIPOVE VEROVATNOSNIH UZORAKA

Verovatnosni uzorci	Prednosti	Nedostaci
Prost slučajni uzorak	Jednostavan	Visoki troškovi, zahtevan
Stratifikovani uzorak	Veća preciznost	Složen, visoki troškovi
Klaster uzorak	Niski troškovi, jednostavan	Manja preciznost

TABELA 3 – PREDNOSTI I NEDOSTATCI ZA POJEDINE TIPOVE NEVEROVATNOSNIH UZORAKA

Neverovatnosni uzorci	Prednosti	Nedostaci
Kvota uzorak	Veća reprezentativnost, manja uzoračka greška	Složen, visoki troškovi
Prigodni uzorak	Jednostavan, niski troškovi, vremenski efikasan	Nije reprezentativan, pristrasnost prilikom izbora
Namerni uzorak	Veća kontrola prilikom izbora	Složen, nije reprezentativan
Uzorak grudve snega	Pogodan za ispitivanja retkih karakteristika populacije	Dugotrajan, zahtevan

U praksi se najčešće kombinuju različiti tipovi verovatnosnih i neverovatnosnih uzoraka u cilju veće efikasnosti, boljih procena i smanjenja troškova.

DIJAGRAM GRANANJA VEROVATNOĆA, FORMULA POTPUNE VEROVATNOĆE I BAJESOVA FORMULA

Marko Obradović
Univerzitet u Beogradu
Matematički fakultet

Postoje različita mišljenja o tome kako predavati matematički sadržaj studentima nematematičarima. Po jednima, pošto tim studentima u budućoj profesiji matematika može biti samo alat, ne treba im objašnjavati suštinu, već im je dovoljno predstaviti matematiku kao niz gotovih formula koje treba zapamtiti i kao takve primenjivati. Drugi zagovaraju pristup da im treba objasniti, prilagođavajuć način izlaganja njihovom nivou znanja i potreba.

Razmotrižemo ta dva pristupa na primeru karakterističnih zadataka koji se javljaju u svakom osnovnom (uvodnom) kursu verovatnoće, koji uključuju upotrebu formule potpune verovatnoće ili Bajesove formule.

1. FORMULA POTPUNE VEROVATNOĆE

Zadatak koji zahteva primenu formule potpune verovatnoće može se shvatiti kao eksperiment s više etapa. Na primer:

- Najpre se slučajno bira kutija, pa se iz nje izvlači kuglica
- Najpre se slučjano ubacuju dve kuglice, pa se slučajno zatim izvlači kuglica

Ograničimo se na zadatke u kojima imamo dve etape. Onošto je u ovakvim zadacima dato, ili se (jednostavno) može izračunati, su verovatnoće potpunog sistema događaja za prvu etapu eksperimenta i uslovne verovatnoće događaja iz druge etape, pod uslovom onih iz prve. Traže se verovatnoće događaja iz druge etape.

Neka imamo n ishoda u prvoj etapi i m ishoda u drugoj. Kako su događaji iz druge etape dati svojim uslovnim verovatnoćama u odnosu na prvu etapu, to se njihova verovatnoća računa kao zbir verovatnoća preseka tog događaja i događaja iz prve etape, tj.

$$P\left(B_j\right) = P\left(A_1 B_j\right) + P\left(A_2 B_j\right) + \cdots + P\left(A_n B_j\right), j = 1, \ldots, m. \tag{1}$$

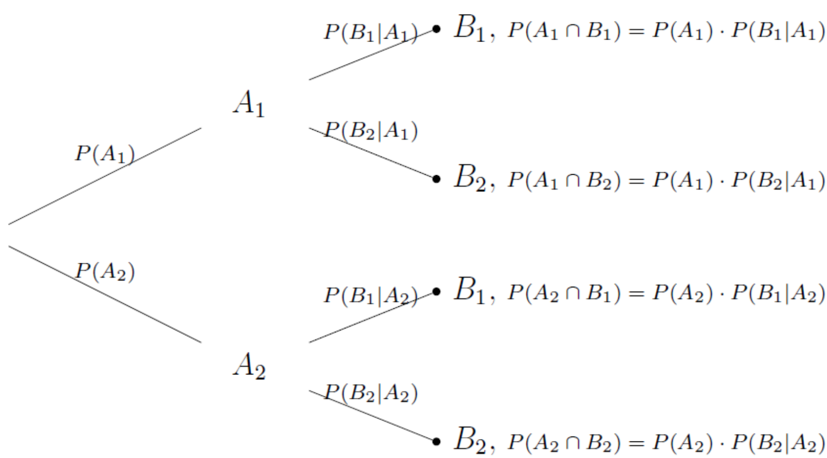
Kako se verovatnoća preseka dva događaja računa kao $P(AB) = P(A)P(B|A)$, tada važi

$$P\left(B_j\right) = \sum_{k=1}^{n} P\left(A_k\right) P\left(B_j | A_k\right), \tag{2}$$

što je poznato pod imenom *formula potpune verovatnoće*.

2. DIJAGRAM GRANANJA VEROVATNOĆA

Možemo se odlučiti da etape ovog eksperimenta predstavimo grafički. Tada je najpogodniji dijagram grananja verovatnoća (drvo verovatnoća) prikazan na Slici 1.



SLIKA 1. - DIJAGRAM GRANANJA VEROVATNOĆA ZA *n*=*m*=2

Na granama su ispisane verovatnoće koje vode tom granom, a u "čvorovima" događaji do kojih odgovarajuće grane vode. Na kraju svakog puta od "korena" do "lista" drveta dolazimo do preseka svih događaja koji su se našli na tom putu. Množeći verovatnoće na pređenim granama dobijamo verovatnoće preseka iz (1).

Sada verovatnoću događaja B_j računamo sabiranjem po svim eksperimentima u kojim se dogodio događaj B_j, tj. kao zbir svih preseka u kojima događaj B_j učestvuje. Ovaj zbir je, naravno, identičan onom u (2).

Pogledajmo kako to izgleda na primeru.

Primer 1. U jednoj kutiji se nalaze tri bele i jedna crna kuglica, u drugoj dve bele i tri crne, a u trećoj jedna bela i dve crne. Slučajno se bira jedna kutija i iz nje izvlači kuglica. Izračunati verovatnoću da je bela.

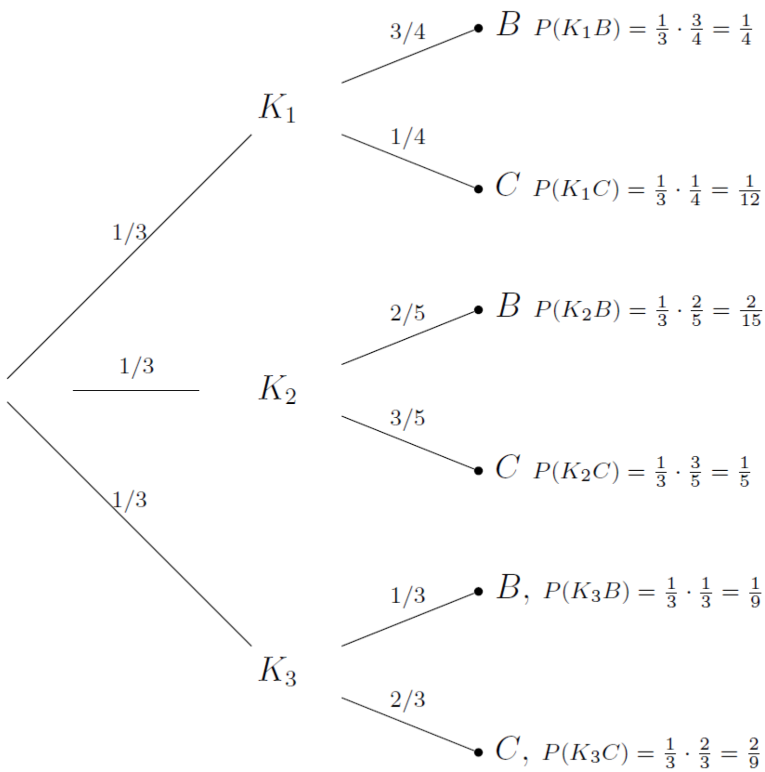
Rešenje Kako se kutije biraju slučajno, verovatnoće događaja prve etape (biranja kutije) su jednake i iznose $\frac{1}{3}$. Sve uslovne verovatnoće prikazane su na dijagramu na Slici 2.

Sada korišćenjem dijagrama traženu verovatnoću računamo kao zbir verovatnoća preseka na granama koje dolaze do B tj.

$$P(B) = \frac{1}{4} + \frac{2}{15} + \frac{1}{9} = \frac{89}{180}$$

Rešenje korišćenjem formule potpune verovatnoće bilo bi

$$P\,(B) = \sum_{k=1}^{3} P\,(K_k)\,P\,(B|K_{\,k}) = \frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{2}{5} + \frac{1}{3}\cdot\frac{1}{3} = \frac{89}{180}$$



SLIKA 2 -. DIJAGRAM GRANANJA IZ PRIMERA 1, K-KUTIJA, B-BELA, C-CRNA

3. BAJESOVE VEROVATNOĆE

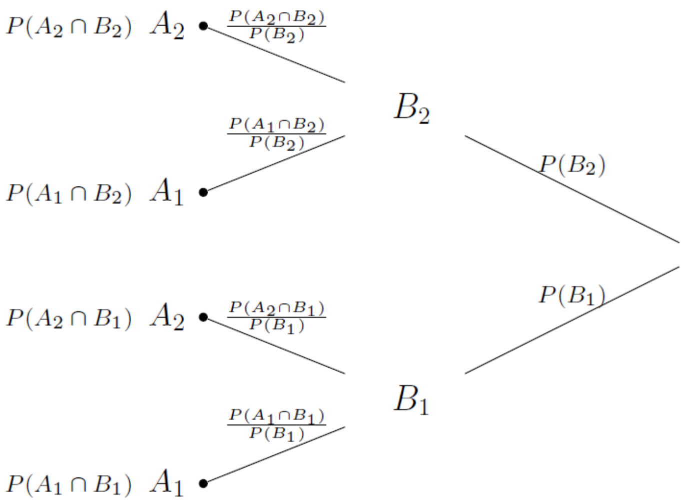
U zadacima navedenog tipa, vrlo često se traže “obrnute” verovatnoće, tj. verovatnoće događaja iz prve etape pod uslovom onih iz druge. Ove verovatnoće nazivamo Bajesovim verovatnoćama.

Standardan pristup njihovom računanju je Bajesova formula:

$$P\left(A_i \middle| B_j\right) = \frac{P(A_i)P\left(B_j|A_i\right)}{\sum_{k=1}^{n} P(A_k)P\left(B_j|A_k\right)} \tag{3}$$

Bajesove verovatnoće možemo računati konstrukcijom dijagrama “obrnutog” eksperimenta. Kako iz dijagrama (Slika 2.) možemo izračunati verovatnoće svih preseka i svih događaja B_j, formiranjem obrnutog dijagrama grananja na odgovarajućim granama tada nam se pojavljuju i Bajesove verovatnoće kao količnici verovatnoća preseka i verovatnoća događaja iz druge etape.

S obzirom na "obrnuti" eksperiment i grananje predstavljamo u obrnutom smeru (Slika 3).

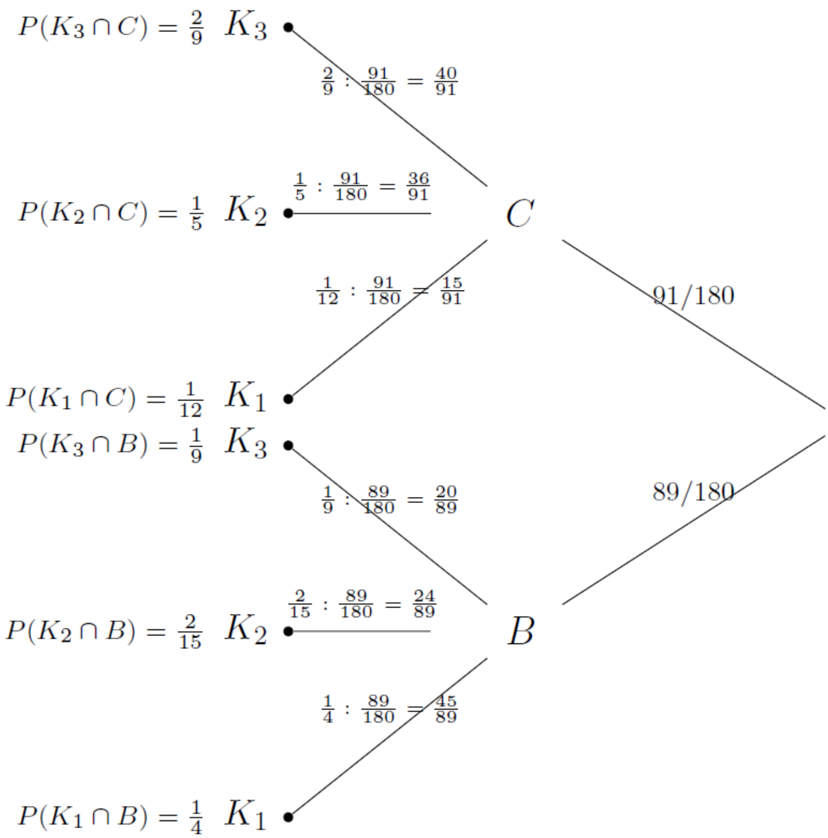


SLIKA 3. - OBRNUTI DIJAGRAM GRANANJA VEROVATNOĆA ZA $n=m=2$

Pogledajmo kako to izgleda na primeru.

Primer 2. Neka imamo isti ekperiment iz primera 1. Ako se zna da je izvučena crna kuglica, kolika je verovatnoća da je izvučena iz prve kutije?

Rešenje. Korišćenjem rezultata iz primera 1, formiramo obrnuti dijagram grananja (Slika 4).



SLIKA 4. - OBRNUTI DIJAGRAM GRANANJA IZ PRIMERA 2, K-KUTIJA, B-BELA, C-CRNA

Tražena verovatnoća nalazi se na grani koja spaja crnu kuglicu i prvu kutiju, a to je $\frac{15}{19}$.

Primenom Bajesove formule rešenje bi bilo:

$$P\left(K_1|C\right) = \frac{P(K_1)\cdot P(C|K_1)}{\sum_{k=1}^{3} P(K_k)\cdot P(C|K_k)} = \frac{\frac{1}{3}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot\frac{1}{5}+\frac{1}{3}\cdot\frac{2}{9}} = \frac{15}{91}$$

4. ZAKLJUČAK

Navedimo sada prednosti jednog i drugog metoda.

Prednosti dijagrama grananja verovatnoća:

- pogodan za školske primere i razumevanje suštine problema
- može se sagledati eksperiment u celini
- moguće je računati više relevantnih verovatnoća pomoću jednog dijagrama
- biće manje grešaka u radu jer nema pamćenja formula
- može se jednostavno primeniti i na primere sa više etapa

Prednosti klasičnog metoda (formule)

- pogodan je za veliki broj ishoda u pojedinim etapama jer se formula za razliku od dijagrama manje komplikuje
- zapis je kraći i računanje je brže
- deluje više “matematički”
- ima teorijski značaj

Iz iskustva stečenog na vežbama koje sam držao studentima biologije i meteorologije, poznato mi je da neki studenti lakše shvataju problem korišćenjem grafičkog prikazivanja, dok drugima više leži učenje formula. To zavisi kako od njihovih prirodnih sklonosti, tako i od prethodnog školovanja koje se dosta razlikuje od škole do škole. Iz ovog razloga, a kako svaki od pristupa ima svoje objektivne prednosti, mislim da je najbolje, ukoliko postoje uslovi, kombinovati i jedan i drugi.

ODREĐIVANJE VELIČINE UZORKA

Dušan Rakić
Poljoprivredni fakultet, Univerzitet u Novom Sadu

U različitim istraživanjima dolazi se do potrebe da se ispitaju izvesne karakteristike elemenata nekog skupa. Nemogućnost pristupa svakom elementu, vremenska ili materijalna ograničenja su najčešći razlozi zbog kojih su istraživači primorani da tražene karakteristike ispituju na nekom podskupu početnog skupa. Početni skup nazivamo populacija, njegovi elementi su jedinice (koje mogu biti ljudi, biljke, predmeti...), a sa N označavamo broj jedinica u populaciji. Traženi podskup populacije nazivamo uzorak i sa n označavamo broj jedinica u uzorku. Jasno, odabrani uzorak treba da bude dostupan i reprezentativan da bi mogli izvršiti istraživanje na njemu i na osnovu dobijenih rezultata izvesti zaključke za čitavu populaciju. Odabir veličine (obima) uzorka je veoma bitna odluka u istraživanju i bilo da su u pitanju ispitivanja koja uključuju rad sa ljudima ili ne, veličina uzorka može značajno da utiče na rezultat i opravdanost ispitivanja. Na primer, jako veliki uzorak povećava vreme i cenu ispitivanja, dok mali obim uzorka može značajno da smanji preciznost ispitivanja i pouzdanost rezultata. Mada se čini da je određivanje veličine uzorka prvi korak u istraživanju najčešće nije tako, posebno u dobro osmišljenim i pripremljenim ispitivanjima. Na primer, pogrešno je početi istraživanje pitanjem: "Koliki procenat od ukupne populacije treba da čini uzorak?", jer nam nekad broj jedinica populacije nije ni poznat (ako se ispituju izvesne karakteristike biljaka u šumi, ukupan broj biljaka u čitavoj šumi nam je verovatno nepoznat). "Magična formula" za obim uzorka ne postoji i broj jedinica koje će biti uključene u ispitivanje zavisi od više faktora. Dobro izvedeno istraživanje bi trebalo da obuhvati sledeće faze:

- preliminarna ispitivanja u kojima se analizira populacija u odnosu na traženu karakteristiku, kao i uslovi pod kojim se vrši istraživanje (vremenska i materijalna ograničenja),

- odabir metoda kojima ćemo prikupljati podatke iz uzorka i metoda kojima ćemo ih obrađivati,

nakon kojih sledi odluka o veličini uzorka. Navedene faze ćemo detaljnije objasniti, posebno naglašavajući kako one utiču na odabir veličine uzorka.

1. PRELIMINARNA ISPITIVANJA

Na početku rada istraživač bi trebalo da se detaljno upozna sa specifičnostima studije i uslovima u kojima se ona obavlja. Ova ispitivanja su često presudna za preciznost konačnog rezultata i cenu istraživanja i obično vremenski traju duže nego kasnije prikupljanje i obrada podataka. Faktore koje treba razmatrati u ovoj fazi rada u velikoj meri definiše samo istraživanje i njegove osobenosti, a kao smernice možemo navesti neke od njih:

- priroda populacije i jedinica na kojima se vrši ispitivanje,
- kolikim materijalnim sredstvima se raspolaže i koja su vremenska ograničenja,
- koji su ciljevi istraživanja,
- da li imamo nekih etičkih ili zakonskih ograničenja u radu.

1.1. Priroda populacije

Verovatno najveći uticaj na odabir veličine uzorka ima poznavanje prirode (karakteristika) populacije i jedinica koje je sačinjavaju. Pre svega veličina populacije i dostupnost njenih jedinica. Ako je ona izrazito mala i imamo pristup svim jedinicama, onda se postavlja pitanje da li ispitati čitavu populaciju ili uzimati uzorak. Dalje, vrlo je značajna procena homogenosti populacije. Ako je populacija homogena (jedinice su ravnomerno raspoređene u odnosu na ispitivanu karakteristiku), tada se može uzeti manji obim uzorka, kao i koristiti jeftinije metode uzorkovanja (klaster metoda). Na primer, ako kuvamo pasulj u loncima od 10 i 100 litara i ako je on dobro izmešan, tada je u oba slučaja dovoljno probati po jednu kašičicu da bi se proverilo da li je pasulj dobro začinjen. S druge strane, ako je populacija heterogena tada treba uzeti veći uzorak i koristiti metode uzorkovanja prilagođene ovom slučaju (stratifikovan uzorak).

1.2. Dostupna sredstva

Veoma važne informacije potrebne za dobro planiranje istraživanja su: kolika materijalna sredstva su obezbeđena, za koje vreme treba obaviti ispitivanje, kao i sa koliko brojnim i kako obučenim ljudskim potencijalom raspolažemo. Od ovih parametara direktno zavisi kvalitet istraživanja, posebno ako su uzorkovane jedinice fizički odvojene, ako je metoda prikupljanja podataka složena i zahteva posebnu stručnost (rad na biljkama, sa životinjama, delikatna pitanja u anketama). S obzirom da je često veličina uzorka direktno srazmerna troškovima obrade jedne jedinice (tada se do obima uzorka jednostavno dolazi deljenjem ukupnih sredstava sa kojim raspolažemo sa troškovima obrade jedne jedinice) preporučuje se da se pri ispitivanju jedinice posmatra više karakteristika, čime za gotovo iste troškove vršimo više istraživanja. Takođe, u cilju racionalizacije, često se primenjuje klaster metoda u kojoj u uzorak ulaze čitave podgrupe prostorno bliskih jedinica (svi đaci jednog razreda, svi stanovnici jedne zgrade, sve biljke na jednom delu parcele), pri čemu je važno napomenuti da klaster metoda, kao jedna od nepreciznijih metoda, traži povećanje veličine uzorka da bi se dobijeni podaci mogli smatrati relevantnim.

1.3. Ciljevi istraživanja

Veličina uzorka u velikoj meri zavisi od ciljeva i zadataka istraživanja. Ako istraživanje neće dati finalnu odluku o nekom pitanju, već će biti deo neke studije ili teorijskog razmatranja, uzorak može biti mali, dok u slučaju da ishod istraživanja može imati ozbiljne finansijske i društvene posledice (poslovanje banaka, metode lečenja u medicini, reforme u školstvu...), uzorak koji se posmatra mora biti velikog obima da bi rezultati bili što pouzdaniji. Takođe, uzorak treba da bude velikog obima ako želimo da dobijemo odgovore na pitanje kako reaguju podgrupe na ispitivanu karakteristiku, a ne samo ukupna populacija (ukoliko se ispituje dejstvo nekog medikamenta bitno je znati kako on deluje na ljude različite starosne dobi, polova, pacijente obolele od drugih bolesti...). Neke podgrupe koje su nam od interesa mogu biti retke ili teško dostupne i potreban je veći uzorak da bismo došli do što je moguće više podataka o njima. Dakle, veliki značaj i odgovornost istraživanja, kao i potreba za detaljnim i preciznim analizama, dovešće nas do većeg obima uzorka.

1.4. Etičke norme

U slučaju da se pri istraživanju (najčešće ispitivanju ljudi) izložimo riziku da se naruše etičke ili zakonske norme (ugrožavanje privatnosti, iznošenje podataka koji mogu ispitaniku narušiti društveni ugled (pitanja vezana za seksualnu orjentaciju, verski i politički stavovi) ili stvoriti probleme na poslu (odavanje poslovnih tajni)), tada se treba ograničiti na što je moguće manji obim uzorka, koji će nam biti dovoljan da postignemo traženu pouzdanost rezultata.

2. METODE PRIKUPLJANJA I OBRADE PODATAKA

Nakon detaljne analize problema i uslova u kojim radimo, a pre odluke o veličini uzorka, treba se odlučiti koju od metoda (ili kombinaciju više njih) koristiti u istraživanju, na koji način prikupljati i obrađivati podatke. Veoma je bitno znati da li se do podataka dolazi telefonskim putem, elektronskom poštom, neposrednom anketom, kao i da li se koristi verovatnosno ili neverovatnosno uzorkovanje. U zavisnosti od odabranog pristupa ukupan obim uzorka se određuje na početku rada (ako se koristi metoda u kojoj na početku odaberemo uzorak i dalje ga ne menjamo) ili se do njega dolazi u toku istraživanja (ukoliko se na osnovu nekih zadatih pravila i izbora u toku istraživanja menja obim uzorka).

2.1. Verovatnosno i neverovatnosno uzorkovanje

Pri odabiru metode kojom se bira uzorak prvo treba odlučiti da li se koristi verovatnosno ili neverovatnosno uzorkovanje. Među metodama verovatnosnog uzorkovanja (koje smo u tekstu već i spominjali) izdvajaju se: metoda prostog slučajnog uzorka, stratifikovanog uzorka i klaster metoda. Najrasprostranjenije metode kod neverovatnosnog uzorkovanja su: metoda pogodnog, namernog i kvota uzorkovanja.

Kod verovatnosnog uzorkovanja može se očekivati postojanje formula za veličinu uzorka koje se obično izvode iz formula za interval poverenja i one su sledećeg oblika:

$$n = \frac{z^2 \cdot s^2}{e^2},$$

gde z određuje nivo poverenja (obično se uzima 95%-tna sigurnost), s^2 je procena disperzije populacije oko srednje vrednosti, a e margina greške koju ne želimo da pređemo u ispitivanju. Vrednosti z i e se zadaju spram zahteva preciznosti i pouzdanosti rezultata istraživanja. Tipične vrednosti za njih su $z = 1{,}96$ (podrazumeva 95%-tni nivo poverenja, pri zakonitostima normalne raspodele) i $e = 0{,}03$ (gde posebno treba obratiti pažnju na to da li je e dato kao apsolutna ili relativna margina greške). Najteže je doći do vrednosti s^2, jer na početku istraživanja nemamo znanja o disperziji populacije oko srednje vrednosti i nju možemo dobiti na jedan od sledećih načina:

- koristeći pilot uzorak veličine 20 do 150 jedinica koji nam predstavlja preliminarno istraživanje iz kojeg dobijamo procenu veličina potrebnih za određivanje veličine uzorka,
- koristeći slična istraživanja, ranije obavljena, koja su dostupna u literaturi (vrlo je verovatno da se neko već bavio sličnom problematikom),
- jednostavno pogađajući vrednost disperzije na osnovu iskustva i poznavanja osobina populacije (pre svega homogenosti).

Ukoliko se ispituje proporcija neke karakteristike u odnosu na čitavu populaciju koristimo formule oblika:

$$n = \frac{z^2 \cdot p \cdot (1 - p)}{e^2},$$

gde je p procena proporcije (dobijena na jedan od navedenih načina za procenu disperzije). U Tabeli 1 su date veličine uzorka za neke karakteristične vrednosti proporcije i margine greške pri vrednosti $z = 1{,}96$.

TABELA 1 – VELIČINE UZORKA ZA NEKE KARAKTERISTIČNE VREDNOSTI PROPORCIJE I MARGINE GREŠKE PRI VREDNOSTI $z = 1{,}96$.

p \ e	0,01	0,03	0,05	1
0,01	380	42	15	4
0,1	3457	384	138	35
0,25	7203	800	288	72
0,5	9604	1067	384	96

Ako nismo u mogućnosti da dobijemo procenu *p*, tada je najsigurnije uzeti vrednost $p = 0{,}5$ za koju funkcija $p \cdot (1 - p)$ dostiže maksimum. Iz tog razloga se vrednost $n = 1067$ često preporučuje za veličinu uzorka (kada se ispituje proporcija, pri margini greške $e = 0{,}03$), bez obzira na to kolika je brojnost ukupne populacije.

Napomenimo i to da analizirajući faktor korekcije

$$\text{fpc} = 1 - \frac{n}{N},$$

koji je činilac u većini formula za ocenu parametara u različitim metodama teorije uzoraka, možemo doći do zanimljivih zaključaka u vezi sa obimom uzorka. Primetimo da je za malu vrednost N faktor korekcije mali (približava se vrednosti 0) i značajno utiče na procene, dok ako je N velika vrednost i broj jedinica u uzorku je značajno manji od N (situacija koju uglavnom imamo u ispitivanju javnog mnjenja), tada je faktor korekcije zanemarljiv i postaje značajan tek ako se u uzorak uzima više od 5% populacije. Dakle, često nam je bitnija apsolutna vrednost obima uzorka, a ne relativna vrednost u odnosu na ukupnu populaciju. Pokazaćemo još jedan primer (metoda slučajnog prostog uzorka) gde jako veliki obim populacije nije glavni faktor u određivanju obima uzorka. Varijansa srednje vrednosti uzorka, koja nam direktno određuje preciznost rezultata, računa se na sledeći način:

$$V(\bar{y}) = \frac{S^2}{n} \cdot \left(1 - \frac{n}{N}\right),$$

gde je S^2 varijansa populacije oko srednje vrednosti. Vidimo da je za fiksirano S^2 i $n = 100$ vrednost $V(\bar{y})$ približno ista za N = 100.000 i N = 100.000.000, odnosno da sa uzorkom od 100 jedinica postižemo približno istu preciznost u slučajevima gde se veličine ukupne populacije značajno razlikuju.

Metode verovatnosnog uzorkovanja pri kojima se ista jedinica obrađuje više puta u zadatim vremenskim intervalima zahtevaju uzorak većeg obima, jer ne možemo biti sigurni da ćemo sa svakom jedinicom uspeti da izvršimo čitav ciklus istraživanja (npr. kontrola dejstva nekog leka na pacijente tokom duže terapije, gde sprečenost ili, u ekstremnim slučajevima, smrt pacijenta onemogućavaju da se svi pacijenti podvrgnu tretmanu u svim fazama lečenja).

Kod neverovatnosnog uzorkovanja se mogu očekivati samo preporuke za veličinu uzorka u vidu konkretnih vrednosti u odnosu na tip istraživanja. Do ovih vrednosti se dolazi iskustveno, tokom dužeg perioda rada na sličnim problemima. Na primer, poznato je da za etnografska istraživanja u uzorak treba uzeti 35 do 50 jedinica, za marketinška istraživanja 200 do 2.500 jedinica, za nacionalna istraživanja 10.000 do 15.000 jedinica, itd.

Na osnovu iznetih činjenica o faktorima koji utiču na obim uzorka može se zaključiti da je dobar odabir veličine uzorka posledica izvršene detaljne analize populacije u odnosu na karakteristiku koja se ispituje, kao i dobrog odabira metode kojom se prikupljaju i obrađuju podaci u odnosu na prirodu problema, materijalne i ljudske resurse kojima se raspolaže.

JEDNA LEKCIJA – STOTINU STATISTIČKIH POJMOVA

Vesna Jevremović
Matematički fakultet
Univerzitet u Beogradu, Srbija

Uniformna raspodela iako najjednostavnija neprekidna raspodela ima važnu ulogu u Statistici jer je neophodna za modeliranje slučajnih veličina i stoga se koristi u metodama Monte Karlo. U ovom tekstu koji se može shvatiti i kao jedna lekcija iz Statistike, date su osnovne osobine uniformne raspodele, kao i jedan neparametarski test koji koristi uniformnu raspodelu, a zasnovan je na Lorencovoj krivoj i indeksu koncentracije.

1. UVOD

I podučavanje i učenje Statistike zahtevaju poznavanje raznih oblasti matematike: analize (granične vrednosti, redovi, integrali, funkcije jedne i više promenljivih), algebre (matrice, sistemi jednačina), numeričke analize (približno rešavanje jednačina),... ali takođe i korišćenje statističkih paketa i (poželjno) znanje da se na nekom programskom jeziku pišu novi programi.

U svakoj lekciji iz Statistike ima puno statističkih termina, od kojih su mnogi međusobno povezani. Stoga su razumevanje lekcije i kasnija primena sadržaja te lekcije mogući samo ako student dobro poznaje definicije statističkih pojmova, njihove osnovne osobine i međusobne relacije.

Skoro svaka lekcija iz Statistike sadrži puno matematičkih i statističkih pojmova, ali ponekad budemo i iznenađeni neočekivano velikim dijapazonom pojmova u nekoj lekciji, kao što je slučaj sa lekcijom o uniformnoj raspodeli koja sledi u nastavku ovog teksta. Da bismo skrenuli pažnju na korišćene pojmove, svaki od njih je posebno naglašen kurzivom pri prvom pojavljivanju u tekstu.

2. OSNOVNA SVOJSTVA NEPREKIDNE UNIFORMNE RASPODELE

Osnovna svojstva uniformne raspodele (ili pravougaone raspodele) su dobro poznata, pa ćemo ih ukratko nabrojati. Slučajna veličina *X* ima uniformnu raspodelu na intervalu [a, b], što označavamo *X: U(a, b),* ako je njena gustina raspodele jednaka $f(x)=\dfrac{1}{b-a},\ x\in[a,b]$, i 0 inače. Odatle se dobija da je funkcija raspodele $F(x)=\dfrac{x-a}{b-a},\ x\in[a,b]$.

Momenti uniformne raspodele su $m_r=\dfrac{1}{r+1}\dfrac{b^{r+1}-a^{r+1}}{b-a},\ r\in N$, a centralni momenti су $\mu_{2k-1}=0,\ \mu_{2k}=\dfrac{1}{2k+1}\left(\dfrac{b-a}{2}\right)^{2k},\ k\in N$. Mod raspodele nije jedinstven, a medijana je očigledno *(a+b)/2* zbog simetrije uniformne raspodele. Iz simetrije proizilazi i da je koeficijent asimetrije jednak 0, dok je koeficijent spljoštenosti једнак *−1.2*.

Ako se broj X na *slučajan način bira* iz intervala (0,1), onda *X: U(0, 1)*.

Uniformna raspodela je specijalni slučaj *beta raspodele*. Tačnije, uniformna raspodela *U(0, 1)* je raspodela $B_2(1, 1)$. Dodatno, ako je $(X_1,\dots,X_n)$ *slučajni uzorak* iz *populacije* sa *U(0,1)* raspodelom, tada je *statistika poretka k-tog ranga* $X_{(k)}$ slučajna veličina sa beta raspodelom $B_2(k,n-k+1)$.

Karakteristična funkcija za *X: U(a, b)* je $\varphi(t)=\dfrac{e^{itb}-e^{ita}}{it(b-a)},\ i=\sqrt{-a}$, dok je *generatorna funkcija momenata* $M(t)=\dfrac{e^{tb}-e^{ta}}{t(b-a)}$.

3. DISKRETNA I NEPREKIDNA UNIFORMNA RASPODELA

Slučajna veličina (s.v) *X* koja ima konačno mnogo vrednosti $x_1,\dots,x_n,$ a svaku sa jednakom *verovatnoćom* $P(X=x_j)=1/n$, gde je *n* pozitivan prirodan broj, se naziva diskretna uniformna raspodela.

Neke od veza diskretnih i neprekidnih uniformnih raspodela daju sledeće teoreme.

Teorema 1. *Neka su B: U(0,1) i Bj, j=1,2,...,N-1 nezavisne s.v. sa P(Bj=0) = P(Bj=1) =1/2. Tada* $\sum_{j=1}^{N-1} Bj/2^j + B/2^{N-1} : U(0,1)$ *i*

$\lim_{N\to\infty} \sum_{j=1}^{N-1} Bj/2^j : U(0,1)$.

Veza uniformne raspodele *U(0,1)* i diskretne uniformne raspodele oblika

$$Y : \begin{pmatrix} 0 & 1 & 2 & \cdots & 9 \\ 0.1 & 0.1 & 0.1 & \cdots & 0.1 \end{pmatrix} \qquad (*)$$

data je u sledećoj teoremi

Teorema 2. *Neka je* $\gamma = 0.u_1u_2...u_n...$ *realizacija s.v. X: U(0,1). Tada su* $u_1, u_2,...,u_n,...$ *nezavisne realizacije diskretne uniformne raspodele Y, i obrnuto.*

Teorema 2 može biti uopštena:

Neka je dat broj u sistemu sa brojnom osnovom r: $\gamma = 0. v_1v_2\ldots v_n\ldots$, tada su $v_1, v_2,\ldots,v_n,\ldots$ nezavisne realizacije diskretne uniformne raspodele V, pri čemu je

$$P(V=j)=1/r, j=0,1,\ldots,r-1,$$

i obrnuto..

4. UNIFORMNA RASPODELA I MAKSIMALNA VERODOSTOJNOST

Ako *X: U(a,b)*, gde su *a* i/ili *b* nepoznati, možemo ih *oceniti* na osnovu uzorka $(X_1,\ldots,X_n)$ iz te raspodele. S obzirom da uniformna raspodela nije *regularna u smislu Rao-Kramera*, tada *ocene* zasnovane na *metodi maksimalne verodostojnosti* mogu imati interesantna svojstva.

Primer 1. Razmotrimo prvo slučaj raspodele *U(0,b)*. Po metodi maksimalne verodostojnosti (MV) dobijamo ocenu $\hat{b} = \max_{1\le j\le n} X_j = X_{(n)}$, dok *metoda momenata* (MM) daje $\bar{b}_n = 2\bar{X}_n$. Imamo da je

$$E(\hat{b}) = nb/(n+1), D(\hat{b}) = nb^2/(n+2)(n+1)^2$$

$$E(\bar{b}_n) = b, D(\bar{b}_n) = b^2/3n\,.$$

Zaključujemo da je ocena po metodi MV *pristrasna*, ali *asimptotski nepristrasna i stabilna*. Ona je takođe *efikasnija* od *nepristrasne* ocene dobijene po MM. Nova ocena dobijena kao linearna transformacija ocene MV je $\frac{n+1}{n}\hat{b}$ i ona jeste nepristrasna, a njena *disperzija* je manja od disperzije nepristrasne ocene po MM. Pri tome, disperzija navedene ocene je manja od b^2/n što predstavlja *donju granicu diperzije* nepristrasnih ocena koju bismo dobili po *nejednakosti Rao-Kramera*. A što se tiče statistike $\hat{b} = \max_{1\le j\le n} X_j = X_{(n)}$ ona je *kompletna dovoljna statistika* za parametar *b*; takođe je, kao što smo već napomenuli *konzistentna* (stabilna) ali i *stabilna u srednje-kvadratnom* jer je $\lim_{n\to\infty} E(\hat{b}-b)^2 = 0$. Koristeći *statistike poretka* možemo odrediti i druge nepristrasne ocene za b, kao što je $\hat{b}_1 = (n+1)\min_{1\le j\le n} X_j = (n+1)X_{(1)}$, ili, u slučaju neparnog *n*, *n=2k+1*, ocena zasnovana na *uzoračkoj medijani* $\hat{b}_2 = 2X_{(k+1)}$. Uz sve navedeno, pošto je raspodela ocene po MV poznata, mogu se određivati i *intervali poverenja* za parametar b, sa željenim *nivoom poverenja*.

Primer 2. U ovom primeru posmatramo raspodelu obeležja *X: U(a–1/2,a+1/2)*. Tada ocena po metodi MV *nije jedinstvena*, i to tako da svaka statistika *V* koja zadovoljava nejednakosti

$$\max_{1\le j\le n} X_j - \frac{1}{2} \le V \le \min_{1\le j\le n} X_j + \frac{1}{2},$$

može biti ocena po metodi MV za *a*.

Primer 3. U slučaju *familije raspodela X: U(–a,a), a>0*, koja *nije kompletna*, imaćemo da je *dovoljna statistika* za *a* dvodimenzionalna $(X_{(1)},X_{(n)})$. Interesantno je napomenuti da je *koeficijent korelacije* s.v. $X_{(1)}$ i $X_{(n)}$ jednak *1/n,* što znači da se zavisnost tih s.v. smanjuje kad se *n* uvećava. Ocena po metodi MV za *a* data je izrazom $\hat{a}=\max(-X_{(1)},X_{(n)})$ i predstavlja *minimalnu dovoljnu statistiku* za taj parametar.

5. UNIFORMNA RASPODELA I MONTE KARLO METODE

Monte Karlo metode koriste *pseudo-slučajne brojeve* radi odredjivanja osobina nekih funkcija ili skupova funkcija, izračunavanja vrednosti odredjenih integrala, ili proučavanja nekih složenih sistema. Pseudo-slučajni brojevi su kompjuterski generisani nizovi brojeva koji imaju *statistička svojstva* kao *nizovi nezavisnih slučajnih veličina* koje sve imaju istu uniformnu raspodelu, diskretnu ili neprekidnu.

Modeliranje slučajnih veličina je prvi korak u Monte Karlo metodama, a za taj korak je neophodna uniformna raspodela. Metode modeliranja slučajnih veličina zavise od prirode tih slučajnih veličina, ali se u svakom slučaju koriste modelirane vrednosti uniformne raspodele. Ako je to vrednost diskretne uniformne raspodele (*) onda je nazivamo *"slučajna cifra"*. Veoma je interesantna činjenica da transcedentni brojevi kao što su e, π,… predstavljaju generatore slučajnih cifara, u smislu da nizovi njihovih cifara imaju statistička svojstva nizova nezavisnih jednako raspodeljenih slučajnih veličina sa uniformnom raspodelom (*).

Svaka vrednost slučajne veličine sa uniformnom raspodelom *U(0,1)* se naziva *"slučajan broj"*, i u daljem će biti označena sa γ. Radi preciznosti treba reći da se u Monte Karlo metodama ne koriste slučajni, nego pseudoslučajni brojevi, tj. nizovi brojeva iz intervala *(0,1)* koji imaju statistička svojstva kao slučajni uzorak iz uniformne raspodele. Ti brojevi se generišu kompjuterski primenom odgovarajućih formula. Uobičajena oznaka/naredba kojom se dobijaju takvi brojevi je RAND ili RND. Tačnost Monte Karlo metoda se, u opštem slučaju, poboljšava sa povećanjem broja modeliranih vrednosti slučajne veličine, a to znači i porast broja upotrebljenih pseudoslučajnih brojeva.

Osobine uniformne raspodele date u sledećoj teoremi se često koriste pri modeliranju slučajnih veličina.

Teorema 3.

1) *Ako X: U(0,1), onda je 1−X: U(0,1),*
2) *Ako slučajna veličina X ima funkciju raspodele F(x), onda slučajna veličina F(X) ima uniformnu raspodelu na intervalu [0,1], tj. F(X): U(0,1) i*
3) *Ako X: U(0,1), onda je a· X+b: U(a,b), za proizvoljne realne brojeve a i b.*

Metode modeliranja slučajnih veličina su ukratko opisane u nastavku:

a) Neka je *X* diskretna slučajna veličina (s.v.) sa raspodelom

$$X:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ p_1 & p_2 & p_3 & \cdots & p_n \end{pmatrix},\ \sum_{k=1}^{n} p_k = 1\,. \tag{1}$$

Neka je γ jedan (pseudo)slučajan broj. Ako je $\gamma \le p_1$, onda smatramo da se realizovala vrednost x_1 s.v. *X*. Ako je $p_1 < \gamma \le p_1 + p_2$, onda smatramo da se realizovala vrednost x_2 s.v. *X*. Ako je $p_1 + p_2 < \gamma \le p_1 + p_2 + p_3$, onda smatramo da se realizovala vrednost x_3 s.v. *X*, itd. Na taj način za dobijanje svake realizacije s.v. *X* se koristi po jedan (pseudo)slučajan broj .

Ista ideja se primenjuje za modeliranje ishoda (realizacija) nekog *slučajnog događaja* koristeći odgovarajuću s.v. Neka je *p* verovatnoća događaja *A* u nekom eksperimentu, i neka je definisana s.v. $I_A:\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, tzv. *indikator*. Realizacije te s.v. dobijamo kao što je prethodno opisano, pa ako se dobije 1, smatramo da se *događaj A realizovao* u tom eksperimentu, a ako dobijemo 0, smatramo da se *A* nije realizovao.

To je opšti pristup modeliranju diskretnih s.v, mada ima i puno specijalnih pristupa u zavisnosti od svojstava s.v. koja se modelira. Tako na primer, ako treba modelirati vrednosti s.v. sa *binomnom raspodelom B(n,p),* onda je pogodno koristiti reprezentaciju te s.v. kao zbira n nezavisnih, jednako raspodeljenih indikatora, nego računati verovatnoće iz zakona raspodele, pa postupati kao za s.v. (1).

Primer 4. "Bacanje kockice"

a) Ishode takvog eksperimenta možemo dobiti korišćenjem neprekidne uniformne raspodele. Ako imamo niz slučajnih brojeva, zaokruženih na dve decimale: 0.23, 0.32, 0.98, 0.85,… onda dobijamo rezultate: 2, 2, 6, 6, … (jer je $1/6<0.23\leq 2/6$, što znači da je ishod "2", itd…)
b) Drugi način da se modeliraju ishodi eksperimenta je korišćenjem diskretne uniformne raspodele. Ako imamo niz slučajnih cifara: 2, 3, 3, 2, 9, 8, 8, 5,…, onda dobijamo ishode: 2, 3, 3, 2, 5,… (0, 7, 8 i 9 se iz očiglednih razloga ne koriste).

b) Neka je *X diskretna s.v. sa beskonačno mnogo vrednosti*

$$X:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots & p_n & \cdots \end{pmatrix},\ \sum_{k=1}^{\infty} p_k = 1.$$

Neka je *n* prirodni broj za koji važi $p_{n+1}+p_{n+2}+\ldots<\delta$, gde je δ proizvoljno mali pozitivan broj. To je uvek moguće odrediti, jer je red $\sum_{k=1}^{\infty} p_k = 1$ konvergentan. Dalje, umesto s.v. *X* koristimo

$$X_Z:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ p_1^* & p_2^* & p_3^* & \cdots & p_n^* \end{pmatrix},$$

gde je $p_1^* = p_1, \ldots, p_{n-1}^* = p_{n-1}$, $p_n^* = 1-(p_1+\ldots+p_{n-1})$. Realizacije te s.v. X_Z (*zasečena s.v.* X) modeliramo prema proceduri opisanoj pod a) i smatramo ih za realizovane vrednosti s.v. *X*. Na taj način ne možemo dobiti ni jednu od vrednosti $x_{n+1}, x_{n+2}, \ldots$ za *X*, ali je verovatnoća da s.v. *X* ima neku od tih vrednosti manja od δ, koji je izabran da bude mali pozitivan broj, npr. $\delta=0.001$ ili slično, a u zavisnosti od ciljeva i potreba simulacije.

c) Neka je *X* neprekidna s.v. sa funkcijom raspodele *F(x)*, i neka ta funkcija ima inverznu. Realizacije s.v. *X* se tada mogu modelirati na osnovu sledeće teoreme:

Teorema 4. *Neka je data s.v. X čija je funkcija raspodele F(x) neprekidna, strogo monotona, sa inverznom funkcijom* F^{-1}. *Neka s.v. Y ima uniformnu raspodelu U(0,1). Tada s.v.* $F^{-1}(Y)$ *ima funkciju raspodele F(x).*

U ovom slučaju se za dobijanje svake modelirane vrednosti *x* za s.v. *X* koristi po jedan slučajni broj *γ*, po formuli $x = F^{-1}(\gamma)$.

Primer 5. Kao primer primene Teorema 3 i 4 dajemo formulu za modeliranje vrednosti slučajne veličine sa *eksponencijalnom raspodelom*. Ako *X: ε*(*λ*), onda je njena modelirana vrednost $x = \frac{-1}{\lambda} ln(\gamma)$.

Navedeni postupak se može, sa manjim izmenama, primeniti i na slučaj prekidnih funkcija raspodele, kao i onih koje imaju intervale konstantnosti. Ako pak nema inverzne funkcije raspodele, ili je izraz za nju suviše komplikovan, primenjujemo neku drugu metodu.

d) Neka je *X* neprekidna s.v. sa gustinom raspodele (g.r.) *g* koja je definisana i ograničena na konačnom intervalu. Realizacije takve s.v. mogu se modelirati tzv. *metodom odbacivanja ili Nojmanovom metodom* koja je zasnovana na sledećem tvrđenju:

Teorema 5. *Neka je g.r. g(x) neke s.v. X definisana na konačnom intervalu (α,β) i neka postoji pozitivan broj M tako da* $g(x) \leq M, x \in (\alpha, \beta)$. *Neka su* x_T *i* y_T *modelirane realizacije s.v. U(α,β) i U(0,M), u tom redosledu. Ako važi* $y_T < g(x_T)$, *tada je* x_T *realizovana vrednost s.v. X.*

Ako ne važi $y_T < g(x_T)$, tada treba modelirati nov par x_T i y_T. Stoga se broj potrebnih (pseudo)pseudoslučajnih brojeva za modeliranje jedne realizacije neke s.v. po ovoj metodi ne može unapred predvideti.

Realizacije s.v. *U(α,β)* i *U(0,M)* se dobijaju na osnovu relacije *a*· *γ*+*b* kojom se dobija realizacija s.v. *U(a,b)* ako je *γ* vrednost s.v. *U(0,1).*

Ovu metodu dugujemo Nojmanu (Neumann), i ona, uz izvesne izmene, može da se primeni na proizvoljnu g.r. Ideja je da se odredi s.v. X_Z koja se sa verovatnoćom bliskom 1, poklapa sa s.v. X koju treba modelirati, a da pritom za X_Z važe uslovi Teoreme 5. Neka *X* ima g.r.

g(x), a<x<b, pri čemu *a* može biti $-\infty$, i/ili *b* može biti ∞, pri čemu *g(x)* ne mora biti ograničena na intervalu *(a,b)*. U svakom slučaju mora važiti $\int_a^b g(x)dx = 1$. Pretpostavimo da je na konačnom intervalu (a',b'), $(a',b') \subset (a,b)$ funkcija *g(x)* ograničena. Ukoliko je potrebno posmatra se unija svih takvih konačnih intervala na kojima je gustina raspodele *g(x)* ograničena. I sada definišemo s.v. X_Z (zasečena s.v. X) tako da je $X_Z = X$ na (a',b'), i 0 inače. Gustina raspodele $g_Z(x)$ s.v. X_Z zadovoljava

$$g_Z(x) = c \cdot g(x) = \left[\int_{a'}^{b'} g(t)dt \right]^{-1} \cdot g(x),\ a' < x < b'$$

Konstanta c je veća od 1, i $g_Z(x) > g(x)$ važi na (a',b'). Interval (a',b') je izabran tako da važi $1 - \int_{a'}^{b'} g(x)dx < \delta$, gde je δ proizvoljno mali pozitivan broj.

Tada se postupak modeliranja s.v. *X* ostvaruje tako što se Nojmanovom metodom modeliraju vrednosti X_Z i dobijene vrednosti smatraju za realizacije s.v. *X*, s obzirom da se *X* i X_Z poklapaju sa verovatnoćom *1*− δ.

e) <u>Neka je *X* slučajna veličina sa *normalnom raspodelom*.</u> Metoda inverzne funkcije nije primenljiva, a ni Nojmanova metoda se ne koristi za modeliranje normalne raspodele. Postoji više različitih metoda za modeliranje normalne raspodele od kojih je jedna zasnovana na *centralnoj graničnoj teoremi*, i korišćenju (pseudo)slučajnih brojeva na način koji ćemo objasniti u nastavku. Neka su $Y_1, Y_2, \ldots$ nezavisne s.v. sa uniformnom raspodelom *U(0,1)*. Zbir $S_n = \sum_{j=1}^{n} Y_j$ će imati očekivanje i disperziju redom $E(S_n)=\frac{n}{2}$, $D(S_n)=\frac{n}{12}$, a prema centralnoj graničnoj teoremi za s.v.

$$S_n^*=\frac{S_n-E(S_n)}{\sqrt{D(S_n)}}=\sqrt{\frac{3}{n}}\sum_{j=1}^{n}\left(2Y_j-1\right)$$

važi

$$P\left\{S_n^*<x\right\}\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt,\ n\to\infty$$

Ova *konvergencija u raspodeli* je brza i već za *n=12* razlika između S_n^* i *N(0,1)* je dovoljno mala, tako da uzimamo $\xi^{(12)}=\sum_{j=1}^{12}\gamma_j-6$ kao realizovanu vrednost s.v. sa *N(0,1)* raspodelom. Na taj način je potrebno 12 (pseudo)slučajnih brojeva za jednu realizaciju s.v. sa *N(0,1)* raspodelom. Napominjemo da se već za *n=6* dobija zadovoljavajuća preciznost. U opštem slučaju preciznost se još povećava ako se koristi: $S_n^*+\frac{1}{20n}\left((S_n^*)^3-3S_n^*\right)$, za svako n.

Ima i mnogo drugih metoda za modeliranje normalne raspodele, i u nastavku dajemo jednu od tih metoda koja koristi unformnu raspodelu.

Neka su α_1 i α_2 nezavisne s.v. sa U(0,1). Tada su ξ i η određene relacijama :

$$\xi=\sqrt{-2\log\alpha_1}\cdot\cos(2\pi\alpha_2)$$
$$\eta=\sqrt{-2\log\alpha_1}\cdot\sin(2\pi\alpha_2)$$

nezavisne s.v. sa *N(0,1)* raspodelama. Ovaj postupak se može uopštiti na *n-dimenzionalni slučaj.*

6. DOBIJANJE UZORAKA I REGENERISANJE UZORAKA

Ako treba dobiti *slučajni uzorak sa vraćanjem obima n* iz konačne populacije koja ima N elemenata, onda se elementi populacije numerišu od 1 do N, pa se modelira n realizacija s.v (1) sa vrednostima 1, 2, ..., N, na prethodno opisan način. A ukoliko je potreban *uzorak bez vraćanja* tada od svih modeliranih vrednosti uzimamo samo one koje su međusobno različite dok ne dobijemo željeni broj elemenata u uzorku.

Metoda modeliranja diskretnih s.v. sa konačno mnogo vrednosti se primenjuje i u metodi testiranja koja se zasniva na korišćenju uzoraka generisanih na osnovu jednog raspoloživog uzorka, tzv. *butstrep postupak (“bootstrap” procedure).* Pretpostavimo da treba “generisati” uzorak sa vraćanjem obima n_1 na osnovu raspoloživog uzorka obima *n*. Tada vrednosti datog uzorka $x_1, \ldots, x_n$, uzimamo kao vrednosti s.v. *X* u (1), bez obzira da li među njima ima istih, i svakoj vrednosti dodeljujemo verovatnoću p_j, $j=1,n$ koja je jednaka *1/n*.

Primer 6. Pretpostavimo da imamo uzorak 2, 3, 5, 4, 4. Odgovarajuća diskretna s.v. X ima raspodelu P(X=2)=P(X=3)=P(X=5)=1/5, i, naravno, P(X=4)=2/5. Neka su dati slučajni brojevi: 0.23, 0.42, 0.15, 0.78, 0.91, 0.32, 0.47, 0.11, 0.77, 0.08. Na osnovu njih “novi” uzorak obima 10, dobijen od datog uzorka je: 3, 5, 2, 4, 4, 3, 5, 2, 4, 2.

7. UNIFORMNA RASPODELA I STATISTIČKI TESTOVI

Ukoliko hoćemo da *testiramo hipotezu* da su podaci iz uzorka iz populacije sa uniformnom raspodelom *U(a,b)*, gde su *a* i/ili *b* nepoznate vrednosti, tada se može primeniti neki *neparametarski test* kao što je *χ2 test*, ili, ako su *a* i *b* poznati, *test Kolmogorova*. Moguće je koristiti i neke varijante *parametarskih testova* i porediti *moć* tih testova.

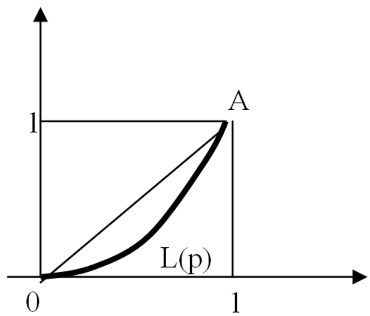
Pri testiranju *generatora slučajnih brojeva* neki testovi koriste uniformnu raspodelu kao raspodelu dobijenih brojeva/cifara. To je slučaj sa *testom frekvencija* (kad je raspodela (*) raspodela prema nultoj hipotezi)*, testom parova, testom trojki, …*

Pri proučavanju *raspodele prihoda* M.O. Lorenc je uveo novu funkciju, koja je kasnije po njemu nazvana: Lorencova linija ili Lorencova kriva. Ta kriva i sa njom povezan *koeficijent koncentracije* mogu biti primenjeni u statistici i kao osnova za *test saglasnosti sa datom raspodelom*. Gail i Gastwirth razvili su 1978. jedan takav test za *Laplasovu raspodeli*, a zatim proširili taj rezultat i na testiranje eksponencijalne raspodele. Na osnovu tog pristupa ovde ćemo otići korak dalje i pokazati kako se može u nultoj hipotezi uzeti i uniformna i bilo koja druga raspodela kao raspodela obeležja.

Neka je *X nenegativna s.v.* sa vrednostima u *[0,a)*, pri čemu *a* može biti beskonačno i neka postoji *očekivanje* μ*=E(X)*. Neka je *F(x)* funkcija raspodele za X, pri čemu pretpostavljamo da je ta funkcija strogo rastuća i *neprekidno diferencijabilna*, i označimo sa *G(x)* njenu inverznu funkciju. Lorencova krive je funkcija

$$L(p)=\frac{\int_0^{G(p)} t\,dF(t)}{\mu},\ 0\le p<1.$$

Veoma važno svojstvo Lorencove krive je da svaka raspodela koja ima konačno očekivanje jedinstveno određuje svoju Lorencovu krivu do na transformaciju skaliranja (Thompson, 1976).



SLIKA 1. - LORENCOVA KRIVA

Indeks koncentracije za populaciju sa raspodelom *F(x)* i Lorencovom linijom *L(p)* jednak je dvostrukoj površini izmedju *L(p)* i odsečka *OA*, tj.

$$C(X)=2\int_0^1 (p-L(p))dp \quad .$$

Odgovarajuća *uzoračka statistika*, često označena kao *Đinijev indeks*, je

$$G = \frac{\sum_{i,j=1}^{n} \left| X_i - X_j \right|}{2n(n-1)\overline{X}_n},$$

gde je $(X_1,\dots,X_n)$ prost slučajni uzorak obima n, a $\overline{X}_n$ je *uzoračka sredina*.

Indeks koncentracije za uniformnu raspodelu je *1/3*, pa ako hoćemo da testiramo *nultu hipotezu* H_o da je uzorak iz populacije sa *U(0,1)* raspodelom, a sa *pragom značajnosti* α možemo primeniti proceduru: najpre izračunati realizovanu vrednost g indeksa koncentracije na osnovu realizovanog uzorka $(x_1,\dots,x_n)$ i ako je $\left| g - 1/3 \right| < \alpha$ tada ne odbacujemo hipotezu H_o. *Alternativna hipoteza* H_1 je da raspodela nije uniformna. Dalje, ako s.v. X ima funkciju raspodele F, tada s.v. $Y=F(X)$ ima uniformnu raspodelu *U(0,1)* pa jednakost

$$P\left(X \in \left[a,b\right)\right) = P\left(F(X) \in \left[F(a),F(b)\right)\right)$$

očigledno važi za sve realne brojeve a, b, $a<b$. Na taj način ako hoćemo da testiramo hipotezu H_o da X ima funkciju raspodele F_o, a na osnovu uzorka $(X_1,\dots,X_n)$, tada možemo koristiti indeks koncentracije i testirati hipotezu da $F_o(X)$ ima uniformnu raspodelu *U(0,1)* za dati prag značajnosti α.

8. LITERATURA

1. Dagum, C. (1985) *Lorenz curve*, Encyclopaedia of Statistical Sciences, Vol. 5, S. Kotz; N. L. Johnson and C. B. Read (editors), p. 156-161, New York, J. Wiley
2. Gail, M.H. and Gastwirth J.L. (1978) *A scale-free goodness of fit test for the exponential distribution based on Gini statistic*, J. R. Statist. Soc. B, 40, No. 3, p. 350-357

3. Gail, M.H. and Gastwirth J.L. (1978) *A scale-free goodness of fit test for the exponential distribution based on the Lorenz curve*, J. Amer. Statist. Assoc., 73, No. 364, p. 787-793
4. Jevremovic, V. (1998) *The index of concentration and the goodness of fit*, Compstat 98, XII Symposium on Computational Statistics, Bristol, England, 24-28 August 1998
5. Hogg, R.V., McKean J.W., Craig, A.T. (2005) *Introduction to Mathematical Statistics*, Pearson Education International, printed in USA
6. Larsen, R. J., Marx, M.L. (2006) *An Introduction to Mathematical Statistics and Its Applications*, Pearson International Edition International, printed in USA
7. Соболь, И. И. (1973) *Численные методы Монте Карло*, Наука, Москва (на руском)
8. Đorić, D., Jevremović, V. i drugi (2007) *Atlas raspodela*, Građevinski fakultet, Beograd (на српском)
9. Jevremović, V. (2010) *Uniform Distribution in Statistics*, u International Encyclopedia of Statistics (red. M. Lovrić), Springer