V. STATISTIČKI OPIS UZORKA U POGLEDU JEDNE KATEGORIČKE VARIJABLE I GRAFIČKO PRIKAZIVANJE PODATAKA NA JEDNOJ KATEGORIČKOJ VARIJABLI

Neophodni matematički pojmovi za razumevanje teksta u ovoj glavi:

Osnovni pojmovi teorije verovatnoće

Logaritamska funkcija

Operator sabiranja (sumacioni operator) Σ

Kao što smo to već naveli u Glavi 2 podaci na kategoričkoj varijabli potiču uglavnom sa tzv. nominalnog nivoa „merenja“ zbog čega se često ove varijable zovu i nominalnim varijablama.[5] Dakle, kategoričke varijable su varijable koje sadrže određeni broj iscrpnih i uzajamno isključivih kategorija. Primeri takvih podataka su pol (sa kategorijama muški i ženski),[6] rasna pripadnost (sa kategorijama “kavkaska”, “afroamerička”, “azijatska”...), etnička pripadnost (sa kategorijama “Rom”, “Mađar”, “Srbin”, “Hrvat”...), radni status (sa kategorijama “zaposlen na određeno vreme”, “zaposlen na neodređeno vreme”, “nezaposlen”, “u penziji”), bračni status (sa kategorijama “oženjen/udata”, “zajednički život/nevenčan-a”, “udovac/udovica”, “neoženjen/neudata”), preferencija ruke ili “rukost” (sa kategorijama “levoruk”, “ambidekstar”, “desnoruk”), zanimanje (sa kategorijama “domaćica”, “poljoprivrednik”, “nekvalifikovani manuelni radnik”, “kvalifikovani manuelni radnik”, “zanatlija”, “administrativni radnik”, “stručnjak”, “službenik”, “nastavnik”...), konfesionalna pripadnost (sa kategorijama “pravoslavac”, “katolik”, “grkokatolik”, “protestant”, “budista”, “musliman”...), grupna pripadnost u eksperimentu (sa kategorijama “eksperimentalna grupa” i “kontrolna grupa” ili “ciljna grupa” i “komparativna grupa”) ili dijagnoza mentalnog poremećaja (sa kategorijama “šizofrenija”, “paranoja”, “depresija”, “granični poremećaj ličnosti”...). Kod nekih primera kategoričkih varijabli nisu navedene sve moguće kategorije jer taj broj kategorija može ponekad biti veoma veliki (na primer, na varijabli etničke pripadnosti, zanimanja, dijagnoze mentalnog poremećaja). Kategoričke varijable koje sadrže samo dve kategorije (na primer, pol) nazivaju se dihotomnim a kategoričke varijable sa više kategorija (na primer, zanimanje) predstavljaju politomne varijable.[7] „Merenje“ se kod pravih kategoričkih varijabli zapravo svodi na pridruživanje svakoj jedinici posmatranja oznake kategorije na varijabli kojoj data jedinica posmatranja pripada. Pridruživanje oznake kategorije (koja je najčešće numerička mada može biti i alfanumerička) datoj jedinici posmatranja naziva se kategorisanjem ili klasifikacijom entiteta. Klasifikovanje entiteta u kategorije na kategoričkoj varijabli mora biti iscrpno (svaki entitet mora biti moguće svrstati u neku od kategorija) i nedvosmisleno (kategorije varijable moraju biti uzajamno isključive te dati entitet mora biti moguće klasifikovati samo u jednu kategoriju).[8] Na osnovu oznake koja se pridružuje svakom entitetu na nominalnoj varijabli dobijamo informaciju o tome koji entiteti su međusobno isti a koji se međusobno razlikuju u pogledu kategorijske pripadnosti. Numeričke oznake koje se uobičajeno koriste za označavanje kategorija na nominalnoj varijabli nemaju kvantitativno značenje. Na primer, to što smo muškarce označili cifrom 1 a žene cifrom 2 nikako ne znači da žene imaju više “pola” a muškarci manje “pola”.[9]

[5] U statističkim udžbenicima na našem jeziku ove varijable često se zovu i atributivnim obeležjima (cf. Žižić i sar., 2000).

[6] Premda se pol uobičajeno tretira kao dihotomna kategorička varijabla, tj. varijabla sa dve kategorije, pitanje je koliko to zaista odgovara realnosti. Pojedini autori ističu da pol biološki posmatrano ima barem pet kategorija: muškarci, žene, hermafroditi, muški pseudohermafroditi i ženski pseudohermafroditi (cf. Fausto-Sterling, 2000).

[7] Ovi nazivi potiču od grčkih reči dicha (dvostruka), polys (mnogo, višestruko) i tomia (presecanje, podela).

[8] Veoma podroban prikaz problema klasifikacije može se pročitati u Todorović, 2008, str. 61–67.

[9] Upravo zbog toga što je numeričko kodiranje podataka na kategoričkim varijablma uobičajeno u analizama podataka, mi u ovom tekstu kvantitativne varijable nikada ne zovemo „numeričkim“, iako se termin numeričke varijable često koristi kao sinonimni naziv za kvantitativne varijable.

Prema tome, izbor numeričke oznake za označavanje pojedinih kategorija sasvim je proizvoljan.[10] Jedino što je važno jeste to da različite kategorije imaju različite oznake. Dakle, potpuno je svejedno da li ćemo kategoriju "oženjen/udata" na varijabli bračni status označiti, na primer, cifrom 1 ili cifrom 3. Ukoliko kategorije na dihotomnoj kategoričkoj varijabli označavamo oznakama 0 i 1 tada tu varijablu zovemo binarnom varijablom. Često se, međutim, u kategoričke varijable svrstavaju i varijable sa tzv. "uređenim kategorijama" (na primer, obrazovanje sa kategorijama "nepotpuno osmogodišnje", "osmogodišnje", "srednje", "više", "visoko", "postdiplomsko", procena sopstvenog zdravstvenog stanja sa kategorijama "veoma loše", "loše", "ni loše ni dobro", "dobro", "veoma dobro"), kao i varijable koje se dobijaju kategorisanjem podataka na kontinuiranoj ili diskretnoj kvantitativnoj varijabli (na primer, uzrast sa kategorijama "manje od 18 godina", "od 18 do 24 godine", "od 25 do 34 godine", "od 35 do 54 godine", "od 55 do 64 godine", "65 i više godina" ili broj dece sa kategorijama "nijedno", "jedno" "dva ili tri" i "više od tri"). Kategoričke varijable sa uređenim kategorijama često se nazivaju *uređenim kategoričkim* (engl. *Ordered categorical*) ili *ordinalnim* varijablama, a kategoričke varijable koje se dobijaju kategorizacijom kvantitativnih varijabli *kategorisanim* varijablama. Pri numeričkom označavanju kategorija na ovim varijablama najbolje je kategorije koje imaju "viši rang" označiti većom numeričkom oznakom. Na primer, na varijabli obrazovanje, kategoriju "srednje" prirodnije je označiti većom cifrom nego kategoriju "osmogodišnje". Na taj način se pri analizi podataka na ovim varijablama ne gubi važna informacija o uređenosti kategorija. U ovom poglavlju ćemo uglavnom prikazati statističke mere i postupke koji se mogu primeniti na kategoričke varijable nominalnog tipa.

Različiti načini kodiranja podataka na kategoričkoj varijabli

Kao što smo već istakli, uobičajeno se podaci na kategoričkoj varijabli sastoje u cifarskim oznakama kategorijske pripadnosti za svaku jedinicu posmatranja. Ukoliko, na primer, na varijabli "preferencija ruke" oznakom 1 označimo levoruke ispitanike, oznakom 2 ambidekstre, a oznakom 3 desnoruke, tada će podaci na toj varijabli predstavljati jednu kolonu jedinica, dvojki i trojki u matrici podataka. Ovakav način kodiranja podataka naziva se *kodiranje kategoričke varijable kao faktora*. Naziv ovog načina kodiranja potiče iz statističkih postupaka analize varijanse, u kojima se kategorička varijabla kodirana na ovaj način koristi kao nezavisna varijabla ili faktor. (Postupcima analize varijanse statistički se proverava da li nezavisna kategorička varijabla, tj. faktor ima efekta na zavisnu kvantitativnu varijablu). Premda je za statistički opis uzorka u pogledu kategoričke varijable dovoljno kodirati kategoričku varijablu kao faktor, objasnićemo ovom prilikom i neke druge načine kodiranja kategoričkih podataka koje je neophodno poznavati kako bi kategorički podaci mogli da se koriste u pojedinim složenijim statističkim analizama (generalni linearni model, multipla regresiona analiza, diskriminaciona analiza...).

Podaci na kategoričkoj varijabli mogu se kodirati i u tzv. *kompletni disjunktni oblik*. U tom slučaju se na osnovu jedne kategoričke varijable pravi g binarnih varijabli, pri čemu je g broj kategorija na početnoj kategoričkoj varijabli. Određena jedinica posmatranja na onoj binarnoj varijabli koja predstavlja kategoriju kojoj pripada ta jedinica posmatranja dobija oznaku 1 a na svim preostalim binarnim varijablama dobija oznaku 0. Na primer, kodiranje varijable "rukost" na kojoj su, kada se ona kodira kao faktor, levoruki označeni jedinicom, ambidekstri dvojkom a desnoruki trojkom, podrazumeva pravljenje tri binarne varijable. Na prvoj binarnoj varijabli svi levoruki bivaju označeni sa 1, a ambidekstri i desnoruki sa 0. Na drugoj binarnoj varijabli svi ambidekstri dobijaju oznaku 1 a levoruki i desnoruki oznaku 0, dok na trećoj binarnoj varijabli desnoruki dobijaju oznaku 1, a svi ostali oznaku 0. Na taj način, umesto jedne kolone početnih podataka koja sadrži jedinice, dvojke i trojke

[10] Iz tehničkih razloga u analizama podataka najbolje je i najjednostavnije za označavanje kategorija koristiti numeričke a ne alfanumeričke oznake, tj. cifre a ne slova. Za označavanje kategorija na dihotomnim varijablama najbolje je koristiti oznake 0 i 1 jer korišćenje binarnih varijabli, tj. dihotomnih varijabli na kojima su kategorije označene ciframa 0 i 1 ima određenih prednosti koje ćemo objasniti na odgovarajućim mestima u knjizi.

dobija se tzv. selektorska ili indikatorska matrica sa tri kolone i onoliko redova koliko ima jedinica posmatranja.

Kodiranje u kompletni disjunktni oblik mogli bismo shematski predstaviti na sledeći način:

Rukost kodirana kao faktor	Selektorska matrica		
	Binarna varijabla 1 (Levoruki)	Binarna varijabla 2 (Ambidekstri)	Binarna varijabla 3 (Desnoruki)
1	1	0	0
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0
2	0	1	0

Dakle, levoruki ispitanici su u ovom načinu kodiranja predstavljeni kodom 1 0 0, ambidekstri kodom 0 1 0, a desnoruki ispitanici kodom 0 0 1. Uočimo da broj jedinica u određenoj koloni selektorske matrice odgovara učestalosti date kategorije u početnom nizu podataka. Uočimo, isto tako, da je poslednja binarna varijabla redundantna, tj. izlišna jer se na osnovu prvih dveju binarnih varijabli može iscrpno i nedvosmisleno kodirati kategorijska pripadnost bilo kojeg entiteta: kôd 10 bi označavao levoruke, kôd 01 ambidekstre a kôd 00 desnoruke. Ipak, u pojedinim statističkim analizama pogodno je baš kodiranje u kompletni disjunktni oblik, bez obzira na redundantnost poslednje binarne varijable u odnosu na prethodnih g - 1 binarnih varijabli.

Konstrukcija selektorske ili indikatorske matrice, u oznaci **S**, matematički se može opisati na sledeći način:

$$\mathbf{S} = (s_{ik}) = \begin{cases} 0, & e_i \notin q_k \\ 1, & e_i \in q_k \end{cases} \qquad i = 1, \ldots, n;\ k = 1, \ldots, g$$

Dakle, element s_{ik} u selektorskoj matrici jednak je 0 za jedinicu posmatranja koja ne pripada kategoriji q_k kategoričke varijable, a jednak je 1 za jedinicu posmatranja koja pripada kategoriji q_k.

Budući da ima n jedinica posmatranja i g kategorija matrica **S** je reda n x g.

"Dàmi" kodiranjem (engl. dummy = lažan, kvazi) kategoričke varijable izbegava se redundantnost poslednje binarne varijable: ovim postupkom kodiranja se podaci sa kategoričke varijable koja ima g kategoorija prevode u g - 1 binarnih ("dami") varijabli.[11] Koja kategorija će biti izostavljena, tj. za koju kategoriju neće postojati odgovarajuća binarna varijabla zavisi od istraživača. Uobičajeno se jedna kategorija uzima kao referentna kategorija, tj. kategorija u odnosu na koju će se u statističkim analizama posmatrati preostale kategorije. Jedan od mogućih načina "dami" kodiranja podataka na varijabli "rukost" shematski se može predstaviti na sledeći način:

Rukost kodirana kao faktor	"Dami" varijabla 1 (Levoruki)	"Dami" varijabla 2 (Ambidekstri)
1	1	0
2	0	1
3	0	0
2	0	1
1	1	0
2	0	1

[11] Reč dàmi izgovara se sa kratkosilaznim akcentom na a, kao a u našoj reči *stani.*

Uočimo da je kategorija "desnoruki" u ovom načinu kodiranja predstavljena kodom 00 na dvema "dami" varijablama: prema tome ova kategorija predstavlja referentnu kategoriju. Ovaj način kodiranja kategoričkih podataka pogodan je onda kada se kategorička varijabla koristi u statističkom postupku koji se zove multipla regresiona analiza.

Kodiranje efekata (engl. effect coding) i *kontrastno kodiranje* (engl. contrast coding) predstavljaju način kodiranja kategoričkih podataka koji se često koristi onda kada se statističkim postupcima (na primer, analizom varijanse) ispituje postoji li efekat (dejstvo) kategoričke varijable kao nezavisne varijable na neku zavisnu varijablu. U ovim načinima kodiranja podaci na kategoričkoj varijabli kodiraju se tako da omogućuju međusobno poređenje pojedinih kategorija neke kategoričke varijable u pogledu neke kvantitativne varijable uz istovremeno isključivanje preostalih kategorija iz ovog poređenja, ili, pak, poređenje grupacija kategorija međusobno. Ove vrste kodiranja omogućuju, naprimer, da se levoruki i desnoruki ispitanici uzeti zajedno porede sa ambidekstrima u pogledu nekog kvantitativnog obeležja ili da se levoruki i desnoruki ispitanici porede u pogledu nekog kvantitativnog obeležja pri čemu su iz ovog poređenja (statističkim terminima rečeno *kontrasta*) isključuju ambidekstri. Budući da razumevanje odvijanja ovih načina kodiranja podrazumeva poznavanje statističkih postupaka (na primer analize varijanse i regresione analize) unutar kojih se najviše koriste, nećemo ih ovde dalje obrazlagati.[12]

1. Statistički opis uzorka u pogledu jedne kategoričke varijable

Statistički opis uzorka u pogledu jedne kategoričke varijable sastoji se uglavnom u utvrđivanju učestalosti i relativne učestalosti pojedinih kategorija. Ukoliko je uzorak probabilistički, relativna učestalost kao proporcija određene kategorije predstavlja ocenu verovatnoće te kategorije.

Od svih mera centralne tendencije koje smo opisali u Glavi 4, a koje se koriste za statistički opis uzorka u pogledu kvantitativne varijable, za kategoričke varijable nominalnog tipa smisaono je koristiti jedino mod. Mod u slučaju kategoričkih varijabli ukazuje na "tipičnu" ili najučestaliju kategoriju u uzorku. Dakle, mod je kod kategoričkih varijabli predstavljen modalnom kategorijom, tj. kategorijom sa najvećom učestalošću. Ukoliko je reč o kategoričkim varijablama sa uređenim kategorijama moguće je kao meru centralne tendencije pored moda koristiti i medijanu. Budući da su *kategorisane* varijable u osnovi kvantitativne varijable za računanje mera centralne tendencije na ovim varijablama treba koristiti nekategorisane podatke i primeniti postupke prikazane u Glavi 4.

Premda pojam varijabilnosti, u smislu odstupanja rezultata od njihove aritmetičke sredine, nema mnogo smisla koristiti kada je reč o pravim kategoričkim varijablama, smisaono je upotrebiti neki pokazatelj varijabilnosti u smislu raznovrsnosti ili raznolikosti (engl. *diversity* ili *unalikeability*) uzorka u pogledu kategoričke varijable. Pri prikazivanju standardne devijacije kao mere varijabilnosti pokazali smo da postoji bliska matematička veza između definicije standardne devijacije na osnovu odstupanja rezultata od njihove aritmetičke sredine i definicije ove mere na osnovu razlika tj. distanci između samih rezultata. Pojam raznolikosti kod kategoričkih varijabli zasnovan je na potonjem pristupu. U Glavi 2 istakli smo da je u skupu kategoričkih podataka, tj. podataka sa nominalnog "nivoa merenja" jedino smislena relacija relacija ekvivalencije, te da korišćenjem ovakvih podataka ima smisla jedino utvrditi da li su jedinice posmatranja iste ili različite u pogledu kategorijske pripadnosti. Stoga je kod kategoričkih varijabli pojam varijabilnosti smisaono definisati u smislu raznolikosti ili raznovrsnosti podataka (cf. Perry & Kader, 2005) .

[12] Podrobna objašnjenja različitih pristupa i postupaka kodiranja kategoričkih varijabli mogu se naći u tekstovima posvećenim analizi varijanse ili regresionoj analizi, na primer, u Serlin & Levin, 1985 i Cohen, Cohen, West, & Aiken, 2003, str. 302–353.

Za ilustraciju "varijabilnosti" u smislu raznolikosti kategoričkih podataka poslužićemo se vrlo jednostavnom ilustracijom (inspirisano idejama izloženim u Kader & Perry, 2007). Zamislimo da smo na dva uzorka od po 15 ispitanika dobili sledeće rezultate u pogledu "rukosti", tj. preferencije jedne ruke u obavljanju složenih motornih veština (1 = levoruk, 2 = ambidekstar, 3 = desnoruk):

Zamišljeni uzorak 1: **1 3 2 1 3 2 1 3 2 1 2 3 1 2 3**

Zamišljeni uzorak 2: **1 3 2 1 3 2 3 3 3 3 2 3 3 3 3**

Pre čitanja nastavka teksta predlažem čitaocu da pogleda svaki od ova dva niza podataka i da odluči za sebe koji od tih nizova mu se čini "šarenijim" ili raznolikijim?
Verujem da je ogromna većina čitalaca odabrala niz podataka za Zamišljeni uzorak 1 kao "šareniji", raznovrsniji ili raznolikiji. U svakom slučaju čitaoci kojima je vizuelna raznolikost veća za podatke iz Zamišljenog uzorka 1 imaju implicitnu definiciju vizuelne raznolikosti nalik statističkoj definiciji "varijabilnosti", ili, bolje rečeno, raznolikosti, tj. raznovrsnosti podataka na kategoričkoj varijabli.
Odakle potiče veća raznolikost skupa podataka za Zamišljeni uzorak 1? Rasporedi učestalosti pojedinih kategorija za ova dva uzorka prikazani su u sledećoj tabeli:

	Zamišljeni uzorak 1	Zamišljeni uzorak 2
Kategorija	f_k	f_k
1 (levoruk)	5	2
2 (ambidekstar)	5	3
3 (desnoruk)	5	10
Ukupno	15	15

Očigledno, veća raznolikost Zamišljenog uzorka 1 potiče od toga što su frekvencije za različite kategorije jednake, dok je manja raznolikost Zamišljenog uzorka 2 posledica toga što su učestalosti kategorija dramatično različite. Većina mera raznolikosti za kategoričke varijable zasnovana je na razlikama u učestalosti pojedinih kategorija: što su ove učestalosti sličnije to je veća raznolikost. Ukoliko se raznolikost posmatra u pogledu rasporeda relativnih frekvencija tada važi sledeće: što su relativne frekvencije iskazane proporcijama bliže recipročnoj vrednosti broja kategorija na varijabli to postoji veća raznolikost. (Budući da je za vizuelnu ilustraciju raznolikosti imalo smisla koristiti veoma mali broj rezultata u tabeli nisu prikazane relativne već samo obične frekvencije. Naime, relativne frekvencije na uzorcima manjim od 100, a pogotovu na uzorcima manjim od 50, nemaju mnogo smisla jer su tada male promene u frekvencijama praćene velikim promenama u relativnim frekvencijama).

Mere raznolikosti (raznovrsnosti) nominalne varijable

Od mnogih mera raznolikosti ili raznovrsnosti kategoričkih varijabli prikazaćemo indeks raznovrsnosti, indeks kvalitativne varijacije, koeficijent raznolikosti i entropiju. Ove mere se znatno češće koriste u sociologiji, ekologiji i ekonomiji nego u psihologiji.

Indeks raznovrsnosti (engl. Index of diversity)

Indeks raznovrsnosti, u oznaci D, definisan je na sledeći način:

$$D = 1 - \sum_{k=1}^{g} p_k^2$$

pri čemu je p_k relativna frekvencija ili proporcija kategorije k (k = 1, ..., g). Dakle, ovaj indeks se vrlo jednostavno računa: potrebno je sabrati kvadrirane relativne frekvencije iskazane kao proporcije u svim kategorijama kategoričke varijable i potom dobijeni zbir oduzeti od 1. Očigledno, minimalna moguća vrednost ovog indeksa je 0 i desiće se onda kada sve jedinice posmatranja pripadaju samo jednoj kategoriji, tj. kada je $p_r = 1$ a $p_s = 0$ za svako $s \neq r$. Maksimalna vrednost indeksa D zavisi od broja kategorija i data je sledećim izrazom:

$$D_{max} = 1 - \frac{1}{g} = \frac{g-1}{g}$$

.

Ovaj indeks će, za dati broj kategorija, imati maksimalnu vrednost onda kada je $p_k = 1/g$ za svako k jer je tada

$$D = 1 - \sum_{k=1}^{g} \left(\frac{1}{g}\right)^2 = 1 - g * \frac{1}{g^2} = 1 - \frac{1}{g} = D_{max}$$

Dakle, za dati broj kategorija, D će imati maksimalnu vrednost onda kada su učestalosti svih kategorija varijable jednake.

Pri tumačenju vrednosti ovog indeksa važno je imati na umu da se maksimalna vrednost indeksa D asimptotski bliži jedinici sa povećanjem broja kategorija. Tako, ako varijabla ima samo dve kategorije D može biti najviše 0.5, ako varijabla ima pet kategorija D može dostići vrednost 0.8, a za varijablu sa 10 kategorija maksimalna vrednost ovog indeksa može biti 0.9. Prema tome, ista vrednost indeksa D za varijablu sa manjim brojem kategorija ukazuje na veću raznolikost nego za varijablu sa više kategorija.

Vrednost indeksa D za dva zamišljena dovoljno velika uzorka na varijabli “rukost” sa istim relativnim učestalostima kao što je to slučaj u primeru kojim smo vizuelno ilustrovali pojam raznolikosti izračunali bismo na sledeći način:

Zamišljeni uzorak 1: $D = 1 - (0.333^2 + 0.333^2 + 0.333^2) = 0.67$

Zamišljeni uzorak 2: $D = 1 - (0.133^2 + 0.2^2 + 0.667^2) = 0.50$.

Dakle, veća raznovrsnost u pogledu “rukosti” postoji u Zamišljenom uzorku 1.

Nismo slučajno u prethodnim redovima isticali da je reč o zamišljenim dovoljno velikim uzorcima sa istim relativnim učestalostima kao što je to slučaj u vizuelnoj ilustraciji raznolikosti. Naime, za vizuelnu ilustraciji raznolikosti nije bilo svrsishodno koristiti veliki broj podataka. Međutim, indeks raznolikosti, kao i preostale statističke mere raznolikosti koje ćemo prikazati u ovoj glavi zasnivaju se na računanju relativnih frekvencija, tj. proporcija. Nije opravdano ni proporcije niti procente računati u situacijama kada raspolažemo sa malim ukupnim brojem rezultata. Smisleno korišćenje proporcija i procenata podrazumeva da je ukupan broj rezultata dovoljno veliki: najbolje je da ukupan broj rezultata bude veći od 100. Ukoliko procente ili proporcije računamo na malom broju rezultata dolazimo u apsurdnu situaciju da vrlo mala promena u frekvenciji dovodi do nesrazmerno velike promene u proporcijama, tj. procentima. Na primer, ako imamo ukupno 20 rezultata, promena sa učestalosti 5 na učestalost jednaku 7, koja je realno veoma mala, dovodi do promene u relativnoj učestalosti kao proporciji sa 0.25 na 0.35, odnosno do promene u relativnoj učestalosti u procentima sa 25% na 35%!

Ako se indeks raznovrsnosti računa korišćenjem relativnih frekvencija u procentima, koristi se sledeći oblik obrasca:

$$D = \frac{100^2 - \sum_{k=1}^{g} P_k^2}{100^2}$$

Kada se koristi kao deskriptivna statistička mera indeks raznovrsnosti se uobičajeno zaokružuje na dve decimale.

Indeks kvalitativne varijacije

Indeks kvalitativne varijacije (engl. Index of qualitative variation), u oznaci IQV, definisan je na sledeći način:

$$IQV = \frac{g}{g-1}\left(1 - \sum_{k=1}^{g} p_k^2\right) = \frac{g}{g-1} D$$

Indeks kvalitativne varijacije zapravo predstavlja količnik dobijenog i maksimalno mogućeg indeksa raznovrsnosti, tj. standardizovani indeks raznovrsnosti. Ovaj indeks, prema tome, za bilo koji broj kategorija dostiže maksimalnu vrednost jednaku 1 onda kada su učestalosti svih kategorija varijable jednake. Budući da je reč o meri koja je standardizovana s obzirom na broj kategorija na varijabli, indeks kvalitativne varijacije je jednostavniji za tumačenje od indeksa raznovrsnosti. Dobijena vrednost indeksa kvalitativne varijacije se može pomnožiti sa 100 i tumačiti kao procenat prisutne raznovrsnosti u uzorku od maksimalno moguće raznovrsnosti. Na primer, ako dobijeni indeks kvalitativne varijacije od 0.62 pomnožimo sa 100 dobijamo 62%. To bismo mogli tumačiti kao da je u uzorku prisutno 62% maksimalno moguće raznovrsnosti u pogledu date kategoričke varijable.

Međutim, neosetljivost na broj kategorija može istovremeno predstavljati i nedostatak ove mere jer bi veći broj kategorija prirodno trebalo da doprinosi raznovrsnosti. Na primer, za varijablu sa dve podjednako zastupljene kategorije (relativna frekvencija svake kategorije jednaka 0.5) IQR bi imao vrednost 1, dok bi za varijablu sa pet kategorija sa relativnim frekvencijama 0.2, 0.2, 0.2, 0.15 i 0.25 ovaj indeks iznosio 0.99 i ukazivao bi na nešto manju raznolikost u drugom slučaju, što je svakako problematično (cf. Agresti & Agresti, 1978). Stoga je najbolje upotrebu i tumačenje indeksa kvalitativne varijacije ograničiti na situacije kada se prikazuje raznovrsnost različitih grupa u pogledu relativne zastupljenosti istih kategorija.

Pri računanju indeksa kvalitativne varijacije treba uzeti u obzir teorijske kategorije, tj. kategorije sa kojima se krenulo u istraživanje. Na primer, ako se u istraživanje krenulo sa kategoričkom varijablom sa 5 kategorija ali je na konkretnom uzorku jedna kategorija ostala prazna, tj. nijedan ispitanik nije pao u tu kategoriju, pri računanju indeksa kvalitativne varijacije treba uzeti da je g = 5 a ne da je g = 4.

Za dva zamišljena dovoljno velika uzorka na varijabli "rukost" sa istim relativnim učestalostima kao što je to slučaj u primeru kojim smo vizuelno ilustrovali pojam raznolikosti indeks kvalitativne varijacije bismo izračunali na sledeći način:

Zamišljeni uzorak 1: IQR = (3/2) [1 - (0.333^2 + 0.333^2 + 0.333^2)] = 1

Zamišljeni uzorak 2: IQR = (3/2) [1 - (0.133^2 + 0.2^2 + 0.667^2)] = 0.75.

Dakle, veća raznovrsnost u pogledu rukosti postoji u Zamišljenom uzorku 1. U Zamišljenom uzorku 1 postoji maksimalno moguća raznovrsnost, dok raznovrsnost u Zamišljenom uzorku 2 jednaka 75% maskimalno moguće raznovrsnosti.

Ako se indeks kvalitativne varijacije računa na osnovu relativnih frekvencija u procentima koristi se sledeća varijanta obrasca:

$$IQV = \frac{g}{g-1}\left(\frac{100^2 - \sum_{k=1}^{g} P_k^2}{100^2}\right)$$

Kada se koristi kao deskriptivna statistička mera indeks kvalitativne varijacije uobičajeno se zaokružuje na dve decimale.

Entropijski indeks (engl. Information index ili Enthropy)

Entropijski indeks je statistička mera koja odgovara entropiji, meri neizvesnosti ili meri količine informacija izvedenoj u okviru matematičke teorije komunikacije, tj. terije informacije Kloda Šenona. U okviru ove teorije entropija predstavlja meru neizvesnosti ili nepredvidljivosti informacija sadržanih u nekoj poruci. Kakve veze ima neizvesnost iz teorije informacija i raznovrsnost kategoričke varijable? Za razumevanje ove veze korisno je da sebi postavimo sledeće pitanje: kada postoji veća neizvesnost, u situaciji kada treba da pogodimo da li je na slučaj uzeta osoba iz populacije desnoruka, ambidekstar ili levoruka ili u situaciji kada treba da pogodimo da li na slučaj uzeta osoba iz iste populacije ima dve noge? Verujem da bi većina ljudi lako pogodila da je nepredvidljivost ili neizvesnost u potonjem slučaju znatno manja budući da ogromna većina ljudi ima obe noge, znatno je manje onih sa jednom nogom, a veoma su retke osobe bez obeju nogu. Dakle, manja raznovrsnost populacije u pogledu neke kategoričke varijable zapravo znači da u tom slučaju kategorička varijabla nosi sobom manju količinu informacija, tj. manju neizvesnost.

Entropijski index (engl. enthropy), u oznaci H, definiše se na sledeći način:[13]

$$H = -\sum_{k=1}^{g} p_k \ln p_k$$

Kao što se iz obrasca može uočiti, radi dobijanja ove mere potrebno je za svaku kategoriju na varijabli izračunati proizvod relativne frekvencije, tj. proporcije za datu kategoriju i logaritma te relativne frekvencije. Potom se proizvodi za sve kategorije saberu i i tako dobijenom zbiru promeni predznak. Predznak minus ispred zbira je neophodan kako bi entropijski indeks imao pozitivan predznak. Naime, logaritmi proporcija, tj. brojeva između 0 i 1 su negativni brojevi. Prema tome, svi sabirci će takođe imati negativan predznak te će i ukupan zbir svih sabiraka biti negativan. [14]

U definiciji entropijskog indeksa moguće je koristiti, osim logaritma za osnovu e i logaritam za osnovu 10 ili logaritam za osnovu 2. Ipak, u statistici se uobičajeno pri računanju entropijskog indeksa koristi prirodni logaritam, tj. logaritam za osnovu e.

Minimalna vrednost entropijskog indeksa jednaka je nuli. To se može desiti samo u teorijskom slučaju, tj. onda kada bi sve jedinice posmatranja u uzorku bile u jednoj kategoriji kategoričke varijable. Tada je u jednoj kategoriji proporcija jednaka 1, dok su u svim ostalim kategorijama proporcije jednake nuli. Prema tome, H = -[1*ln(1)] = 0.

Budući da entropijski indeks predstavlja zbir proizvoda relativne frekvencije i logaritma relativne frekvencije za sve kategorije, maksimalna vrednost entropijskog indeksa zavisi od broja kategorija.

U sledećoj tabeli date su maksimalne vrednosti entropijskog indeksa koji se računa korišćenjem prirodnog logaritma za kategoričke varijable koje sadrže od 2 do 10 kategorija:

Broj kategorija	Maksimalno H

[13] Oznaka H predstavlja veliko grčko slovo eta budući da termin entropija potiče od grčkog entrope = okret ka unutra.

[14] Radi razumevanja matematičke pozadine entropijskog indeksa korisno je razmišljati na sledeći način (modifikovano prema Ivković, 1997, str. 68): Kada je uzorak slučajan, pripadnost entiteta iz uzorka određenoj kategoriji kategoričke varijable može se posmatrati kao slučajan događaj, a dešavanje bilo koje kategorije iz skupa svih mogućih kategorija varijable kao izvestan ili siguran događaj jer entitet mora pripadati nekoj od kategorija na varijabli, ako je varijabla valjano definisana. Ako je verovatnoća dešavanja neke kategorije, ocenjena proporcijom entiteta u toj kategoriji, jednaka p_k, tada se „neodređenost“ u vezi sa tom kategorijom može meriti nekom monotono opadajućom funkcijom verovatnoće te kategorije tako da što je verovatnoća kategorije veća „neodređenost“ bude manja. Na primer, takva funkcija može biti $-\ln p_k$. Verovatnoća pojavljivanja takve neodređenosti je onda p_k. Podelu kategoričke varijable na kategorije možemo posmatrati kao razbijanje izvesnog događaja na *k* događaja. Stoga $-\ln p_k$ možemo posmatrati kao moguće vrednosti slučajne varijable, a kao meru „neodređenosti“ na toj varijabli možemo uzeti očekivanu vrednost slučajne varijable. Očekivana vrednost slučajne varijable u tom slučaju zapravo predstavlja entropijski indeks H. Zaista, ako pogledamo obrazac za H, i ako $-\ln p_k$ posmatramo kao moguće vrednosti slučajne varijable, tada lako možemo uočiti istovetnost obrasca za H i obrasca za očekivanu vrednost iz poglavlja 3, tj. obrasca **.

2	0.69
3	1.10
4	1.39
5	1.61
6	1.79
7	1.95
8	2.08
9	2.20
10	2.30

Za fiksiran broj kategorija maksimalna vrednost entropijskog indeksa dobija se onda kada su relativne frekvencije u svim kategorijama jednake.

Entropijski indeks za zamišljene uzorke izračunali bismo na sledeći način:

Zamišljeni uzorak 1: H = - (0.333 * ln 0.333 + 0.333 * ln 0.333 + 0.333 * ln 0.333) = 1.10

Zamišljeni uzorak 2: IQR = -(0.133 * ln 0.133 + 0.2 * ln 0.2 + 0.667 * ln 0.667) = 0.86.

Radi donošenja zaključka o raznovrsnosti uzorka u pogledu kategoričke varijable možemo dobijenu vrednost entropijskog indeksa posmatrati u odnosu na maksimalno moguću vrednost ovog indeksa za dati broj kategorija. Dakle, prema entropijskom indeksu u Zamišljenom uzorku 1 postoji maksimalno moguća raznovrsnost, dok je raznovrsnost u Zamišljenom uzorku 2 jednaka 78% maksimalno moguće raznovrsnosti. Zaključak do kojeg smo došli na osnovu entropijskog indeksa sličan je onom do kojeg smo došli na osnovu indeksa kvalitativne varijacije.

Kada se koristi kao deskriptivna statistička mera entropijski indeks se uobičajeno zaokružuje na dve decimale.

Varijabilnost uređenih kategoričkih (ordinalnih) varijabli

Premda je u principu kao pokazatelje varijabilnosti na uređenim kategoričkim (ordinalnim) varijablama moguće koristiti one mere varijabilnosti koje smo prikazali u Glavi 4 a koje se zasnivaju na kvantilima (interkvartilni raspon i kvartilno odstupanje) u novije vreme razvijene su i statističke mere posebno namenjene opisu varijabilnosti uzorka u pogledu ovih varijabli. Prikazaćemo u ovom tekstu samo normiranu meru ordinalne koncentracije.[15]

Normirana mera ordinalne koncentracije (engl. Normed measure of ordinal concentration)

Normirana mera ordinalne koncentracije, u oznaci L^2, definisana je na sledeći način:[16]

$$L^2 = \frac{\sum_{k=1}^{g-1}\left(cp_k - \frac{1}{2}\right)^2}{\frac{g-1}{4}}$$

Pri tome, cp_k je kumulativna relativna frekvencija (iskazana kao proporcija) u kategoriji k a g je ukupan broj kategorija. Ova mera zapravo predstavlja količnik dobijene ordinalne koncentracije i maksimalno moguće ordinalne koncentracije za dati broj uređenih kategorija. Iz obrasca treba uočiti da se zbir u brojiocu računa za za sve kategorije osim poslednje, tj. izuzimanjem kategorije za koju je kumulativna relativna frekvencija jednaka jedinici. Izraz u brojiocu predstavlja dobijenu ordinalnu koncentraciju čije računanje počiva na postavci prema kojoj maksimalna varijabilnost na ordinalnoj varijabli sa g kategorija postoji onda kada su svi rezultati smešteni u prvoj (k = 1) i poslednjoj (k = g) kategoriji. (Uočimo da to nije slučaj kod kategoričkih varijabli nominalnog tipa kod kojih maksimalna raznovrsnost postoji onda kada su rezultati uniformno raspoređeni, tj. podjednako raspoređeni na sve

[15] Podrobniji prikaz ovih mera može se naći u Blair & Lacy, 2000.

[16] Premda je u označavanju ove mera u originalnom radu korišćeno malo slovo l, odlučili smo da u skladu sa prinicpom označavanja statističkih mera uzoraka primenjenim u ovoj knjizi koristimo u oznaci veliko slovo L.

kategorije).[17] Prema tome, kumulativne relativne frekvencije za prvih g - 1 kategorija kod ordinalne varijable koja ima maksimalnu varijabilnost jednake su 0.5, tj. 1/2. Na primer, ako na nekoj uređenoj kategoričkoj, tj. ordinalnoj varijabli sa četiri kategorije ima 20 rezultata, maksimalna varijabilnost ili polarizovanost na toj varijabli postoji onda kada su učestalosti ovih 20 rezultata smeštene po kategorijama redom na sledeći način: 10, 0, 0, 10. U tom slučaju će kumulativne relativne frekvencije za svaku od prve tri kategorije biti jednake1/2. Izraz u brojiocu normirane mere ordinalne koncentracije predstavlja zbir kvadriranih odstupanja relativnih kumulativnih frekvencija za prvih g - 1 kategorija od vrednosti 1/2, tj. od vrednosti kumulativnih relativnih frekvencija kada postoji maksimalna varijabilnost. Uočimo da je izraz u brojiocu jednak nuli kada su kumulativne relativne frekvencije u svih prvih g - 1 kategorija jednake 1/2. U tom slučaju će i vrednost mere L^2 biti jednaka nuli. Dakle, vrednost mere L^2 jednaka nuli odgovara maksimalnoj varijabilnosti ordinalne varijable. Izraz u brojiocu predstavlja zbir za prvih g - 1 kategorija. Zbir u brojiocu, dakle, zavisi od broja kategorija koje varijabla ima. Izraz u imeniocu predstavlja maksimalnu vrednost koju zbir u brojiocu može postići za dati broj kategorija, te taj izraz služi za normiranje mere koncentracije. Maksimalna vrednost zbira u brojiocu za dati broj kategorija – vrednost jednaka (g - 1) / 4 – postiže se onda kada su svi rezultati na varijabli koncentrisani u jednoj kategoriji. U tom slučaju će i vrednost normirane mere ordinalne koncentracije biti jednaka jedinici. Zato se ova mera zove merom koncentracije a ne merom varijabilnosti.

Naglasimo još jednom: pri tumačenju normirane mere ordinalne koncentracije treba voditi računa da vrednost ove mere jednaka nuli govori o minimalnoj koncentraciji, odnosno maksimalnoj varijabilnosti na varijabli, dok vrednost ove mere jednaka jedinici govori o maksimalnoj koncentraciji, odnosno minimalnoj varijabilnosti. To može lako da zbuni. Radi jednostavnijeg tumačenja, vrednost ove mere može se oduzeti od 1: ukoliko je vrednost izraza $1 - L^2$ jednaka nuli varijabilnost na ordinalnoj varijabli je minimalna, a ukoliko je vrednost ovog izraza jednaka jedinici varijabilnost na varijabli je maksimalna.

Računanje normirane mere ordinalne koncentracije pokazaćemo na primeru podataka o obrazovanju oca i majke za jedan uzorak studenata Univerziteta u Beogradu:[18]

	Otac			Majka		
Obrazovanje	f_k	cf_k	cp_k	f_k	cf_k	cp_k
6 (postdiplomsko)	55	450	1	48	450	1
5 (visoko)	161	395	.8778	140	402	.8933
4 (više)	21	234	.5200	37	262	.5822
3 (srednje)	202	213	.4733	209	225	.5000
2 (osmogodišnje)	11	11	.0244	15	16	.0356
1 (nepotpuno osmogodišnje)	0	0	.0000	1	1	.0022
Ukupno	450			450		

(Kumulativne relativne frekvencije u koloni cp_k prikazali smo u tabeli na više decimala nego što je to uobičajeno jer ih koristimo za računanje L^2).

Za obrazovanje oca:

$$L^2 = \frac{(0-0.5)^2 + (0.0244-0.5)^2 + (0.4733-0.5)^2 + (0.52-0.5)^2 + (0.878-0.5)^2}{\frac{6-1}{4}} = 0.50$$

Za obrazovanje majke:

[17] Upravo se ovaj argument često koristi u objašnjenju neadekvatnosti korišćenja mera raznolikosti nominalnih varijabli kao mera varijabilnosti za uređene kategoričke varijable (cf. Blair & Lacy, 2000).

[18] Uzorak je iz istraživanja koje je sprovedeno 2014. i 2015. godine.

$$L^2 = \frac{(0.0022-0.5)^2 + (0.0356-0.5)^2 + (0.5-0.5)^2 + (0.5822-0.5)^2 + (0.8933-0.5)^2}{\frac{6-1}{4}}$$

$$= 0.50$$

Kada dobijene mere oduzmemo od 1, dobijamo:
Obrazovanje oca: $1 - L^2 = 0.5$
Obrazovanje majke: $1 - L^2 = 0.5$
Dakle, i u pogledu obrazovanja oca i u pogledu obrazovanja majke varijabilnost ovog uzorka je osrednja i svakako manje izražena nego što je varijabilnost obrazovanja u opštoj populaciji.

Kada se koristi kao deskriptivna statistička mera, L^2 se uobičajeno zaokružuje na dve decimale.

*****

Statističke mere raznovrsnosti nominalnih podataka i varijabilnosti ordinalnih podataka ima smisla uglavnom koristiti u situacijama kada nominalne ili uređene kategoričke varijable imaju veći broj kategorija (više od 3) jer tada nije uvek jednostavno samo pregledom rasporeda frekvencija prosuđivati o raznolikosti ili varijabilnosti podataka. Prema tome, nema mnogo opravdanja koristiti ove mere za varijable sa malim brojem kategorija. Pri demonstriranju računanja pojedinih od ovih mera mi smo koristili veoma jednostavne primere, tj. varijable sa malim brojem kategorija iz praktičnih razloga. Korist od korišćenja ovih mera može biti posebno dragocena u situacijama kada varijabla ima veći broj kategorija i kada u istraživanju imamo veći broj uzoraka. U tim situacijama možemo korišćenjem ovih mera lakše steći predstavu o sličnostima ili razlikama među uzorcima u pogledu raznovrsnosti ili varijabilnosti.

Podaci koji nedostaju

Nedostajanje podataka na kategoričkoj varijabli predstavlja mnogo složeniji problem nego kada podaci nedostaju na kvantitativnoj varijabli. Naime, podatke koji za neku jedinicu posmatranja nedostaju na kvantitativnim varijablama često je moguće nadomestiti ocenjenim vrednostima koje se dobijaju na osnovu specijalizovanih algoritama koji koriste informacije sadržane u preostalim podacima za datu jedinicu posmatranja. Takav postupak praktično nije moguć kada je reč o podacima na kategoričkim varijablama. U slučaju nedostajanja podataka na kategoričkoj varijabli najsmisaonije je uvesti zasebnu kategoriju za podatke koji nedostaju i tu kategoriju nazvati “bez podatka”.

2. Grafički prikaz podataka na jednoj kategoričkoj varijabli (atributivnom obeležju)

Za grafičko predstavljanje kategoričke ili nominalne varijable koriste se najčešće štapićasti dijagram (engl. barchart) i pitasti dijagram (engl. piechart).

Štapićasti dijagram

Ključna informacija koju grafički prikaz podataka na kategoričkoj varijabli treba da prenese jeste učestalost (apsolutna ili relativna) pojedinih kategorija. Štapićasti dijagram je veoma efikasan način za grafičko prikazivanje kvantitativnih informacija (frekvencija ili relativnih frekvencija) po kategorijama kategoričke varijable jer u sebi spaja dve vrste vizuelnih atributa koje se od mogućih vizuelnih atributa pokazuju kao najbolji načini za prenos kvantitativne informacije: dvodimenzionalnu lokaciju i dužinu linije.

Štapićasti dijagram, kako mu i samo ime nagoveštava, sastoji se od štapića, tj. pravougaonika čija dužina odgovara učestalosti određene kategorije. Za razliku od histograma na kojem su stubići, tj. pravougaonici postavljeni bez razmaka, štapićasti dijagram sadrži pravougaonike koji su uobičajeno razmaknuti. Pored toga, dok kod histograma stubići počinju od prave matematičke X-ose, vertikalno su postavljeni i ne mogu proizvoljno menjati mesta, pravougaonici ili štapići na štapićastom dijagramu postavljeni su arbitrarno, tj. mogu proizvoljno menjati mesta i mogu biti čak svi postavljeni

horizontalno. Dakle, linija od koje počinju pravougaoni oblici na štapićastom dijagramu ne predstavlja pravu matematičku osu.

Na primer, na osnovu sledeće tabele učestalosti

Primer: Grafički prikaz "rukosti" (preferencije jedne ruke u obavljanju složenih motornih veština) štapićastim dijagramom (U Editoru grafike SPSS-a odabiranjem opcije **Bar label style / Framed** u meniju **Attributes** postignuto je to da na stubićima piše procenat slučajeva koji pripada datoj kategoriji)

Pri korišćenju štapićastog dijagrama treba se pridržavati sledećih pravila:

- Štapićasti dijagram treba koristiti za predstavljanje kategoričke varijable sa većim brojem kategorija. Da bi ovaj grafički prikaz imao ikakvog smisla potrebno je da varijabla ima najmanje tri kategorije: nema nikakvog smisla grafički predstavljati dihotomnu kategoričku varijablu budući da se iz dva broja kojima su u tom slučaju predstavljene apsolutne ili relativne učestalosti može lako sve videti;
- Štapići bi trebalo da počinju od nulte tačke kako bi njihova dužina dala ispravnu predstavu o frekvenciji ili relativnoj frekvenciji kategorija;
- Štapiće je bolje postaviti horizontalno nego vertikalno. To je posebno bitno kada su na samom grafiku dati nazivi kategorija, što je inače preporučljivo: nazivi kategorija lakše se čitaju kada su postavljeni sa leve strane štapića nego ispod njega;
- Ukoliko je važno da posmatrač može lako sa grafika da uoči precizne vrednosti zastupljenosti pojedinih kategorija može se koristiti mreža sastavljena od razmaknutih paralelnih vertikalnih linija u pozadini štapića (ako su štapići horizontalno postavljeni), pri čemu ta mreža ne sme biti suviše gusta i perceptivno upadljiva u odnosu na štapiće. Dakle, štapići treba da budu perceptivna figura, a mreža perceptivna pozadina. Druga mogućnost je da se precizne brojke o zastupljenosti kategorija postave na štapiće (na primer, procenti) i tada nikakva mreža nije potrebna;
- Treba izbegavati trodimenzionalne štapiće jer se dodavanjem treće dimenzije nepotrebno vizuelno usložnjava grafik, a da se time ne prenosi nikakva informacija koja već nije sadržana u dvodimenzionalnim štapićima.

Pitasti dijagram

Pitasti ili kružni dijagram predstavlja podelu kružne površine na odsečke kojima su uobičajeno predstavljene relativne frekvencije pojedinih kategorija.[19] Koliki odsečak kružne površine će pripasti određenoj kategoriji određeno je veličinom ugla koji grade prave linije koje zajedno sa lukom kruga čine dati odsečak. Budući da ceo krug ima 360 stepeni, ugao (izražen u stepenima) koji grade prave linije koje na površini kruga ograničavaju odsečak za kategoriju *k*, u oznaci α_k, računa se prema sledećem obrascu:

$$\alpha_k = \frac{f_k}{n} * 360°$$

Pri tome, f_k je frekvencija kategorije *k*, a n je zbir frekvencija za sve kategorije na pitastom dijagramu. Iz prikazanog obrasca lako je uočiti da se ugao koji grade prave linije koje na površini kruga

[19] Ponekada se ovaj dijagram duhovito zove i „burek" (cf. Todorović, 2008, str. 207).

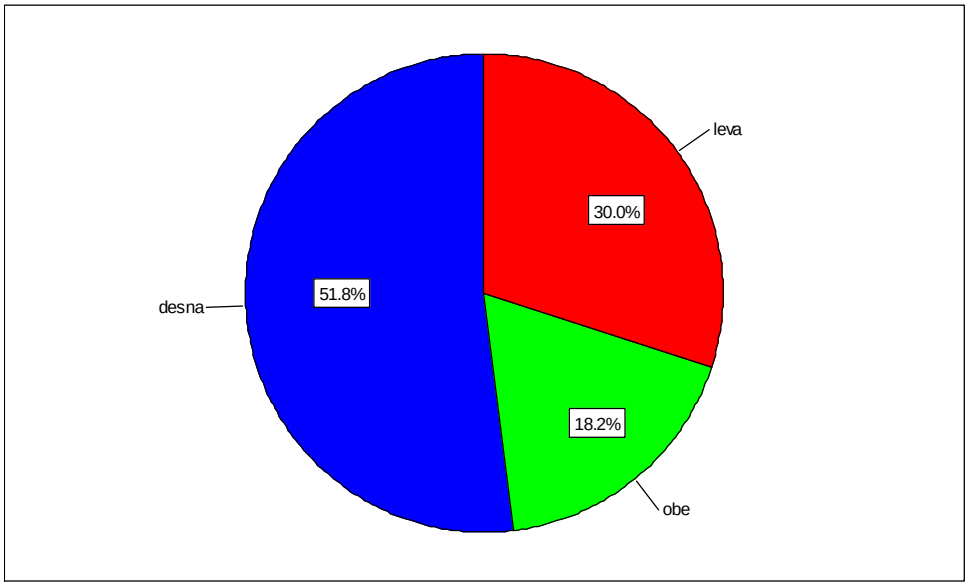
ograničavaju odsečak za kategoriju *k* može jednostavno izračunati i množenjem relativne frekvencije kategorije izražene u procentima sa 3.6.

Pitasti dijagram ima najviše smisla koristiti za varijable sa relativno malim brojem kategorija (od 3 do 6). Prema tome, pitasti dijagram nije uputno koristiti kada kategorička varijabla ima veliki broj kategorija (na primer, više od 7). Baš kao ni štapićasti dijagram, ni pitasti dijagram nema nikakvog opravdanja koristiti za prikazivanje dihotomnih varijabli.

Pri korišćenju pitastog dijagrama treba izbegavati:

- pseudotrodimenzionalne prikaze (odabirom 3-D opcije u statističkim paketima može se dodati lažna treća dimenzija koja ne nosi nikakvu informaciju a koja izvitoperuje grafik i usložnjava percepciju);
- perceptivno teško distinktivna senčenja i šrafure odsečaka;
- korišćenje legende (bolje je kategorije direktno označiti na samim odsečcima ili oznake direktno povezati linijama sa odgovarajućim odsečcima, kao i izdvajanje odsečaka iz kruga (korišćenjem tzv. “rasprsnutog” pitastog dijagrama).

Primer: Grafički prikaz atributivnog obeležja rukost, tj. preferencije jedne ruke u obavljanju složenih motornih veština pitastim dijagramom. (U Editoru grafike SPSS-a odabiranjem opcije **Options** u meniju **Charts**, uključivanjem **Percents** u okviru **Labels** i potom klikom na **Formats** i odabirom u polju **Position** varijante **Numbers inside text outside** postignuto je to da na "kriškama" pitastog dijagrama piše procenat slučajeva koji pripada datoj kategoriji).



****

***

Dugo se među istraživačima koji se bave percepcijom grafika vodila polemika o tome koji je grafički prikaz – štapićasti ili pitasti dijagram bolji i jednostavniji za izvlačenje informacija čijem prikazivanju su ovi grafici namenjeni. Ishodi ovih polemika bili su različiti zavisno od metodološko-teorijskih polazišta sprovedenih istraživanja. Ranija istraživanja su na osnovu psihofizičkih nalaza o tačnijem procenjivanju dužine nego površine prednost davala štapićastom dijagramu. Međutim, ključna informacija koju treba prikazati štapićastim ili pitastim dijagramom nije toliko precizna informacija o preciznoj veličini (apsolutnoj ili relativnoj frekvenciji) pojedinih kategorija (za te svrhe su najbolje statističke tabele) koliko informacija o relativnim veličinama (koje su kategorije zastupljenije a koje manje zastupljene, da li je zastupljenost određenog podskupa kategorija veća nego zastupljenost drugog podskupa kategorija i slično). Pojedina novija istraživanja u kojima je eksperimentalno direktno ispitivana brzina i tačnost izvlačenja relevantnih informacija iz ove dve vrste grafika daju izvesnu prednost štapićastom dijagramu (cf. Simkin & Hastie, 1987) dok su druga na

strani pitastog dijagrama samo za specifične situacijeu kojima je potrebno vršiti složenija poređenja (na primer, kada je potrebno porediti čitave skupove kategorija po zastupljenosti, cf. Spence & Lewandowski, 1991). Dakle, u većini situacija u kojima je uloga grafika da prikaže relativnu zastupljenost pojedinih kategorija praktično je svejedno koji od ova dva grafička prikaza ćemo koristiti u prikazivanju podataka na kategoričkoj varijabli. U određenim situacijama štapićastom dijagramu treba dati prednost: kada postoji veliki broj kategorija, kada su relativne frekvencije po kategorijama relativno ujednačene, Ukoliko, pak, grafikom želimo da kontrastiramo zastupljenost određenih skupova kategorija (npr. obrazovanje niže od srednjeg naspram srednjeg i višeg ili visokog obrazovanja) onda bismo prednost u grafičkom prikazivanju podataka mogli dati pitastom dijagramu.

Agresti, A., & Agresti, B. F. (1978). Statistical Analysis of Qualitative Variation. Sociological Methodology, *9*, 204–237.

Blair, J., & Lacy, M. G. (2000). Statistics of ordinal variation. *Sociological methods and research*, *28*(3), 251–280.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences, Third edition.* London: Lawrence Erlbaum Associates, Inc.

Fausto-Sterling, A. (2000). The five sexes, revisited. *The Sciences, July/August*, 19–23.

Ivković, Z. (1989). *Teorija verovatnoća sa matematičkom statistikom, IV izdanje*. Beograd: Naučna knjiga.

Kader, G. D., & Perry, M. (2007). Variability for Categorical Variables. *Journal of Statistics Education*, *15*. http://www.amstat.org/publications/jse/v15n2/kader.html (skinuto 19. 11. 2011.)

Perry, M., & Kader, G. D. (2005). Variation as Unalikeability. *Teaching statistics*, *27*(2), 58–60.

Serlin, R. C., & Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of educational statistics*, *10*(3), 223–238.

Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, *82*(398), 454–465.

Spence, I., & Lewandowski, S. (1991). Displaying propoortions and percentages. *Applied Cognitive Psychology*, *5*, 61–77.