

MJERNE SKALE I KONVERZIJE JEDNE U DRUGU

Vlado Simeunović
Univerzitet Sinergija
Bijeljina

Rezime

Jedno od početnih pitanja svake statističke obrade je nivo mjerenja, odnosno tip mjerne skale na kojoj su izmjerene varijable iz dotičnog istraživanja. Teorijski uslovi korišćenja pojedinih statističkih tehnika su uglavnom jasni i, sa stanovišta istraživača restriktivni. Oni se, pored tipa mjerne skale, odnose i na vrstu distribucije po kojoj se individualne razlike pretpostavljene populacije raspođeljuju.

U praksi se zadovoljenost tih uslova obično ne provjerava, a eventualna saznanja o odstupanju se zanemaruju. Prvenstveno se zanemaruje pretpostavka normalne raspodjele i prateći uslov linearnosti veza, dok se tip mjerne skale koliko toliko poštuje. To poštovanje se, međutim, svodi na razdvajanje nominalnih varijabi od ostalih, dok se skoro rutinski ordinalne varijable obrađuju kao intervalne.

Ključne riječi: mjerne skale, konverzija, mjerenje, normalna raspodjela.

Uvod

Problematika mjerenja je generalno poznata stvar i, slično pojmu slučaja, mjerenje to bolje služi što se manje o njemu razmišlja. Jer mjerenje ima razne aspekte i "suštine" i kao i svi trajni problemi ljudskog uma, zaokuplja um utoliko više ukoliko se više njime bavimo. Mjerenje je i pridruživanje brojčanih oznaka objektima i preslikavanje skupa objekata na skup brojeva i objektivna i numerička konceptualizacija prirode... Ono može imati potpuno formalnu sadržinu, ako ga tumači matematičar, ili potpuno epistemološku, ako ga tumači filozof, ili pragmatičku, ako ga tumači istraživač-praktičar. Operacija mjerenja je u formalnom smislu preslikavanje skupa objekata na skup brojeva. Samo preslikavanje mora biti takvo da zadovoljava određene uslove među kojima je najvažniji onaj da je ponovljivo i jednako za sve objekte. Rezultat mjerenja takode mora biti takav da dobijene mjere omogućavaju da dva identična objekta imaju istu mjeru i dva različita objekta različitu mjeru.

Ovakva formalna definicija ne objašnjava šta se mjeri i ne ulazi u problem definisanja atributa kao komplementa objekta, odnosno supstancije. Mi ne možemo da spoznamo objekt bez atributa, niti može da postoji atribut bez objekta. Ovaj nedostatak formalnih definicija ima stalno povratno dejstvo na teoriju mjerenja kroz npr. različite modele mjerenja i različite modele grešaka. Formalnim definisanjem mjerenja smo unaprijedili model (teoriju), ali ne treba nikako zaboraviti da kad se izbací predmet mjerenja iz definicije mjerenja, onda se izlažemo riziku da sva svojstva postanu svojstva modela, a ne predmeta mjerenja.

Kasnije ćemo objasniti kako su normalna distribucija i kontinuiranost mjerenja upravo takva svojstva-svojstva matematičkog modela.

Međutim, nama nije cilj da doprinosimo konfuziji i uvećavamo potencijalno nepotrebna znanja. Zato ćemo prestati da raspravljamo o "suštini" mjerenja i ograničićemo se samo na ono što je po našem uvjerenju neposredno vezano za temu ovog rada. Naime, nas ovdje mjerenje interesuje utoliko što je zapravo ono sâmo razlog za postojanje ovog rada. Mjerenje je, naime, "krivac" za postojanje nominalnih, nenumeričkih ili kategorijalnih

podataka kojima se bavimo. Izrada i razrada nenumeričkih algoritama ima za cilj da učini dostupnim multivarijatne statističke metode istraživačima koji imaju nenumeričke podatke. Zato ćemo u ovom poglavlju u osnovnim crtama iznijeti naše stavove koji se tiču mjernih skala, posebno nenumeričkih.

U praksi istraživača se već odavno ustalio problem da "statistika" postavlja uslove pod kojima se neka metoda za obradu podataka može primjenjivati. Neki uslovi su lako shvatljivi, prihvatljivi i razumljivi, kao što je naprimjer problem koreliranosti podataka kod istraživačkih nacrta sa ponovljenim mjerenjima. Istraživači mogu lako da identifikuju da li su njihovi podaci ponovljena mjerenja na istim ispitanicima i što je možda još važnije, da razumiju smisao postavljenih uslova.

Postoji međutim, jedna grupa uslova koji već decenijama predstavljaju noćnu moru za istraživača. Oni se prvenstveno odnose na pitanje mjerne skale i, donekle povezano s time, na pitanje modela raspodjele vjerovatnoća. Što se tiče prvog, fundamentalni pristup zahtijeva, kao prvo, da se mjerna skala precizno identifikuje, a kao drugo, da se primijenjene statističke metode striktno saobraze skali. Problemi počinju već od toga kako se pravi podjela skala, zatim šta su osobine koje pojedina skala mora imati, na šta se zahtjevi odnose itd.

1. Problem korišćenja mjernih skala

Da bismo izlaganje učinili kratkim i jasnim potrebno je da odmah definišemo nekoliko pojmova koje ćemo stalno koristiti.

Preziciranje pojmova ćemo obaviti u dva koraka: u prvom ćemo fiksirati nekoliko najčešćih termina, a kada budemo izvodili matematičke dokaze izvršićemo i sva dodatna, a potrebna preciziranja. Objektom mjerenja ćemo zvati entitet čije karakteristike mjerimo i posmatramo. U praktičnoj primjeni statistike ovi termini se "slučaj" (case) ili "posmatranje" (observation). Svi ovi termini su ekvivalentni sa našim objektom.

Atribut je svojstvo (ponašanje, funkcija...) objekta koje mjerimo. Ako mjerimo neko svojstvo "attrib" na n objekata dobićemo varijablu "attrib". Ta varijabla se standardno prikazuje vektorom dužine n , tj. sadrži n ishoda pojedinačnih mjerenja. Uobičajeno je u istraživanjima da postoji više mjerenih atributa, recimo m , pa dakle kao rezultat imamo m vektora dužine n . Uobičajeno je, takode, da se svi ti podaci predstavljaju matricom, recimo X , čije su dimenzije $n \times m$.

Vektor varijable jeste dužine n , ali to ne znači da ima n različitih vrijednosti. U zavisnosti od nivoa mjerne skale, kao i prirode samog atributa, broj različitih vrijednosti k može biti $2 < k < n$. Različite vrijednosti varijable, odnosno različita značenja atributa zvaćemo modalitetima ili kategorijama.

U praksi istraživača se već odavno ustalio problem da "statistika" postavlja uslove pod kojima se neka metoda za obradu podataka može primjenjivati. Neki uslovi su lako shvatljivi, prihvatljivi i razumljivi, kao što je naprimjer problem koreliranosti podataka kod istraživačkih nacrta sa ponovljenim mjerenjima. Istraživači mogu lako da identifikuju da li su njihovi podaci ponovljena mjerenja na istim ispitanicima i što je možda još važnije, da razumiju smisao postavljenih uslova.

Postoji međutim, jedna grupa uslova koji već decenijama predstavljaju problem za istraživače. Oni se prvenstveno odnose na pitanje mjerne skale i, donekle povezano s time, na pitanje modela raspodjele vjerovatnoća. Što se tiče prvog, fundamentalni pristup zahtijeva, kao prvo, da se mjerna skala precizno identifikuje, a kao drugo, da se primijenjene statističke metode striktno saobraze skali. Problemi počinju već od toga kako se pravi podjela skala, zatim šta su osobine koje pojedina skala mora imati, na šta se zahtjevi odnose itd.

Recimo, s obzirom na dopustivost transformacija, skale se dijele na nominalne, ordinalne, intervalne i ratio skale (sa nekim kombinacijama), a obično se posebno izdvajaju dihotomne (binarne) varijable kao univerzalne. Postoji zatim i podjela na kategorijalne skale, koje nastaju klasifikacijom objekata, i mjerne skale, koje nastaju mjerenjem. Ove prve se nekad zovu nenumeričke, odnosno kvalitativne, a ove druge numeričke, odnosno kvantitativne. Nekad se numeričke mjerne skale razlikuju s obzirom na neposredni postupak dodjeljivanja brojeva, pa imamo brojanje i mjerenje. S obzirom na drugu grupu problema, primjenjivost matematičkih modela distribucija vjerovatnoća, razlikuju se diskretne i kontinuirane skale, od kojih se one prve nazivaju takode i kategorijalnim. Sva ta šarolikost se može pročitati kod Nunnally-a (1967), Ghiselli-a (1964), Burr-a (1974), Stilson-a (1966), Younga (1981) itd.

Kao što se vidi iz godina izdanja, klasifikacija skala pripada samoj klasici statistike. No već onda je, recimo kod Nunnally-a (1967), bilo razumnog kriticisma u pogledu fundamentalističkog pogleda na mjerenje i zahtjeva koji su iz takvog pristupa proistekli. On smatra, recimo, da konvencija mora biti veoma važna osobina kod skala. Poput temperaturnih skala, Celzijusove i Farenhajtove npr., koje se uspješno koriste i danas bez obzira što je određivanjem apsolutne nule skala u Kelvinima postala ratio skala. Pridržavanje konvencije ima "društveni" kvalitet jer omogućava uporedivost rezultata i komunikaciju između istraživača, a pošto ne smijemo smetnuti s uma da je nauka društvena djelatnost, kriterijumi socijalne interakcije i komunikacije su realno utemeljeni.

U međuvremenu, praktično je pokazano da monotone transformacije skala ne mijenjaju značajno koeficijente korelacije, F i t testove. Osim toga, razvijeno je u praksi dosta ležernosti i "tolerancije" u odnosu na to da li su neki konkretni podaci prihvatljivi za neku konkretnu statističku metodu za analizu podataka - u pogledu skale. Interpretabilnost i stabilnost rezultata su samo potvrdili da se to isplati.

Dalje, razvoj multivarijantnih tehnika za obradu nominalnih varijabli je učinio da istraživači danas mogu kvalitetno analizirati sve tipove podataka. Upadljiv primjer za to je set programa iz tzv. holandske škole koji se po matematičko-statističkom supstratu svrstavaju pod različite naslove kao što su analiza homogenosti, optimalno skaliranje, multipla korespondentna analiza i komponentna analiza, a prihvataju varijable svih tipova (de Leeuw, 1997). Vodeći autor iz te škole, de Leeuw, je razvio vlastiti formalizam za definiciju tzv. multivarijable, koji uopšte ne uključuje pitanja mjerne skale (de Leeuw, 1993, 1997 i de Leeuw, Michailides, Wang, 1997, Young, 1981).

Nama se čini da je današnje stanje stvari takvo da su podaci sa ordinalne skale dovoljno dobri za primjenu bilo koje tehnike koja počiva na pretpostavci multivarijantne normalne raspodjele. U pogledu nominalnih varijabli, kod njih treba koristiti binarizaciju da bi se dovele u "bez-skalni" oblik koji je pogodan za bilo kakvu dalju analizu. U gornjim pasusima smo spomenuli da su koeficijenti korelacije, pa tako i sve metode koje se baziraju na njima, robusni na monotone transformacije. Otuda se može zaključiti da je poredak (rang) objekata ono što presudno utiče na visinu koeficijenta.

Postoji još jedan ugao gledanja na vrstu mjerne skale. Naime, kad bolje pogledamo, potpuno nominalne varijable nisu tako česte u ekonomiji. Većina varijabli koje tretiramo kao nominalne su parcijalno nominalne, odnosno parcijalno ordinalne. To znači da je opseg varijabli (sve moguće vrijednosti koje ona zauzima) sastavljen od jednog ili više podopsega koji su ordinalni i jedne ili više vrijednosti koje su van tog (tih) redosljeda. Na primjer, kod mjerenja stavova, sve alternative su poredane u niz po jačini slaganja sa nekim iskazom, osim

alternative "ne zna". Slično važi za zanimanje (...,"NK", "KV", "VKV", "SSS", "VŠS",..., "privatnik", "učenik",...), bračni status ("neoženjen/neudata" nasuprot svih drugih podvrsta) itd.

2. Kontinuiranost i diskontinuiranost

Razmotrićemo ovdje još jedno svojstvo podataka koje je za naš rad neposredno interesantno, a dozvoljava da usput razmotrimo značaj statističkih uslova generalno. To je pitanje kontinuiranosti, odnosno diskontinuiranosti. U jednom generalnom, da kažemo ontološkom smislu, to pitanje nam se već nametalo u radu, a nametaće se i kasnije. Stalno se, naime, srećemo sa konstatacijom da postoji imanentna diskrepancija između strukture našeg znanja koja je diskontinuirana, tj. podijeljena u pojmove; i strukture svijeta, pa i naše prakse, koja je kontinuirana. No, ovdje se nećemo baviti tim najširim aspektom, nego jednim specifičnijim koji se odnosi na rezultate mjerenja.

Naime, predmet ovog rada je komparacija postupaka rješavanja određenih statističkih problema koji operišu kategorijalnim varijablama. Izraz "kategorijalan" u osnovi znači prekidnost - diskontinuiranost, ali kad se primijeni na mjerenje vidimo da razvlači jednu dimenziju koja se u tom, ili sličnom obliku proteže i u svim drugim aspektima mjerenja. Na jednom polu te dimenzije su objekti koje mjerimo i njihove osobine. Naime, osobina koju posmatramo može biti takva da dozvoljavava samo klasifikaciju objekata u dvije ili više kategorija. Recimo pol, bračni status, postojanje ili nepostojanje neke bolesti, riješenost ili neriješenost nekog slučaja itd. Na drugom polu te dimenzije su prekidne brojne (numeričke) skale. Njihova prekidnost (diskretnost) nastaje tako što nam mjerni instrumenti imaju ograničenu preciznost, što nam nije potrebno da mjerimo preciznije, ili što koristimo brojanje kao mjerenje.

Dakle, s jedne strane imamo osobine, a s druge mjerenje. I jedno i drugo može biti bilo kontinuirano, bilo diskretno. Ako mjerimo osobine koje su kategorijalne, rezultirajući podaci će biti diskretni bez obzira na mjerne skale i instrumente koje koristimo. Takođe, ako nam mjerni instrumenti prave samo grube razlike između objekata, podaci će biti kategorijalni, bez obzira što osobina nije. Izabrali smo da ovo svojstvo nazovemo dimenzijom kako bi ukazali da postoje prelazi i mješavine. Nedovoljno precizni instrument, npr., primijenjen na kategorijalnu osobinu može rezultirati u manjem broju kategorija nego što osobina ima.

Da li će tretman osobine koja je generički kategorijalna biti u principu nominalan? Nažalost ne, jer smo odmah na početku rekli da je mjerenje utoliko jasnije ukoliko se manje analizira. Posmatrajmo recimo varijablu koja se zove plaćanje rente. Ona je kategorijalna, tj. diskretna, ali nije nominalna. Pošto je svako mjerenje skup pravila za zadavanje brojeva pojavama, ovdje je pravilo koje se koristi prosto brojanje. Međutim, možemo taj mjerni "instrument" zamijeniti jednim grubljim i spomenutu osobinu mjeriti samo kao "plaćena renta - nije plaćena renta".

Da rezimiramo, prekidne mjerne skale i rezultirajući kategorijalni podaci mogu nastati kad su varijable koje se mjere diskretne i/ili kad su mjerni instrumenti i postupci diskretni. Za razliku od njih, kontinuirani podaci moraju biti kontinuirani sa obje strane. Kontinuiranost ima značajno mjesto u statistici, ne toliko zbog mjerne skale, koliko zbog modela normalne raspodjele. Model normalne raspodjele je ugrađen u svu multivarijatnu statistiku tako što su svojstva normalne raspodjele iskorišćena za izvođenje svih relevantnih multivarijatnih koeficijenata i pokazatelja. U univarijatnoj statistici je normalna raspodjela manje prisutna, dok se u statistici zaključivanja ona ne odnosi na podatke, nego na statistike (estimatore), pa otuda i nije relevantna za našu diskusiju.

3. Normalna raspodjela i veze između pojava

Po našem shvatanju slučaja, normalna raspodjela očigledno opisuje ishode koji nastaju na presjeku velikog broja nezavisnih kauzalnih lanaca. Pod nezavisnim podrazumijevamo da su nezavisni od naše praktične aktivnosti u kojoj se slučaj dogodio, kao i da međusobni odnosi lanaca ne stoje u uzročno-posljedičnoj vezi. Zato smatramo da je ona univerzalna teorijska kriva slučajnih događaja. Međutim, mi takođe smatramo da nema "teorijskog" slučaja i kada bi u praksi postavili cilj da provjerimo distribuciju vjerovatnoća nekih ishoda, samo "slučajno" bismo dobili stvarnu normalnu raspodjelu. Teško je, naime, povjerovati u veliki broj (čak šta više bezbroj) kauzalnih lanaca koji su nezavisni kako međusobno tako i od procesa mjerenja. I stvarno, u praksi postoji mnogo podataka o tome da pojave koje bi po svemu morale biti normalno raspodijeljene to stvarno nisu, čak i na vrlo velikim uzorcima. Vjerovatno možemo naše shvatanje odnosa slučaja, individualnih razlika i naučnog istraživanja uopštiti tako da kad god dobijemo bilo kakvu distribuciju posmatranja, posebno ako je ona zvonasta, to znači da je pojava koju mjerimo mnogostruko uslovljena i uzrokovana, ali da je, barem u tom trenutku i na tom uzorku mjerenja, postojalo stabilno kombinovanje svih uslova (uzroka). Naprimjer, čak i ako se radi o mjerenju jedne proste stvari kao što je tjelesna visina, normalna raspodjela koja bi trebala da se dobije govori o velikom broju faktora koji kao svoj rezultat imaju jednu ili drugu visinu organizma. Tu spadaju svakako genetski i kongenitalni faktori, faktori ishrane, doba dana itd., a naravno i ono što standardno zovemo greške mjerenja. Tjelesna visina je zapravo naš koncept, osmišljavanje prirode u smislu geometrizacije, a zvonasta kriva rezultata nam pokazuje njegove slabosti i vrline. Slabosti su u tome što visina organizma nije generička kategorija prirode, nego (slučajni) ishod bezbroj uzroka. Posljedica te slabosti je to što visinu neke osobe ne možemo tačno predvidjeti. Dobra strana koncepta tjelesne visine je u tome što zvonasta kriva svjedoči o zakonitostima i tendencijama u kombinovanju uzroka. Otuda možemo prognozirati očekivanu vrijednost visine svakog pojedinca. Čak štaviše, statistika nam omogućava da varijacije (individualne razlike) iskoristimo za izučavanje uzroka, dovodeći tjelesnu visinu u vezu sa drugim pojmovima.

Veze između varijabli koje se multivarijatno normalno raspodjeljuju imaju još jedno zanimljivo svojstvo - linearnost. Po našem shvatanju slučajnosti i individualnih razlika, linearnost veze je nužna kada do slučajne pojave dolazi usljed brojnih i nezavisnih kauzalnih lanaca. Pod nezavisnošću podrazumijevamo njihovo nezavisno kombinovanje unutar svakog ishoda, u našem konkretnom slučaju objekta, kao i nezavisnost od naših manipulacija sadržanih u operaciji mjerenja. Odstupanje od multinormalne raspodjele i od linearnosti veze znači da naš uzorak objekata nije homogen. Podsjetimo se da mi mislimo da korelacija dvije varijable uvijek ukazuje na postojanje jednog ili više kauzalnih lanaca (uzroka) koji su zajednički tim pojavama i koji dovode do uzajamnog rasipanja (scatter). Kada korelacija nije linearna, smatramo da postoje bar dva poduzorka u našem uzorku objekata u kojima djeluju različiti uzroci, ili različito kombinovani. Drugim riječima, kauzalni lanci nisu nezavisni, nego sami oni imaju korelaciju sa objektima usljed nekog, nama nepoznatog uzroka, ili imaju korelaciju sa samim postupkom mjerenja. Dakle, ako istražujemo vezu između dvije pojave, dobijena korelacija će značiti da na obje pojave vrše sinhroni uticaj neke dublje pojave, koje su ostale van našeg mjerenja. Ako je dobijena korelacija nelinearna, onda možemo biti sigurni da postoji bar jedna, treća varijabla, koja ne djeluje sinhrono na obje varijable. Pod sinhronim podrazumijevamo istosmjerno ili istovjetno djelovanje kod svih objekata.

Ako se podsjetimo našeg piramidalnog topološkog modela, pretpostavka normalne distribucije je bila presudna u horizontalnoj manipulaciji slučajnim pojavama na visokim nivoima piramide. Tako je nastala moćna oblast statistike, ali kad je mentalni tok krenuo naniže, ka istraživačkoj praksi, javio se problem što su podaci diskontinuirani i često nisu normalno raspodjeljeni. U većini istraživanja se ne ispituje eksplicitno i dosljedno tip distribucije podataka, pa čak i kad distribucije "izgledaju" zvonasto one mogu biti bliže nekom drugom modelu, naprimjer modelu beta distribucije. Postoje indicije da se odgovarajućim transformacijama koje uklanjaju greške mjerenja i specifične izvore varijansi može ukloniti i normalnost distribucije iz podataka (Kovačević i Momirović, 1998). Vjerovatno je većini autora iz oblasti statistike intuitivno jasno da je rješenje tog problema "neadekvatno", pa nama nije poznat sistematski i objektivan tretman primjene multivarijantne statistike kad uslovi nisu zadovoljeni, a da on bude recimo na nivou fundamentalnog djela T. W. Andersona (1958).

Istu sudbinu ima i pretpostavka kontinuiranosti. U matematičkom smislu, kontinuiranost koja se pretpostavlja kod normalne distribucije podrazumijeva realni brojni niz (cijeli, racionalni i iracionalni brojevi), odnosno neprebrojivo beskonačni skup brojeva. Tek se očekuje istraživanje koje će imati na raspolaganju podatke sa takve mjerne skale. Sadašnje stanje je takvo da su infinitezimalni podaci (kod kojih razlika između dva susjedna broja teži nuli) nevjerovatni iz više razloga. Kada govorimo o mjernim instrumentima, onda nije samo njihova preciznost ta koja predstavlja ograničenje. Ima indicija, kako je Heissenberg tvrdio u okviru kvantne mehanike, da postoji principijelna nemogućnost nezavisnog mjerenja u mikrosvijetu. Možemo takode da govorimo o tome da su populacije sa kojima mi radimo konačne, ili u najboljem slučaju prebrojivo beskonačne, tako da nikada nećemo imati mogućnost da dokažemo kontinuiranost mjerenja. I na kraju, računanje, a ono se zasada vrši na računarima, dužinom mantise uslovljava prekidnost i podataka i rezultata.

Postoji još jedno, i to vrlo dobro konceptualiziranje razlike između "statističkih uslova" i stvarnih podataka. De Leeuw (1994 i 1995) tumači svoju viziju uloge statistike u nauci, pa statističke metode generalno tretira kao analitičke modele za obradu podataka. U takvoj konceptualizaciji, normalna raspodjela, kontinuiranost itd. su osobine modela, dok osobine podataka s time nemaju ništa. Pitanje pravilne primjene modela je po de Leeuwu heurističko pitanje svake nauke, podjednako kao što je to izbor teorija, hipoteza, metoda itd. On recimo razmatra i mogućnost evaluacije efikasnosti primjene modela i za to nudi pojam kalibriranja (gauges).

Sva prethodna diskusija dopunjuje našu tvrdnju da se multivarijantna statistika mora pomiriti sa ordinalnim varijablama. Očigledno je da su joj sudbina diskontinuirane i ne nužno normalno raspodjeljene varijable, a s obzirom da su kovarijanse robusne na monotone transformacije, smatramo da je tvrdnja dokazana.

Da se sada vratimo na pojam "kategorijalno". Mada postoji izvjesna konfuzija oko upotrebe tog pojma u smislu da se on upotrebljava paralelno sa "diskretno" i "diskontinuirano", kada mi kažemo da su varijable kategorijalne, onda mislimo na to da su one nastale kategorizacijom. Pod kategorizacijom podrazumijevamo razvrstavanje objekata u međusobno isključive klase i dodjeljivanje brojeva tim klasama. Pošto ne podrazumijevamo nikakvu relaciju uređenja između klasa, tretman dodijeljenih brojeva je nominalan, tj. oni imaju status imena tih kategorija.

Ovo ističemo zato što se termin "kategorijalno" može koristiti i na drugi način. Podaci su kategorijalni čim su diskretni (a pošto su uzorci konačni, oni su uvijek diskretni), a mi ih možemo tretirati kao intervalne, ordinalne ili nominalne. Premda je ovaj način gledanja različit od našega, stvarne razlike u pogledu primjene statističkih modela nisu velike. Pri tome

mislimo da svaku intervalnu ili ordinalnu varijablu možemo tretirati kao nominalnu, a da ničim ne oštetimo nominalni model obrade i njegove rezultate. Jedino može biti oštećena intervalna varijabla u smislu gubitka informacija, ali mislimo da ni to ne mora biti u pravilu.

4. Diskretni, kontinualni ili kategorijalni podaci

Osim podjele po tipu skale mjerenja u statistici se koristi i podjela na diskretne i kontinualne podatke (Burr, 1974). Kontinualni podaci nastaju mjerenjem u klasičnom smislu, odnosno upoređivanjem sa standardnom skalom. Maksimalni mogući broj modaliteta zavisi samo od preciznosti mjerenja i teorijski je beskonačan. Takođe, ako imamo bilo koji interval između dvije mjere, možemo ga podijeliti na beskonačno mnogo podintervala. Diskretni podaci obično nastaju prebrojavanjem i u takvim podacima postoji jedan najmanji interval koji se dalje ne može dijeliti (recimo, porodica može imati 0, 1, 2,... djeteta). Prebrojavanje se može odnositi na brojanje /učestalost/pojavljivanja nekog događaja, ali i na određivanje količine prisustva nekog atributa. Za oba tipa podataka postoje posebne funkcije raspodjela, tj. posebni modeli raspodjela podataka u populaciji. Za diskretne podatke to su npr. binomna, Puasonova, multinominalna, a za kontinualne normalna, F, t i druge raspodjele.

Ponekad postoji konfuzija između ove podjele i podjele po tipu skale, tj. nominalni podaci se obično nazivaju diskretnima. Nama se čini da to nije opravdano, odnosno da suština nominalnih podataka nije diskretnost. Konfuzija nastaje upravo zbog različitih mogućnosti primjene prebrojavanja. Npr. brojanjem automobila koji prođu jednu tačku na putu možemo formirati varijablu "frekvencija saobraćaja" na dva načina: tako da njene vrijednosti budu izražene u broju automobila, ili da prosto sadrži "1" i "2" u zavisnosti da li u toku sata zamišljenu liniju prođe više od 100 vozila, ili manje. U drugom slučaju možemo dalje prebrojavati koliko je prođe vozila po pojedinim tipovima u uzorku, koliko muških a koliko ženskih vozača itd. Nominalne varijable su obično numeričke i kao takve, obično su diskretne, ali ne zato što su nastale prebrojavanjem, nego numerisanjem (a najpraktičnije je numerisanje prirodnim brojevima). Za podatke nastale brojanjem učestalosti važe svi statistički modeli koji zahtijevaju kvantitativne varijable, dok za nominalne podatke to nije slučaj.

S druge strane, prebrojavanje količine produkuje najstrože definisane kvantitativne mjerne skale, recimo racio skal, gdje je osim jednakosti intervala zadovoljen i uslov jednakosti odnosa. Prebrojavati se mogu i skupovi koji jesu prebrojivi, ali su i teorijski beskonačni (npr. broj pojavljivanja greške u proizvodnji nekog komada u industriji). Kod takvih (diskretnih) varijabli teorijski nije primjenjiv model normalne raspodjele, no u praksi se to može potpuno zanemariti. Pogotovo ako broj intervala diskretne varijable omogućava da svaki objekat ima jedinstvenu mjeru, ili ako je veličina intervala veća od greške mjerenja.

Prebrojavanje može da produkuje najviši tip mjerne skale, ali ne mora. Naime, ako brojimo pojave koje su efemerne, hipotetičke i sl., ne možemo očekivati da će rezultat našeg mjerenja biti bolji. Posebno ćemo imati problem određivanja nule, jer nećemo moći tvrditi da neki atribut na nekom objektu ne postoji (tj. da je jednak nuli), ako je taj atribut efemeran ili hipotetički... A nijedna skala nije dobra ako nema fiksnu nulu. No, apsolutno isti problem postoji kod bilo kog mjerenja

Dalje terminološko raslojavanje je između kontinualnih i kategorijalnih varijabli. Mada podsjeća na gornju podjelu, ovdje se osobina diskretnosti ne vezuje za prebrojavanje, nego za postojanje manjeg broja distinktivnih klasa. Podjela na kontinualne i diskretne podatke ne razlikuje nominalne od intervalnih varijabli, niti pravi bilo kakvu naročitu praktičnu razliku između tipova podataka, ali je teorijski dobro zasnovana na tipovima raspodjela slučajnih veličina. Za razliku od toga, pod pojmom "kategorijalno" se u ovoj drugoj podjeli uglavnom podrazumijeva nominalni (ili ordinalni) nivo mjerenja, tj. pravi se jedna bitna tipološka razlika, ali teorijska zasnovanost nam se čini

više intuitivna. Razlika između nominalnog i intervalnog nivoa mjerenja nije razlika između prirodnih i racionalnih brojeva, što bi podjela na kontinualne i kategorijalne varijable mogla da sugerise.

5. Definisanje pojmova - uzorci objekata i atributa

S obzirom da smo već obavili prve korake u formalizaciji problema, a i s obzirom na to će u daljem izlaganju potreba za formalizacijom samo rasti, uvešćemo sada osnovne pojmove i označavanje na koje ćemo se kasnije pozivati.

Neka je $\mathbf{e} = (e_i; i=1, \dots, n) = \mathbf{P}$ slučajni uzorak objekata iz neke višestrukoheterogene populacije $\mathbf{P}$. Objekte iz $\mathbf{P}$ možemo svrstati (klasifikovati) u niz grupa, po različitim atributima. Kao što smo ranije rekli, ovo klasifikovanje nije ništa drugo do mjerenje tih atributa na nominalnom nivou. Naime, svaka nominalna varijabla predstavlja jednu moguću klasifikaciju objekata. Pomoću te klasifikacije se objekti svrstavaju u klase-grupe. Recimo zasada da to svrstavanje može biti potpuno i jednoznačno, ali i ne mora. Odnosno, slučaj koji je najpregledniji, najjasniji i koga je matematički najlakše obraditi je onaj kada *svaki* objekat bude svrstan u *samo jednu* klasu. Nasuprot toga, postoji mogućnost da neki objekti uopšte ne budu svrstani ni u jednu klasu, kao i da neki objekti budu svrstani u dvije, ili više klasa. U daljem radu ćemo još spominjati te varijante.

U cilju opštosti i tačnosti, definišimo $\mathbf{c}$ kao skup svih mogućih i dopustivih klasifikacija objekata, odnosno svih mjerljivih atributa u $\mathbf{P}$. Analogno uzorku objekata koga smo označili sa $\mathbf{e}$, recimo neka je $\mathbf{w} = (w_k; k=1, \dots, g) = \mathbf{c}$ uzorak iz skupa svih dopustivih klasifikacija, odnosno atributa. Članovi skupa su, dakle, nominalne varijable. Svaka varijabla iz $\mathbf{w}$ klasifikuje objekte iz $\mathbf{e}$ u grupe, pa, dakle, možemo reći da su $w_k = (\omega_{kj}; j=1, \dots, m_k)$ skupovi kategorija koje su obuhvaćene tim nominalnim varijablama.

Ovime smo, dakle, definisali populacije objekata i atributa i njihove uzorke.

6. Binarizacija

Pod pojmom binarizacije podrazumijevamo prevođenje neke nominalne varijable sa m modaliteta (značenja) u tipično nekoliko varijabli od kojih svaka ima samo dva modaliteta - tipično nulu i jedinicu.

U prethodnom pasusu smo dva puta upotrebili termin "tipično" jer u ovom trenutku imamo namjeru da binarizaciju predstavimo u što opštijem svjetlu. Naime, binarizacija u užem smislu (tipičnom) je razbijanje jedne nominalne varijable na onoliko binarnih - dihotomnih - bulovih varijabli koliko ta nominalna varijabla ima modaliteta, odnosno m njih, pri čemu su dva značenja ovih binarnih varijabli kodirana kao 0 i 1. Ta operacija se negdje naziva još i dihotomizacijom.

Ovakva transformacija, odnosno preslikavanje se može proizvesti automatski, jednostavna je, matematički pregledna jer je jednoznačna i zbog svega toga je smatramo tipičnom. No, ova standardna binarizacija nije jedini mogući način konverzije nominalnih varijabli u intervalne. Da nabrojimo koje sve mogućnosti postoje:

1. Nominalne varijable se mogu ordinalizovati na osnovu istraživačeve stručne procjene, ali i računski, pomoću maksimizacije korelacija nominalne varijable sa jednom, ili više referentnih intervalnih varijabli.
2. Mogu se koristiti postojeće procedure u raspoloživim statističkim paketima koje vrše konverziju skala. Rezultat takvih transformacija su, kao i u prethodnom slučaju, varijable koje u broju i nazivu odgovaraju ulaznim varijablama, ali im je način skorovanja promijenjen tako da zadovoljava uslove intervalne skale (redosljed, jednaki intervali i si.)
3. "Najčistiji" način binarizacije nominalnih varijabli opisali smo u prethodnom dijelu

teksta. Njegov rezultat je matrica, nazovimo je S_k ($k=1, \dots, g$), za svaku od g varijabli, reda (n, x, m_k) . Dakle, za svaku kategoriju varijable w_k postoji jedna kolona u matrici S_k . Elementi matrice su samo nule i jedinice, s time da u jednom redu može biti samo jedna jedinica. Ovakve matrice se nazivaju i binarne, Bulove, indikatorske, ali ponegdje (kao u SAS-u) matricama nacrt (design matrix).

Ovakvo generisana matrica ima elemente koji zadovoljavaju sljedeće uslove (za $i=1 \dots n$; $k=1 \dots g$ i $j=1 \dots m_k$):

$$\left\{ \begin{array}{l} S_{ikj} = 1 | e_i w_{kj} \\ S_{ikj} = 0 | e_i w_{kj} \\ \sum_j S_{ikj} = 1 \end{array} \right.$$

gde treći uslov kazuje da objekat e_i može pripadati samo jednoj kategoriji date varijable w_k . U matematičkoj logici se za takve matrice kaže da su u potpunoj disjunktivnoj formi.

I na kraju, za ovakvu matricu S_k možemo reći da je ortogonalna matrica (ali ne i ortonormalna), tj. da zadovoljava uslov:

$$S_k^t S_k = \text{diag}(n_1, \dots, n_j, \dots, n_{m_k}),$$

gdje su n_i sume odgovarajućih kolona u S_k , a $\sum_i n_i = n$

Za ovakav način binarizacije možemo intuitivno reći da je potpun jer zadržava sve informacije koje nosi inicijalna varijabla. Međutim, radi se o tome da je on i redundantan, jer je jedna od binarnih varijabli, odnosno jedna od kolona matrice S_k suvišna. Ako ova matrica ima svojstva definisana u datim formulama, onda ćemo svaku od njenih kolona uvijek moći rekonstruisati iz ostalih, odnosno svaka od njenih kolona je linearna kombinacija ostalih. U terminima ranga matrice reći ćemo da je rang ove matrice $m_k - 1$.

4. Osim ovakvog načina binarizacije, može se zamisliti još i mnogo drugih. Uzmimo na primjer varijablu "bračno stanje" sa modalitetima: 1=oženjen/udata, 2=neoženjen/neudata, 3=udovac/udovica i 4=razveden/razvedena. Od nje možemo napraviti sljedeće binarne varijable:

$$\begin{array}{ll} \text{A. U braku:} & \left\{ \begin{array}{l} 1 = \text{modalitet 1} \\ 0 = \text{svi drugi modaliteti} \end{array} \right. \\ \text{B. Neoženjen/neudata:} & \left\{ \begin{array}{l} 1 = \text{modalitet 2} \\ 0 = \text{svi drugi modaliteti} \end{array} \right. \\ \text{C. U prekinutom braku:} & \left\{ \begin{array}{l} 1 = \text{modalitet 3 i 4} \\ 0 = \text{svi drugi modaliteti} \end{array} \right. \end{array}$$

Pored ovakve binarizacije možemo spojiti varijable B. i C. u jednu sa nazivom "van braka" i odgovarajuće je skorovati. Svaka od ovih varijanti zadovoljava uslove matrice S_k kako su dati pod 3., osim što broj kolona nije jednak broju kategorija.

5. Binarizacija iz tačke 3. pretpostavlja (i čuva) svojstvo nominalne varijable da grupiše objekte u distinktivne klase. No, u istraživanjima se često javljaju nominalne varijable koje nemaju to svojstvo, kao npr. "koji muziku najčešće slušate (zaokružite najviše 3 odgovora)". Ako se ovakva varijabla binarizuje na način opisan u tački 3., dobijena matrica S_k će imati više od jedne jedinice u redu.

Standardno rješenje u ovakvom slučaju je razbijanje varijable u nekoliko, standardno onoliko koliko je ispitaniku dozvoljeno odgovora, recimo "I odgovor", "II odgovor", "III odgovor" itd., od kojih svaka ima isti broj modaliteta kao i originalna. Na taj način, višestrukim razbijanjem i zatim binarizacijom možemo problem svesti na onaj izložen u tački 3. Ovo rješenje dovodi neke stvari u red, ali ne sve. "Konfiguracija" višestrukih ispitanikovih odgovora je informacija sama po sebi, a u zavisnosti od tipa kasnije obrade (naprimjer, analize kontingencionih tabela) se ta informacija može potpuno izgubiti.

Dalje, kada dobijemo binarnu matricu S_k , prirodno se javlja potreba za nekom vrstom ponderisanja. Recimo, možemo imati osjećaj da je za potrebe našeg istraživanja neuporedivo značajnije ukoliko je ispitanik "u prekinutom braku" (vidi gore tačku 4.), nego u nekom drugom statusu. Sljedstveno tome, bićemo u iskušenju da jedinicu na toj varijabli zamijenimo recimo dvojkom, ili čak nekim razlomljenim brojem. Uopšte, postavlja se pitanje da li matrica S_k mora imati nule i jedinice, ili može imati bilo koje druge vrijednosti, pod uslovom, naravno, da zadovoljava sve druge uslove. Sva ta pitanja nedvosmisleno zaslužuju pažnju.

Postavlja se pitanje, zašto je binarizacija uopšte potrebna? Binarne varijable, naime, imaju posebno svojstvo, a to je da zadovoljavaju uslov za intervalnu skalu, premda svaka od dvije kategorije može imati nominalni "značaj" (odnosno ordinalni). O čemu se radi. Uslov da varijabla bude intervalnog tipa se izvodi iz opšteg pristupa klasifikaciji tipova mjerenja, odnosno skala. Taj opšti pristup kaže da je svaka skala mjerenja definisana skupom dozvoljenih transformacija dotične varijable, tako da nove vrijednosti ne promijene naš sud o bilo kom izmjerenom objektu. Tako su nominalne varijable podložne bilo kakvim transformacijama koje će sačuvati jedinstvenost kategorija. Pol, naprimjer, možemo "mjeriti" kao 1 i 2, 586 i 359, a da ne govorimo o "m" i "ž" itd. Ordinalne varijable se mogu transformisati samo onim transformacijama koje ostavljaju redosljed objekata nepromijenjen (monotone transformacije). Intervalne varijable se mogu transformisati bilo kojom linearnom transformacijom tipa: $f(x) = px + a$ ($p > 0$; a bilo koje), a da naše saznanje o izmjerenom svojstvu ne bude ni na koji način izmijenjeno (biće npr. sačuvana jednakost intervala, redosljed objekata i si.).

Pogledajmo sada slučaj binarne varijable x sa vrijednostima x_1 i x_2 (x_1 i x_2 su različiti, jer u protivnom to nije varijabla nego konstanta). Nju možemo transformisati bilo kojom linearnom transformacijom u drugu varijablu y sa vrijednostima y_1 i y_2 ; kao $y_i = px_i + a$ ($i=1,2$; $p > 0$), a da svojstva nove varijable ne budu promijenjena. U primjeru iz prethodnog pasusa, sa varijablom "pol", transformacija od 1 i 2 na 586 i 359 bi se izvela pomoću $p = -227$ i $a = 813$.

Dakle, za binarne varijable važi uslov dozvoljenosti linearnih transformacija koji je potreban za intervalne skale. Dakle, binarne varijable su uvijek mogu smatrati intervalnim, bez obzira na "tekuću" primjenu (Anderberg, 1973., i Mirkin, 1980.).

Pomoću standardne binarizacije, nominalnu varijablu "bračno stanje": oženjen/ udata, neoženjen/neudata, udovac/udovica, razveden/razvedena možemo razbiti u 4 binarne varijable, tipično sa vrijednostima 1=da i 0=ne. Tako će jedan udovac na te 4 varijable imati skorove 0, 0, 1, 0. Logika binarizacije se može prikazati kao preslikavanje položaja nekog od mjerenih objekta-entiteta u višedimenzionalni prostor čije su koordinate predstavljene binarizovanim varijablama.

U gornjem slučaju našeg udovca i jedne varijable, njegovog bračnog stanja, on će biti lociran tačno na osi predstavljenoj varijablom "udovištvo" (na ostalim osama će imati vrijednost 0). Tu nema nikakve nove informacije, niti bilo čega neobičnog kada se uporedi sa njegovim skorom 3 na varijabli "bračno stanje". Osim što sada raspolažemo intervalnim varijablama. Kad bismo posmatranog subjekta locirali u prostoru još nekih binarizovanih (nominalnih) varijabli, recimo pola, materijalnog statusa, obrazovanja i sl. dobili bismo manje trivijalno rješenje (nego sa samo jednom varijablom), ali bi suština ostala ista.

Ako radimo sa nominalnim varijablama koje grupišu objekte u distinktivne grupe-klase, onda će svaki objekat pripasti samo jednoj (i nijednoj drugoj) klasi. Varijabla iz gornjeg primjera (bračno stanje) razvrstava objekte u 4 klase. Ako se udruži sa varijablom "pol" proizvešće 8 klasa, itd. Ako te dvije varijable binarizujemo, dobićemo 8-dimenzionalni prostor i svaki će objekt imati koordinate u tom prostoru. Dakle, na prethodni način imamo grupisanje, distinktivne klase, klasifikovanje, a na ovaj način imamo mjerenje, lociranje u prostoru (koji je istina diskretan, ali se može aproksimativno tretirati kao euklidovski), koordinate.

Na isti način kao što je varijabla "bračno stanje" tek jedno zajedničko ime za niz pojava, pravno-formalni ekstrakt za niz ponašanja, događaja, emocija i svojstava koja čine udovištvo iz našeg primjera, tako to isto možemo reći za bilo koju drugu varijablu. Ukratko, uvijek možemo zamisliti neki g-dimenzionalni prostor koji stoji iza naše jedne ili m varijabli (gdje je g znatno veće od m), a koji je rastegnut elementima (ili bolje rečeno fundamentima) tih varijabli i u koji možemo locirati naše objekte.

Priroda, odnosno skala mjerenja tih varijabli pri tome nije bitna. U našem slučaju g-dimenzionalni prostor smo dobili binarizacijom m varijabli. Pošto su polazne varijable diskretne, rezultirajuće koordinate ne mogu biti ništa drugo do 0 ili 1.

Dosada sve što je rečeno je imalo za cilj da rehabilituje nominalni nivo mjerenja koji se obično smatra "nižim" nivom, a potreban je kod rada sa velikim bazama podataka. On je niži samo utoliko što, budući da su individualne razlike izmjerene samo kvalitativno, primjenjivi i raspoloživi fond statističkih metoda obrade nije onako kvalitetan kao kod intervalnih varijabli.

No, ipak, binarne varijable po sebi ne mogu biti ideal mjerenja u bilo kojoj nauci. Znanje koje o nekoj pojavi imamo, a možemo ga opisati samo kao "ima je - nema je", nije veliko znanje. A binarne varijable doslovno ne omogućavaju ništa više. Možemo reći da su one najniži nivo mjerenja uopšte i to u dva smisla. Ako je priroda (generalno) binarne prirode, kao što se činilo da kvantna mehanika dokazuje, onda je to krajnji mogući nivo mjerenja - ali samo pod pretpostavkom da dodemo do "elementarnih čestica" posmatranih pojava. Ako to nije slučaj, onda je binarni nivo mjerenja rezervisan za situaciju kada više i dalje o jednoj pojavi ne znamo - osim da je prisutna ili ne. U svim drugim slučajevima, za binarne varijable je lako pokazati da imaju manju saznavnu vrijednost od bilo koje druge varijable koja ima 3 ili više modaliteta. Npr. ako se vrši transformacija neke skale sa više kategorija na binarnu skalu, onda će aritmetičke sredine i standardne devijacije tih novih varijabli biti konzistentno niže od originalnih (Aiken, 1988).

Zaključno razmatranje

Sve pojave kojih smo svjesni podloženi su nekom obliku 'mjerenja'. Nivoi mjerenja zavise od naše sposobnosti shvatanja tih pojava u n-dimenzionalnom prostoru. Neke pojave samo možemo kvalitativno opisati, što uglavnom govori da se ne radi o kvantitetu, nego o kvalitetu, nasuprot tome druge pojave imaju jasno mjesto na određenoj skali i izražavaju određeni kvantitet. Od nivoa 'kvaliteta' mjerenja zavisi izbor statističkih postupaka za izračunavanje traženih podataka. Mi smo u našem radu pokušali prikazati određene prelazne oblike koji se dobijaju transformacijom jedne u drugu mjernu skalu. Prelazni oblici,

kao što je npr. binarizacija omogućavaju određena poboljšanja i proširivanja upotrebe multivarijantnih statističkih postupaka koji su često neophodni radi prodora u bit određene pojave.

LITERATURA

- [1] Agresti, A. (2002). *Categorical Data Analysis*, New York.
- [2] Bogosavljević, S. (1996). Formalno definisanje i uređenje hijerarhijske klasifikacije, u Bogosavljević, S. i Kovačević, M. (red.): *Analiza grupisanja II*, Savezni zavod za statistiku, Beograd, 43-48.
- [3] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20, 37-46.
- [4] Cox D. R. The Regression Analysis of Binary Sequences. *Journal of the Royal.*
- [5] Cox, D.R. and Snell, E. J. (1989). *Analysis of binary data* (2nd edition). London: Chapman & Hall.
- [6] Cox, R.: Social forces, states and vworld orders' millenium, *Millenium: Journal of International Studies*, 10 (1980)2, p.126-155.
- [7] De Leeuw, J. (1993). Some Generalizations of Correspondence Analysis, Tehnički izveštaj, Preprint 117, UCLA Statistics Series, www.stat.ucla.edu/papers/preprints.
- [8] De Leeuw, J. (1994). Statistics and the Sciences, Tehnički izveštaj, Preprint 152, UCLA Statistics Series, www.stat.ucla.edu/papers.
- [9] De Leeuw, J. (1995). Here's Looking at Multivariables, Tehnički izveštaj, Preprint 210, UCLA Statistics Series, www.stat.ucla.edu/papers.
- [10] De Leeuw, J. (1997). Homogeneity Analysis by Alternating Least Squares, Priručnik uz program, [ftp://ftp.stat.ucla.edu/pub/gifi/](http://ftp.stat.ucla.edu/pub/gifi/).
- [11] De Leeuw, J., Michailides, G., Wang, D.Y. (1997) Correspondence and Component Analysis, Tehnički izveštaj, Preprint 217, UCLA Statistics Series, www.stat.ucla.edu/papers.
- [12] Fajgelj, S. (1995). Faktorska analiza nominalnih varijabli, Magistarska teza, Filozofski fakultet, Beograd.
- [13] Fox, J. (2000). Multiple and generalized nonparametric regression. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No.131. Covers nonparametric regression models for GLM techniques like logistic regression. Nonparametric regression allows the logit of the dependent to be a nonlinear function of the logits of the independent variables.
- [14] Momirović, K. (1988). Uvod u analizu nominalnih varijabli, Metodološke sveske, Jugoslovensko udruženje za sociologiju, Ljubljana.
- [15] Preradović, Lj., Simeunović, V. (2007). *Informatika-Metodologija-Statistika*, Visoka škola unutrašnjih poslova, Banja Luka.