Лазар Тењовић

Robustnost statističkih postupaka i robustni ocenitelji parametara lokacije, skale i povezanosti dveju kvantitativnih varijabli

1. Robustnost statističkih postupaka-osnovni pojmovi

Pojam robustnosti statističkih postupaka u ovom tekstu razmotrićemo imajući u vidu sledeća tri aspekta:[1]

1. robustnost statističkih testova za testiranje hipoteza;
2. robustnost parametara i
3. robustnost ocenitelja parametara, tj. statistika kojima ocenjujemo parametre..

Robustnost statističkih testova za testiranje hipoteza

Matematički (probabilistički) modeli klasičnih statističkih postupaka koji se najčešće koriste u psihologiji (t-test, analiza varijanse i regresiona analiza po principu najmanjih kvadrata) uključuju pretpostavke za koje se, samim korišćenjem određenog statističkog postupka, implicitno prihvata da *dopustivo dobro aproksimiraju* realnost ispitivanih fenomena. Naglasak u tome je na *aproksimativnosti*, tj. na istraživačima dobro poznatoj činjenici da realnost psiholoških fenomena manje ili više odstupa od matematičkih modela koji su u osnovi određenog statističkog postupka. Statistička sredstva predstavljaju jedno od osnovnih "oruđa" (srećom, ne jedino) istraživačima u psihologiji. Stoga, nepostojanje "dopustivo dobre aproksimacije" proučavane realnosti pretpostavkama matematičkog modela na kojima se zasniva određeni statistički postupak može imati dalekosežne posledice na razvoj psihološke nauke. Klasični statistički postupci koji se najčešće primenjuju u analizama podataka psiholoških istraživanja izvedeni su postuliranjem određenih idealizovanih uslova od kojih je "normalnost raspodele varijable X u populaciji" sastavni deo matematičkih modela većine klasičnih parametrijskih testova. (Normalna raspodela je, na taj način, psihološkim rečnikom rečeno, postala vrsta "statističkog arhetipa" za ogromnu većinu populacija na čijim uzorcima su primenjivane statističke metode).[2]

---

[1] Termin robustan potiče od latinske reči *robustus* što znači jak, snažan, krepak, čvrst.

[2] Ilustracije radi, nulta distribucija uzorkovanja t-statistika koji se uobičajeno koristio za testiranje pretpostavke da dve subpopulacije (npr. muškarci i žene) imaju "prosečno gledano" jednaku prostornu sposobnost (H_0: $\mu_1 = \mu_2$) izvedena je pod sledećim pretpostavkama:

a) varijabla X (prostorna sposobnost u našem slučaju) ima normalnu funkciju gustine u svakoj od subpopulacija sa parametrima μ i σ^2 (subpopulacija muškaraca: $X \sim N(\mu,\sigma^2)$; subpopulacija žena: $X \sim N(\mu,\sigma^2)$).

b) varijabla X ima jednake varijanse u dvema subpopulacijama: $\sigma^2_1 = \sigma^2_2 = \sigma^2$;

c) subpopulacije imaju jednake aritmetičke sredine na varijabli X: ($\mu_1 = \mu_2 = \mu$);

d) opservacije su nezavisne, a uzorci jedinica posmatranja jednostavni slučajni uzorci;

Statistik za testiranje nulte hipoteze u ovom slučaju, dobro poznati t-statistik, definiše se na sledeći način:

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{SE_{M_1 - M_2}}$$

U kojoj meri će disparatnost skupa pretpostavki koje su u osnovi određenog statističkog postupka i ispitivane realnosti imati (negativni) efekat zavisi, pre svega, od toga kako određena statistička metoda funkcioniše u različitim "nemodelskim", "neidealnim", tj. realnim uslovima. Osetljivost, tj. nerobustnost statističkih postupaka na aproksimativnost u procesu njihove primene, tj. na veliki broj malih ili mali broj velikih odstupanja od pretpostavki na kojima se ovi postupci zasnivaju predstavlja, po mišljenju pisca ovog teksta, jedno od najvažnijih svojstava statističkih postupaka. Poznavanje tog svojstva statističkih postupaka koji se primenjuju u analizama realnih podataka od kritičnog je značaja jer od toga zavisi valjanost naučnih zaključaka do kojih se dolazi njihovim korišćenjem.

Sam termin *robustnost* (i to pod navodnicima) prvi je upotrebio statističar Boks 1953. godine u članku koji je upravo bio posvećen ispitivanju osetljivosti statističkih testova za poređenje varijansi dveju populacija na neispunjenost pretpostavke o normalnosti distribucija u populaciji (Box, 1953). Ovim terminom Boks je označio relativnu neosetljivost statističkih testova za testiranje značajnosti razlika između aritmetičkih sredina na određeni tip nenormalnosti distribucija u populaciji. Vremenom je upotreba termina robustnost u navedenom smislu proširena tako da obuhvati stepen osetljivosti statističkih testova za testiranje hipoteza na neispunjenost bilo koje pretpostavke koja čini sastavni deo matematičkog modela koji je u osnovi izvođenja samog testa. Dakle, statistički postupak ili test koji funkcioniše "razumno dobro" u uslovima blage ili umerene narušenosti pretpostavki probabilističkog modela na kojima se zasniva njegovo matematičko izvođenje smatra se robustnim statističkim postupkom. "Razumno dobro" funkcionisanje robustnog statističkog testa podrazumeva da distribucija uzorkovanja statistika za testiranje nulte hipoteze ne odstupa bitno u uslovima blage (i umerene) narušenosti pretpostavki probabilističkog modela, kao i to da snaga statističkog testa takođe nije bitno narušena. Drugim rečima, verovatnoće grešaka tipa I i II pri testiranju nulte hipoteze

---

U obrascu za t statistik M_1-M_2 je razlika između aritmetičkih sredina uzoraka, a $SE_{M_1-M_2} = \sqrt{S^2(\frac{1}{n_1}+\frac{1}{n_2})}$ je ocena standardne greške za razliku između aritmetičkih sredina. Pri tom se ocena varijanse za kvantitativnu varijablu u populaciji, u oznaci S^2, dobija na osnovu varijansi uzoraka na sledeći način:

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Ukoliko su uslovi pod a, b, c i d iz probabilističkog modela ovog postupka ispunjeni tada, kao što je poznato, statistik t kao slučajna varijabla ima Studentovu distribuciju uzorkovanja koja je definisana parametrom ν, tj. stepenima slobode (pri čemu je $\nu = n_1 + n_2 - 2$, a n_1 i n_2 predstavljaju veličine uzoraka). *Ukoliko su*, dakle, *uslovi uključeni u probabilistički model koji stoji u osnovi* prikazanog postupka t-testa *dopustivo dobra aproksimacija* realnosti ovaj postupak se može primeniti kao valjan postupak za testiranje nulte hipoteze o jednakosti aritmetičkih sredina muškog i ženskog dela ljudske populacije u pogledu varijable X, tj. prostorne sposobnosti u našem slučaju. U tom će slučaju verovatnoća da t statistik uzme na slučajnim uzorcima neku od svojih vrednosti u određenom intervalu – a koja se dobija na osnovu njegove nulte distribucije uzorkovanja – biti adekvatna. Od tačnosti ocene ove verovatnoće zavisiće i da li će odluka o odbacivanju ili neodbacivanju nulte hipoteze, pa i krajnji supstantivni, tj. naučni zaključak o ispitivanom fenomenu biti utemeljeni na tačnim ili pogrešnim brojkama! Naravno, ovde se prirodno nameću barem dva pitanja: 1. Kakve su posledice neispunjenosti uslova probabilističkog modela, tj. da li će zaključci do kojih ćemo doći primenom statističkog postupka – ako pretpostavke na kojima se on zasniva ne aproksimiraju dopustivo dobro realnost – biti nužno pogrešni? 2. Da li su sve pretpostavke statističkog modela podjednako važne, tj. da li se neispunjenost nekih od uslova može zanemariti bez bojazni po ispravnost donetih zaključaka?

korišćenjem robustnog statističkog testa u uslovima narušavanja pretpostavki probabilističkog modela na kojem se test zasniva veoma su blizu nominalnim vrednostima koje slede na osnovu probabilističkog modela.[3] Međutim, u odsustvu preciznog kvantitativnog kriterijuma za dozvoljeno odstupanje verovatnoća dveju vrsta grešaka u statističkom zaključivanju pri narušenosti pretpostavljenih uslova probabilističkog modela nije moguća ni objektivna ocena robustnosti statističkog testa, a zaključci o (ne)robustnosti statističkog testa mogu biti veoma subjektivni i varijabilni. Postoje pokušaji da se ovaj problem reši definisanjem granica u kojima se – u uslovima narušenosti pretpostavki na kojima test počiva – može kretati aktuelna verovatnoća greške I tipa za dati nominalni nivo značajnosti (uobičajeno u oznaci α) pri primeni statističkog testa. Bredli je 1978. godine predložio dva kriterijuma koji se razlikuju po stepenu strogosti (Bradley, 1978): prema strožem kriterijumu statistički test je robustan ako je – pri narušenosti pretpostavki na kojima test počiva – stvarna verovatnoća greške pri odbacivanju tačne nulte hipoteze u rasponu $\alpha \pm 0.1\alpha$, pri čemu je α nominalni nivo značajnosti; prema blažem kriterijumu, koji Bredli ujedno smatra najliberalnijim kriterijumom koji se "može smatrati ozbiljnim", test se može smatrati robustnim ukoliko se stvarna verovatnoća greške I tipa u uslovima narušenih pretpostavki testa kreće u granicama od $\alpha - 0.5\alpha$ do $\alpha + 0.5\alpha$.

Ispitivanja robustnosti statističkih testova izvode se primenom simulacionih eksperimenata u kojima se porede distribucije uzorkovanja statistika za testiranje hipoteza pri uzorkovanju iz "idealnih" populacija (pretpostavljenih statističkim modelom postupka) i distribucije uzorkovanja za isti statistik kada se uzorkovanje vrši iz populacija koje nisu sasvim u skladu sa pretpostavljenim probabilističkim modelom statističkog testa.

Robustnost parametara

Robustnost određenog parametra θ može se definisati na sledeći način: parametar $\theta(F)$ (pri čemu je F funkcija distribucije, tj. kumulativna distribucija verovatnoća) robustan je ukoliko brzina njegove promene, pri proizvoljno malim promenama funkcije distribucije F, ne može biti proizvoljno velika. Drugim rečima, parametar $\theta(F)$ – pri čemu je F funkcija distribucije – predstavlja robustan parametar ako male perturbacije u F ne dovode do velikih promena u $\theta(F)$. Dakle, male promene u F ne bi trebalo da odvedu $\theta(F)$ ka proizvoljno velikim vrednostima. Minimalni zahtev koji se postavlja nekom parametru da bi se on mogao smatrati robustnim jeste da *tačka otkaza* parametra bude *veća od nule* i da njegova *funkcija uticaja* bude *ograničena*. Tačka otkaza ili granica otkaza (engl. breakdown point; breakdown bound) nekog parametra $\theta(F)$ za distribuciju F predstavlja najmanju vrednost ε za koju parametar $\theta(F_\varepsilon)$ može postići proizvoljno veliku vrednost. Vrednost ε predstavlja meru distance ili razlike distribucije F i distribucije F_ε. Za aritmetičku sredinu, varijansu i koeficijent linearne korelacije tačka otkaza jednaka je nuli, tj. najmanjoj mogućoj vrednosti za tačku otkaza, pa se ovi parametri prema njihovoj tački otkaza ne mogu smatrati robustnima. Drugim rečima, za infinitezimalnu distancu ε između dveju distribucija, tj. njihovih funkcija F i F_ε, ovi parametri mogu

[3] Greška tipa I pri testiranju statističkih hipoteza odnosi se na odbacivanje tačne nulte hipoteze, a greška tipa II na neodbacivanje pogrešne nulte hipoteze. Verovatnoća greške tipa I uobičajeno se naziva nivoom značajnosti i označava se sa α, a verovatnoća greške tipa II označava se sa β.

uzeti vrednosti koje se razlikuju za proizvoljno veliki broj. Funkcija uticaja (engl. influence function), pak, predstavlja graničnu vrednost količnika:

$$\frac{\theta(F_\varepsilon) - \theta(F)}{\varepsilon}$$

sa približavanjem ε ka nuli zdesna. Tako je, na primer, funkcija uticaja $IF_{\mu,F}(x)$ za aritmetičku sredinu kao parametar ($\mu = \int x dF(x)$) jednaka $x - \mu$ ($x \in R$) pa – budući da je njena funkcija uticaja neograničena po x – aritmetička sredina prema ovom pokazatelju ne spada u robustne parametre. Isto tako, varijansa distribucije F ($\sigma^2 = \int (x - \mu)^2 dF$), ako varijansa postoji i za poznatu aritmetičku sredinu distribucije μ, ima funkciju uticaja oblika $(x - \mu)^2 - \sigma^2$ ($x \in R$) pa, očigledno, zbog neograničenosti ove funkcije uticaja ni varijansa nije robustan parametar. U vezi s tim, postavlja se, dakle, pitanje opravdanosti korišćenja ovih parametara za opis tipične osobe ili varijabilnosti u populaciji za varijable za koje pretpostavka o normalnosti ne izgleda održiva. Slični problemi postoje i u vezi sa korišćenjem koeficijenta linearne korelacije kao parametra u populacijama za koje pretpostavka o bivarijacionoj normalnoj distribuciji ne izgleda opravdana.

Moguće je postaviti pitanje da li je izbor parametara koji preovlađuje u psihološkim istraživanjima – sa stanovišta robustnosti samih parametara – "najsrećniji" izbor. Mogli bismo reći da je sam izbor "nerobustnih" parametara kojima se najčešće bave istraživači u psihologiji pod uticajem teorijskih pretpostavki o "sveprisutnoj" normalnosti raspodela psiholoških varijabli u populaciji, tj. "mita o normalnosti" koji opstaje uprkos tome što distribucije psiholoških varijabli izgleda skoro nikada nisu normalne! U prilog tome govore i rezultati istraživanja Mićerija iz 1989. godine u kojem su ispitivana svojstva 440 empirijskih distribucija sa 46 psiholoških testova (uglavnom testova sposobnosti, postignuća i ličnosti) koje su dobijene na velikim uzorcima (Micceri, 1989): ne samo da su *sve* ispitivane distribucije bile nesaglasne sa pretpostavkom o normalnosti u populacijama iz kojih su uzorci dobijeni, već je samo 3% realnih distribucija podataka koje su analizirane u ovom istraživanju bilo simetrično! Dakle, rezultati Mićerijevog istraživanja *ne* sugerišu da, barem kada je o psihološkim varijablama reč, "Bog voli normalnu krivu". Blanca i saradnici (Blanca et al., 2013) analizirali su u pogledu asimetrije i izduženosti 693 distribucije realnih podataka iz psiholoških istraživanja na španskim univerzitetima. Podaci su poticali iz uzoraka veličine od 10 do 30 iz 130 različitih populacija i odnosili su se uglavnom na mere kognitivnih sposobnosti, crta ličnosti i kliničkih relevantnih crta (na primer, anksioznosti, depresivnosti). Rezultati pokazuju da je samo 5.5% distribucija imalo vrednosti skjunisa i kurtozisa vrlo bliske onima koji se očekuju pod pretpostavkom da je raspodela mera u populaciji normalna (apsolutne vrednosti skjunisa i kurtozisa manje od 0.25) . Od preostalih distribucija, 74.4% njih je pokazivalo blaga ili umerena odstupanja od normalnosti (apsolutne vrednosti skjunisa i/ili kurtozisa između 0.26 i 1.75), a 20% raspodela izrazito su odstupale od normalne (apsolutne vrednosti skjunisa i/li kurtozisa iznad 1.75). Među potonjim distribucijama 7% njih pokazivalo je ekstremna odstupanja od normalnosti (apsolutne vrednosti skjunisa i/li kurtozisa veće od 2.25). Kao što se moglo i očekivati, odstupanja od normalnosti su izrazitija za distribucije crta ličnosti i klinički relevantnih crta nego za distribucije kognitivnih sposobnosti. Bez obzira na to što autori u ovom istraživanju nisu testirali statističku značajnost odstupanja ovih distribucija od normalne raspodele, rezultati podgrejavaju sumnju u pretpostavku o normalnosti distribucija psiholoških varijabli u populaciji.

Robustnost ocenitelja parametara

Pored testova statističkih hipoteza statistički postupci uključuju i mnoge statistike ocenitelje koji treba da posluže kao kvalitetna ocena populacijskih numeričkih svojstava, tj. parametara. Uzoračka stabilnost (prosečno kvadrirano odstupanje, odnosno varijansa) ovih ocenitelja može biti pod velikim uticajem distribucionih svojstava varijabli kojima se istraživači bave ali i idiosinkratičnih svojstava uzorka na kojem se primenjuju. Tako, aritmetička sredina uzorka predstavlja najprecizniju ocenu parametra, tj. aritmetičke sredine populacije (predstavlja tzv. nepristrasni ocenitelj sa minimalnom varijansom) ukoliko se uzorkovanje vrši iz populacije sa *normalnom* funkcijom gustine. Ukoliko se, pak, uzorkovanje vrši iz distribucije koja odstupa od normalne funkcije gustine tada aritmetička sredina uzorka kao ocenitelj parametra gubi ova svojstva. Na primer, ako se uzorkovanje vrši iz "mešane" normalne raspodele – raspodele koja predstavlja konvoluciju ('mešavinu') više normalnih raspodela i pokazuje, prema tome, malo odstupanje od normalnosti – medijana uzorka kao ocenitelj parametra ima znatno manju varijansu od uzoračke aritmetičke sredine. Pored toga, dobro je poznato da prisustvo malog broja nesaglasnih vrednosti – iznimaka ili autlajera (engl. outlier) u uzorku opservacija koje služe za računanje aritmetičke sredine može da dovede do potpunog otkaza (engl. break down) ovog ocenitelja, tj. do njegove neupotrebljivosti.

Postupak ocenjivanja ili ocenitelj (engl. estimator) parametra smatra se robustnim ukoliko nije osetljiv na "mala odstupanja" od idealizovanih pretpostavki za koje je optimizovan. Pri tome se pod malim odstupanjima podrazumeva mali broj velikih ili veliki broj malih odstupanja. Robustnost ocenitelja obuhvata dva svojstva: *robustnost u efikasnosti* i *rezistentnost*.

Robustnost u efikasnosti datog ocenitelja podrazumeva da je varijansa nepristrasnog ocenitelja ili prosečna kvadrirana greška za pristrasni ocenitelj blizu minimalne vrednosti za različite distribucije.[4] Dakle, da bi ocenitelj bio robustan u efikasnosti on treba da bude visoko efikasan (tj. da ima malu standardnu grešku ili prosečnu kvadratnu grešku) za različite distribucije koje mogu predstavljati plauzibilne modele empirijskih raspodela. Ova osobina je praktično veoma važna jer ona govori o tome koliko će vrednosti ocenitelja dobijene na uzorcima biti blizu vrednosti parametra koji se ocenjuje. Drugim rečima, što je ocenitelj efikasniji to će ocene parametra koje se dobijaju na uzorku biti u principu bliže pravoj vrednosti parametra.

Sa stanovišta analize realnih podataka ovo svojstvo ocenitelja obezbeđuje njegovo dobro funkcionisanje u uslovima kada uzorci potiču iz populacija čije distribucije imaju "gušće" krajeve u odnosu na normalnu raspodelu ili, pak, iz populacija čija distribuciona svojstva nije moguće precizno definisati.

Svojstvo *rezistentnosti* ocenitelja definiše se preko tačke ili granice otkaza za konačni uzorak (engl. finite sample breakdown point ili finite sample breakdown bound), tj. preko najveće proporcije opservacija koje se mogu neograničeno menjati a da promena u vrednosti ocenitelja bude ograničena. Tačka otkaza za konačni uzorak ocenitelja T_n definisana je formalno na sledeći način:

$$\varepsilon_n^*(T_n, Z_n) = \min\left\{\frac{k}{n}; \sup_{Z_n'} \left\| T_n(Z_n') - T_n(Z_n) \right\| = \infty\right\},$$

[4] Definicije pristrasnosti, varijanse i prosečne (srednje) kvadratne greške ocenitelja može se pročitati u Tenjović (2020, glava VI.1.).

pri čemu Z_n' obuhvata raspone podataka na svim skupovima koji se dobijaju zamenom k vrednosti u Z_n arbitrarnim vrednostima. Prema tome, tačka otkaza je najmanja proporcija neograničeno izmenjenih podataka koja može odvesti ocenitelj izvan svih granica, tj. dovesti do toga da se vrednost ocenitelja drastično promeni, da postane proizvoljno velika ili mala. Aritmetička sredina, varijansa i koeficijent linearne korelacije kao ocenitelji imaju najmanju moguću vrednost tačke otkaza za konačni uzorak, tj. $\frac{1}{n}$, pri čemu je n veličina uzorka. Očigledno, samo jedan neograničeno izmenjen podatak može da dovede do otkazivanja ovih ocenitelja, tj. može tako dobijene ocene učiniti neupotrebljivim. Granična vrednost tačke otkaza za konačni uzorak za neki ocenitelj tipično je jednaka tački otkaza parametra koji se ocenjuje. Budući da je $\lim_{n\to\infty}\frac{1}{n}=0$, tačka otkaza parametara koji se ocenjuju navedenim oceniteljima jednaka je nuli, tj. najnižoj mogućoj vrednosti. Prema tome, ocenitelj je utoliko robustniji ukoliko je njegova tačka otkaza veća od $\frac{1}{n}$. Ako se posmatra iz perspektive analize realnih podataka, robustnost ocenitelja, definisana teorijski preko tačke otkaza, govori o *rezistentnosti* ocenitelja *na autlajere* (engl. outlier), tj. iznimke. U osnovi rezistentnost ocenitelja u analizama podataka znači da vrednost koju ocenitelj uzima na nekom uzorku neće biti pod velikim uticajem malog broja rezultata koji su nesaglasni sa glavninom podataka, tj. *autlajera* ili *iznimaka*.

Autlajeri ili iznimci (engl. outliers)

Autlajeri ili iznimci su one vrednosti u skupu podataka koje su neuobičajeno daleko ili koje su veoma različite od glavnine podataka. Prisustvo autlajera u uzorku podataka najčešće se tumači kao posledica grešaka merenja, grešaka u unosu podataka ili (što je sa statističkog aspekta najvažnije) kao odraz intrinzične varijabilnosti izvora podataka. Određenje iznimaka kao odraza intrinzične varijabilnosti podrazumeva da postoji statistički model o izvoru koji generiše podatke, tj. o prirodi distribucije varijable u populaciji iz koje je generisan uzorak. Autlajer bi se u tom slučaju mogao posmatrati kao opservacija koja je uzorkovana iz druge populacije u odnosu na onu iz koje je glavnina podataka. Moguće je, isto tako, postojanje iznimaka u podacima tumačiti i u smislu uzorkovanja iz populacije čija distribucija ima "gušće krajeve" od pretpostavljene raspodele. Autlajeri u bivarijacionom slučaju mogu predstavljati opservacije koje odstupaju od glavnine podataka po veličini ("autlajeri pomerenosti" – engl. shift outliers) ili po strukturi. Dok prva vrsta autlajera prati osnovnu strukturu povezanosti među varijablama i predstavlja prosto daleke opservacije koje prate trend glavnine podataka, dotle druga vrsta iznimaka upravo narušava strukturu povezanosti između varijabli. Ukoliko su, na primer, dve osobine ličnosti u pozitivnoj linearnoj vezi, autlajer "po veličini" predstavljao bi ispitanika koji ima neuobičajeno visoke (ili niske) rezultate na obema crtama, dok bi autlajer "po strukturi" bila jedinica posmatranja sa neobično visokim rezultatom na jednoj, a neuobičajeno niskim rezultatom na drugoj osobini.

Otkrivanje autlajera u podacima neobično je važno. Pre svega, na taj način moguće je otkriti grube greške u merenju ili unosu podataka koje bi, ukoliko ostanu neuočene, mogle potpuno kompromitovati zaključke istraživanja. S druge strane, pravi iznimci, tj. rezultati izrazito netipičnih jedinica posmatranja mogu – budući da se u analizama podataka psiholoških istraživanja još uvek skoro isključivo koriste

statistički postupci koji su nepostojani, tj. nerezistentni na autlajere – dovesti do nesrazmernog izobličenja rezultata i tako dovesti do neadekvatnih ocena ključnih parametara koji su od interesa za istraživača. Zavisno od toga koji matematički model distribucije, po pretpostavci, dobro opisuje raspodelu jedne ili više varijabli u populaciji postoje različiti statistički kriterijumi i raznovrsni grafički postupci za detekciju autlajera. Najčešće primenjivani klasični statistički kriterijumi za iznimke u jednodimenzionalnom slučaju jesu postojanje podataka čije odstupanje od aritmetičke sredine je veće od 2.5 ili 3 standardne devijacije (u zavisnosti od veličine uzorka) ili čija je udaljenost od gornje (ili donje) četvrti (engl. uper and lower forth, vrednosti koje grubo odgovaraju percentilima 75 i 25) veća od izraza koji se dobija množenjem međučetvrtnog raspršenja (razlike gornje i donje četvrti) vrednošću 1.5. U novije vreme sve se više koristi "robustni kriterijum" za deklarisanje jednodimenzionalnih autlajera koji se definiše na sledeći način: rezultat x_i je iznimak ako je

$$\boxed{\frac{|x_i - \mathrm{Mdn}|}{\mathrm{MAD}/0.6745} > 2.24}$$

Pri tome, Mdn je medijana, MAD je medijana apsolutnih odstupanja svih rezultata od medijane (engl. Median Absolute Deviation), a konstanta 0.6745 služi za reskaliranje mere MAD tako da u slučaju normalne raspodele MAD može poslužiti kao ocena standardne devijacije populacije.

Uočavanje potencijalnih iznimaka u univarijacionom i bivarijacionom slučaju relativno je jednostavno i za to je ponekad dovoljno pažljivo i znalačko posmatranje valjanog grafičkog prikaza podataka (npr. kutijastog dijagrama – engl. boxplot, dijagrama raspršenja i slično).

***

Precizne definicije robustnosti na osnovu funkcije uticaja i tačke otkaza mogu se koristiti u formalnom dokazivanju robustnosti ili nerobustnosti parametra i ocenitelja. Utvrđivanje robustnosti ili nerobustnosti na autlajere za ocenitelje korišćenjem empirijskih studija podrazumeva odabir ili definiciju određenih pokazatelja robustnosti na autlajere. Stepen u kojem je ocenitelj rezistentan na autlajere se potom procenjuje uzimanjem u obzir empirijski dobijenih vrednosti ovih pokazatelja.

2. Robustni ocenitelji lokacije

Statističke mere koje ukazuju na određeno mesto (lokaciju) u raspodeli ili sažeto prikazuju distribuciju neke varijable predstavljaju mere lokacije. Na primer, aritmetička sredina, medijana i percentili spadaju u mere lokacije. U opštem slučaju, ako je X kvantitativna varijabla, statistička mera T(X) može predstavljati meru lokacije ukoliko ispunjava sledeće formalne uslove (prema Wilcox, 2005, str. 20):

a) $T(X + a) = T(X) + a$ (lokaciona ekvivarijantnost);
b) $T(-X) = -T(X)$;
c) $(X \geq 0) \Rightarrow T(X) \geq 0$;
d) $T(b * X) = b * T(X)$ (skalna ekvivarijantnost);

Dakle:

a) dodavanje konstante svakoj vrednosti na varijabli X menja statističku meru lokacije za tu konstantu;

b) množenje svake vrednosti na varijabli X sa -1 dovodi do promene predznaka mere lokacije;
c) ako su sve vrednosti na varijabli X jednake nuli ili veće od nule, tj. nenegativne onda je i mera lokacije nenegativna;
d) množenje konstantom svake vrednosti na varijabli X menja statističku meru lokacije za tu konstantu puta.

Prva tri uslova imaju kao posledicu to da mera lokacije za varijablu X može uzeti samo vrednosti koje su unutar raspona vrednosti na varijabli X.

Mere centralne tendencije, koje ukazuju na vrednost kojoj „teže“ rezultati, tj. na mesto oko kojeg se grupišu ili nagomilavaju vrednosti na varijabli predstavljaju podskup mera lokacije.

U poslednjih 50 godina definisan je veliki broj robustnih ocenitelja lokacije koji imaju višu tačku otkaza od aritmetičke sredine i, prema tome, manju osetljivost na autlajere i koji su, istovremeno, u većoj ili manjoj meri robustni u efikasnosti. U ovom tekstu, samo kao ilustraciju, definisaćemo tri jednostavna robustna ocenitelja lokacije: *postriženu aritmetičku sredinu*, *vinzorizovanu aritmetičku sredinu* i *modifikovani M-ocenitelj iz jednog koraka*.

Postrižena aritmetička sredina (engl. Trimmed mean)

Postrižena aritmetička sredina, u oznaci $M_{t(\gamma)}$, definisana je na sledeći način:

$$M_{t(\gamma)} = \frac{\sum_{i=g+1}^{n-g} x_{(i)}}{n-2g}$$

Pri tome, $x_{(i)}$ je redosledni statistik (rezultat koji je na *i*-tom mestu po veličini kada se rezultati poređaju od najmanjeg ka najvećem), n je ukupan broj rezultata, γ je proporcija najnižih i najviših rezultata koji se odbacuju, a g je jednako [γ*n]. Oznaka [ ] znači da, ukoliko je proizvod γ*n decimalan broj, g predstavlja samo njegov celobrojni deo. Na primer, ako je γ*n jednako 9.8 tada je g = 9. Dakle, postrižena aritmetička sredina predstavlja aritmetičku sredinu rezultata koji preostanu kada se g najnižih i g najviših rezultata odbace, tj. „postrižu“. Najčešće se, pri računanju postrižene aritmetičke sredine uzima da je γ jednako 0.20, tj. odbacuje se 20% najnižih i 20% najviših rezultata. Naime, simulaciona statistička istraživanja pokazala su da postrižena aritmetička sredina sa proporcijom strige 0.20 ima najbolja svojstva u pogledu robustnosti.

Vinzorizovana aritmetička sredina (engl. Winsorized mean)

Vinzorizovana aritmetička sredina, u oznaci M_w, definisana je na sledeći način:

$$M_w = \frac{\sum_{i=1}^{n} w_i}{n}$$

Pri tome, w_i je vinzorizovani rezultat za *i*-tu jedinicu posmatranja, a n je ukupan broj rezultata, tj. veličina uzorka. Vinzorizovani rezultati definišu se na sledeći način:[5]

$$w_i = \begin{cases} x_{(g+1)}, & \text{ako je } x_i . \le x_{(g+1)} \\ x_i, & \text{ako je } x_{(g+1)} < x_i . < x_{(n-g)} \\ x_{(n-g)}, & \text{ako je } x_i \ge x_{(n-g)} \end{cases}$$

Dakle, kada se početni rezultati pretvore u redosledne statistike, tj. poređaju po veličini od najmanjeg do najvećeg, potrebno je na osnovu proporcije vinzorizacije, proporcije γ, odrediti g, tj. broj najnižih i najvećih rezultata koji se vinzorizuju, i to prema sledećoj formuli: g = [γ*n]. Oznaka [] znači da, ukoliko je proizvod γ*n decimalan broj, g predstavlja samo njegov celobrojni deo. Na primer, ako je γ = 0.2, a n jednako 18 tada je g = 3. Pošto se odredi g, tada se svi rezultati koji su manji od rezultata koji je na mestu g+1 u rastućem redosledu zamene rezultatom koji je na mestu g+1, a rezultati koji su veći od rezultata koji je po veličini na mestu n-g zamene rezultatom koji je na mestu n-g. Drugim rečima, rezultati manji od redoslednog statistika $x_{(g+1)}$ zamenjuju se redoslednim statistikom $x_{(g+1)}$, a rezultati veći od redoslednog statistika $x_{(n-g)}$ zamenjuju se redooslednim statistikom $x_{(n-g)}$. Svi ostali rezultati ostaju nepromenjeni.

Primer: ako je proporcija vinzorizacije γ = 0.2 a početni rezultati poređani po veličini 1 6 7 8 9 10 12 32, tada je g = [γ*n] = [0.2*8] = 1. Dakle, rezultatom koji je na mestu g + 1, tj. na drugom mestu (u ovom slučaju je to 6) zamenjujemo sve rezultate koji su manji od 6, a rezultatom koji je na mestu n-g, tj. na sedmom mestu (u ovom slučaju je to 12) zamenjujemo sve rezultate koji su veći od 12. Niz vinzorizovanih rezultata bi tada izgledao ovako: 6 6 7 8 9 10 12 12. Vinzorizovanu aritmetičku sredina niza 1 6 7 8 9 10 12 32 bismo, dakle, izračunali kao običnu aritmetičku sredinu niza 6 6 7 8 9 10 12 12.

M-ocenitelj iz jednog koraka i modifikovani M-ocenitelj lokacije iz jednog koraka

M-ocenitelj iz jednog koraka, u oznaci OM (engl. **O**ne-step **M**-estimator), definisan je na sledeći način:

$$OM = \frac{1.28(MAD/0.6745)(i_2 - i_1)\sum_{i=i_1+1}^{n-i_2} x_{(i)}}{n - i_1 - i_2}$$

[5] Ovu operaciju na podacima nazvao je vinzorizacijom Džon Tjuki (John Tukey), poznati američki matematičar i statističar, začetnik robustne statistike, u čast svog preminulog prijatelja Čarlsa Pejna Vinzora (Charles Paine Winsor), inače inženjera i biostatističara, koji je i osmislio sam postupak vinzorizacije (Tukey, 1962).

Pri tome, $x_{(i)}$ je redosledni statistik, i_1 je broj rezultata za koje važi: $\frac{x_i - \text{Mdn}}{\text{MAD}/0.6745} < -1.28$, i_2 je broj rezultata za koje važi: $\frac{x_i - \text{Mdn}}{\text{MAD}/0.6745} > 1.28$. Pri tome Mdn je medijana, a MAD predstavlja medijansko apsolutno odstupanje, tj. medijanu apsolutnih odstupanja svih rezultata od medijane. Konstante 1.28 i 0.6745 u obrascu za OM predstavljaju vrednosti kvantila 0.90 i kvantila 0.75 iz normalne raspodele, tim redom. Vrednost 1.28 koja figurira u obrascu izabrana je kako bi se postiglo da OM ocenitelj ima standardnu grešku sličnu aritmetičkoj sredini ako je distribucija normalna a da standardna greška ocenitelja ne bude jako velika kada distribucija varijable ima „gušće" krajeve od normalne raspodele. Naime, standardna greška aritmetičke sredine dramatično raste kada distribucija varijable u populaciji ima „gušće" krajeve od normalne raspodele. U tom slučaju, dakle, aritmetička sredina uzorka kao ocenitelj parametra ima veliku standardnu grešku i nije efikasna ocena parametra.

Modifikovani M-ocenitelj iz jednog koraka, u oznaci MOM (engl. **M**odified **O**ne-step **M**-estimator), definisan je na sledeći način:

$$\text{MOM} = \frac{\sum_{i=i_1+1}^{n-i_2} x_{(i)}}{n - i_1 - i_2}$$

Dakle, MOM je, u stvari, aritmetička sredina rezultata, koji prema ranije definisanom robustnom kriterijumu ne predstavljaju autlajere.

3. Robustni ocenitelji skale

U opštem slučaju, ako je X kvantitativna varijabla, statistička mera $\tau(X)$ može predstavljati meru skale ukoliko ispunjava sledeće formalne uslove (modifikovano na osnovu Bickel & Lehmann, 1976):

1. Mera skale $\tau(X)$, pri čemu je X kvantitativna varijabla sa distribucijom *F,* mora biti nenegativna. Drugim rečima, mera skale može imati vrednosti jednake nuli ili veće od nule.

2. $\tau(bX + a) = |b|\,\tau(X)$, za $b \neq 0$ i za svako a.

Ovaj se uslov sadrži u sebi dva uslova: skalnu ekvivarijantnost i lokacionu invarijantnost.
Skalna ekvivarijantnost znači da množenje svake vrednosti na varijabli X nenultom konstantom *b* treba da dovede do promene mere skale za apolutnu vrednost multiplikativne konstante puta. S druge strane, lokaciona invarijantnost znači da dodavanje konstante *a* na svaku vrednost varijable X ne bi trebalo da menja meru skale. Odavde sledi da je
$\tau(X + a) = \tau(X)$, $\tau(-X) = \tau(X)$ i $\tau(c) = 0$, pri čemu je *c* konstanta.

Najpoznatija i najčešće korišćena robustna mera skale je medijansko apsolutno odstupanje.

Medijansko apsolutno odstupanje (engl. Median Apsolute Deviation)

Medijansko apsolutno odstupanje, u oznaci MAD (akronim engleskog naziva ocenitelja), definisano je na sledeći način:

$$MAD = Mdn\ \{\ |x_1 - Mdn|\ \ldots\ |x_n - Mdn|\ \}$$

Dakle, MAD predstavlja medijanu apsolutnih odstupanja svih rezultata od medijane tih rezultata. MAD se često reskalira deljenjem sa 0.6745, tj. kvantilom 0.75 iz standardizovane normalne raspodele kako bi, kada je raspodela varijable normalna, mogao da posluži kao ocenitelj standardne devijacije populacije. Reskalirani oblik medijanskog apsolutnog odstupanja označava se uobičajeno sa MADN:

$$MADN = \frac{MAD}{0.6745}$$

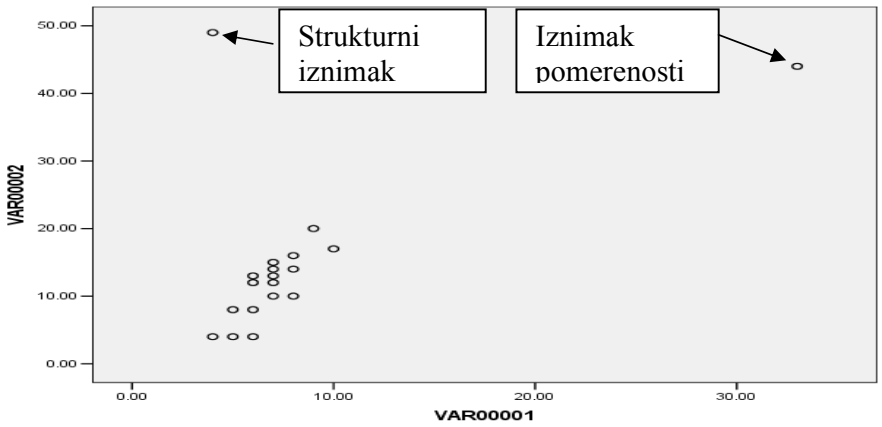
4. Robustni ocenitelji povezanosti dveju kvantitativnih varijabli

Iz statističke teorije poznato je da je koeficijent linearne korelacije adekvatan parametar povezanosti dveju varijabli u populaciji onda kada je zajednička funkcija gustine dveju varijabli bivarijaciona normalna raspodela. Ako želimo da ocenimo taj parametar, tj. linearnu povezanost dveju kvantitativnih varijabli u populaciji tada je Brave-Pirsonov koeficijent korelacije dobijen na uzorku po mnogim kriterijumima najbolji ocenitelj parametra. Međutim, kako smo već istakli, koeficijent linearne korelacije u populaciji ne predstavlja robustan parametar a koeficijent linearne korelacije uzorka nije robustan ocenitelj. To praktično znači da ukoliko zajednička raspodela dveju varijabli u populaciji nije bivarijaciona normalna raspodela postavlja se pitanje opravdanosti razmišljanja o povezanosti dveju varijabli u populaciji u terminima koeficijenta linearne korelacije. S druge strane, koeficijent linearne korelacije uzorka predstavlja nerezistentan ocenitelj, tj. ocenitelj osetljiv na prisutvo iznimaka u podacima. Budući da pri analizi povezanosti dveju kvantitativnih varijabli imamo posla sa tzv. bivarijacionom distribucijom postoji mogućnost postojanja dve vrste iznimaka, tj. autlajera:

1. iznimci „pomerenosti“
2. strukturni iznimci .

Iznimci pomerenosti znatno odstupaju od glavnine podataka ali prate osnovni trend podataka dok strukturni iznimci ne samo da su znatno udaljeni od glavnine podataka već ne slede ni osnovni trend podataka. Iznimci pomerenosti predstavljaju zapravo ekstremne rezultate u marginalnim (univarijacionim) distribucijama obeju varijabli. Strukturni iznimci ne moraju biti ekstremni rezultati kada posmatramo univarijacione distribucije varijabli: ono što definiše njihovu iznimnost je što narušavaju osnovni trend, struktru podataka. Na dijagramu raspršenja na Slici 1 prikazan je po jedan primer ovih dveju vrsta iznimaka. Obe vrste iznimaka mogu bitno da promene vrednost koeficijenta linearne korelacije, svaka na drugačiji način: iznimci pomerenosti, budući da prate glavni trend podataka dovode do povećanja koeficijenta, a strukturni iznimci, budući da narušavaju osnovni trend podataka dovode do smanjenja vrednosti koeficijenta.

Slika 1. Dijagram raspršenja sa iznimkom pomerenosti i strukturnim iznimkom



U skladu sa ovim dvema vrstama iznimaka, robustne ocenitelje parametara povezanosti dveju kvantitativnih varijabli možemo podeliti u dve grupe:

1. ocenitelji M-tipa;
2. ocenitelji O-tipa.

Ocenitelji korelacije M-tipa su rezistentni na autlajere (iznimke) "pomerenosti", tj. na iznimke u pojedinim marginalnim distribucijama dveju varijabli, a ocenitelji korelacije O-tipa su robustni na strukturne autlajere (iznimke).

Robustni ocenitelji M-tipa za povezanost dveju kvantitativnih varijabli

Vinzorizovani koeficijent korelacije

Vinzorizovani koeficijent korelacije, u oznaci r_W, definisan je na sledeći način:

$$r_W = \frac{\sum_{i=1}^{n}(w_{xi} - M_{wx})(w_{xi} - M_{wx})}{nS_{wx}S_{wy}}$$

Pri tome, n je veličina uzorka, w_{xi} i w_{yi} su vinzorizovani podaci na varijablama X i Y, M_{wx} i M_{wy} aritmetičke sredine a S_{wx} i S_{wy} standardne devijacije vinzorizovanih podataka na dvema varijablama.

Dakle, ovaj koeficijent korelacije dobija se računanjem Brave-Pirsonovog koeficijenta korelacije na vinzorizovanim podacima, pri čemu se vinzorizacija podataka izvodi posebno za svaku varijablu.

PB koeficijent korelacije

PB koeficijent korelacije (engl. Percentage bend correlation) se zasniva na odstupanjima rezultata na svakoj varijabli od njihove medijane i na korišćenju M ocenitelja iz jednog koraka kao ocenitelja lokacije. PB koeficijent, u oznaci r_{PB} se računa prema sledećem algoritmu (prema Wilcox, 2005):

1. Definiše se $m = [(1- \beta)n + 0.5]$, pri čemu je $0 \leq \beta \leq 0.5$ (uglaste zagrade znače da se m zaokružuje na ceo najbliži manji broj);

2. Izračunaju se apsolutna odstupanja rezultata na svakoj varijabli od medijane te varijable: $W_i = |x_i - \text{Mdn}|$;

3. Poređaju se odstupanja iz koraka 2 po veličini;

4. Na osnovu koraka 3 odrede se robustne mere lokacije i skale;

5. Tako standardizovani podaci se ponderišu saglasno njihovoj iznimnosti: što više odstupaju od glavnine podataka dobijaju manji ponder;

6. Iz ponderisanih podataka računa se korelacija po standardnom obrascu za Brave-Pirsonovo r.

Podroban prikaz i objašnjenje načina računanja PB korelacije može se videti na sledećoj internet stranici:

http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/pbendcr.htm

Robustni ocenitelj O-tipa za povezanost dveju kvantitativnih varijabli

Najčešće korišćeni ocenitelj O-tipa zasniva se na traženju, u skupu svih n jedinica posmatranja, određene proporcije (na primer 0.5 ili 0.8) jedinica posmatranja čiji rezultati na dvema varijablama daju najmanju determinantu matrice kovarijansi dveju varijabli.[6] To se u suštini svodi na algoritamsko traženje podskupa jedinica posmatranja, tako da na tom podskupu jedinica posmatranja varijanse na obema varijablama budu što manje a istovremeno korelacija između tih dveju varijabli što je moguće veća. Ovaj ocenitelj poznat je kao MCD (od engleskog Minimum Covariance Determinant) ocenitelj korelacije. Najpoznatiji algoritam za računanje MCD ocenitelja korelacije, koji je implementiran i u statističkom paketu R, predložili su Ruseu i Van Drisen (Rousseeuw & Van Driessen, 1999).

Mesto i uloga robustnih ocenitelja u analizama podataka: stanovište autora ovog teksta

S obzirom na to da su „klasični“ ocenitelji lokacije, skale i povezanosti dveju kvantitativnih varijabli statistički optimalni ocenitelji kada su zadovoljeni uslovi za njihovu primenu, autor ovog teksta ne zalaže se za potpunu zamenu „klasičnih“ ocenitelja robustnim. Možda je najbolja strategija u sadašnjem trenutku razvoja robustne statistike uvek primeniti i „klasične“ i robustne ocenitelje. Ukoliko su dobijene vrednosti „klasičnih“ i robustnih ocenitelja relativno bliske, dovoljno je prikazati samo vrednost „klasičnih“ ocenitelja. Ukoliko se pak vrednosti „klasičnih“ i robustnih ocenitelja izrazito razlikuju, potrebno je „pozabaviti“ se opet samim početnim podacima. Ova izrazita razlika u vrednostima „klasičnih“ i robustnih ocenitelja može ukazivati na postojanje iznimaka u podacima kojima se treba „pozabaviti“. Iznimcima se treba podrobno „pozabaviti“ jer oni nisu važni samo

[6] Osnovno objašnjenje determinante matrice može se naći u Tenjović (2020), Dodatak 2.

statistički već i psihološki.[7] Da li, u slučaju postojanja autlajera, treba prikazati samo rezultate dobijene primenom robustnih ocenitelja? Možda je najbolje i u tom slučaju prikazati i rezultate koji se dobijaju primenom „klasičnih" ocenitelja, ali uz ukazivanje na to da se u uslovima postojanja autlajera u podacima u „klasične" ocenitelje može imati manje poverenja nego u robustne ocenitelje.

Reference na koje se upućuje u tekstu

Bickel, P. J., & Lehmann, E. L. (1976). Descriptive statistics for nonparametric models III. Dispersion. *The Annals of statistics*, *4*(6), 1139–1158.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and Kurtosis in Real Data Samples. *Methodology*, *9*(2), 78–84.

Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika*, *40*, 318–335.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31,* 144–152.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Tenjović, L. (2020). *Statistika u psihologiji, drugo dopunjeno i izmenjeno izdanje*. Beograd: Centar za primenjenu psihologiju.

Tukey, J. W. (1962). The Future of Data Analysis. *Annals of Mathematical Statistics*, *33*(1), 1–67. doi:10.1214/aoms/1177704711

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing, Second edition*. Burlington, MA: Elsevier Academic Press.

[7] Na primer, ukoliko na upitniku suicidalne ideacije postoje osobe čiji je rezultat izrazito veći od rezultata svih ostalih ispitanika, to može ukazivati na osobe koje možda planiraju samoubistvo. Ili ako u nekoj grupi dece postoje ona deca čija je inteligencija izrazito visoka a školski uspeh izrazito slab, očigledno se radi o strukturnim iznimcima čija iznimnost može biti signal za intervenciju školskog psihologa.