IV. STATISTIČKI OPIS UZORKA U POGLEDU JEDNE KVANTITATIVNE VARIJABLE I GRAFIČKO PRIKAZIVANJE PODATAKA NA JEDNOJ KVANTITATIVNOJ VARIJABLI

Neophodni matematički pojmovi za razumevanje teksta u ovoj glavi:[1]

Osnovni pojmovi teorije verovatnoće

Logaritamska funkcija

Operator sabiranja (sumacioni operator) Σ

Operator dvostrukog sabiranja (dvojni sumacioni operator) $\Sigma\Sigma$

Operator proizvoda Π

Razmatranje postupaka statističke analize podataka počećemo pretpostavljajući da želimo da sredimo podatke i statistički opišemo uzorak na osnovu podataka dobijenih na jednoj kvantitativnoj varijabli.[2] Veoma su retke situacije u kojima ćemo za dati uzorak raspolagati podacima samo na jednoj varijabli. To se, zapravo, neće desiti nikada u realnim istraživanjima. (Dakle, čitaocu je ova glava jedinstvena prilika da uživa u ovoj najjednostavnijoj mogućoj situaciji). Pri primeni statistike uvek ćemo biti u situaciji u kojoj imamo prikupljene podatke na jednom ili više (najbolje slučajnih) uzoraka u pogledu većeg broja ispitivanih obeležja. Međutim, pre nego što podatke za veći broj obeležja analiziramo nekim od statističkih postupaka koji spadaju u tzv. statistiku zaključivanja (pri čemu uobičajeno koristimo više od jedne varijable istovremeno) potrebno je pažljivo srediti i dobro upoznati podatke na svakoj pojedinačnoj varijabli. Moglo bi se reći da se to podrazumeva i da to nije potrebno posebno naglašavati. Međutim, mi to ističemo zato što ovoj fazi statističke analize podataka istraživači u psihologiji i srodnim oblastima često ne posvećuju dovoljnu pažnju.

Ponavljamo: pre bilo kakve ozbiljnije statističke analize podataka potrebno je pažljivo razglédati podatke. I to baš ovako kako je napisano: razglédati! Dakle, sa nesvršenim oblikom ovog glagola, tj. sa dugouzlaznim naglaskom na slovu e baš kao što je to slučaj sa prvim e u reči dete![3] Pažljivo razgledanje podataka koje želimo statistički da analiziramo jedan je od ključnih preduslova valjane primene statističkih postupaka u analizi podataka. Međutim, svrhovito razgledanje podataka (pogotovu ako je takvih podataka mnogo) nemoguće je dok se podaci ne srede i ne prikažu grafički tako da se

[1] Osnovni pojmovi teorije verovatnoće prikazani su u Glavi 3, a ostale neophodne pojmove čitalac kojem je to potrebno može pronaći pod odrednicama **Operator sabiranja (sumacioni operator)**, **Operator dvostrukog sabiranja (dvojni sumacioni operator)**, **Operator proizvoda** i **Logaritamska funkcija** u Matematičkom pojmovniku u Dodatku **

[2] Postupke koje ćemo opisati u ovoj glavi možemo primeniti i onda kada statistički opisujemo populaciju ukoliko raspolažemo podacima na jednoj kvantitativnoj varijabli za sve članove populacije. Međutim, to je samo teorijska mogućnost, jer pri primeni statistike u psihologiji praktično nikada nemamo podatke za sve članove populacije.

[3] Jedan od najvećih statističara i veliki zastupnik tzv. eksploratorne analize podataka Džon Tjuki, u svom govoru na skupu Američke psihološke asocijacije 1968. godine, ističe da "izgleda da zaista nema zamene za 'razgledanje podataka' ("There really seems to be no substitute for "looking at the data.", Tukey, 1969, str. 83).

njihovo razgledanje učini upotrebljivim. Smisleno razgledanje sređenih i grafički prikazanih podataka treba da nas "zbliži" ili "sprijatelji" sa podacima. Onda kada se "zbližimo" sa podacima koje treba statistički analizirati, statistička analiza podataka postaje nalik učešću u rešavanju životnih problema bliske osobe: može da bude na trenutke zapetljano i teško ali nećemo praviti fatalne propuste jer najčešće imamo, zahvaljujući bliskosti, dobar osećaj šta i kako da uradimo.

Pri sređivanju podataka i statističkom opisivanju uzorka u pogledu jedne kvantitativne varijable možemo da primenimo i određene postupke tzv. eksploratorne analize podataka. Eksploratorna analiza podataka (engl. **E**xploratory **D**ata **A**nalysis – EDA) obuhvata brojne postupke koji su začeti u radovima poznatog statističara Džona Tjukija i njegovih saradnika sedamdesetih godina XX veka (cf. Hoaglin, Mosteller, & Tukey, 1983) i koji se još uvek razvijaju. Ovi postupci namenjeni su prevashodno otkrivanju pravilnosti i složajeva u podacima, kao i neočekivanih odstupanja podataka od pretpostavljenih teorijskih modela. Radi se, dakle, o postupcima koji olakšavaju upoznavanje sa osnovnim karakteristikama skupa podataka i omogućuju fleksibilno otkrivanje struktura u podacima. Osnovna karakteristika ovih postupaka je velika fleksibilnost i nezahtevnost u pogledu teorijskih pretpostavki za njihovu primenu. Ovi postupci nisu ograničeni samo na situacije kada posmatramo podatke na jednoj varijabli već se mogu primenjivati i za podatke sa više varijabli istovremeno. Rezultati eksploratorne analize podataka mogu biti od velike pomoći u izboru odgovarajućeg statističkog modela (i posledično konkretnog statističkog postupka) za finalnu analizu podataka – analizu podataka koja treba da ponudi empirijske argumente za odgovor na pitanje zbog kojeg su podaci i prikupljeni.

U nastavku teksta u ovoj glavi prikazaćemo postupke koje možemo koristiti za sređivanje podataka, podrobnije upoznavanje sa podacima, statistički opis uzorka i grafičko prikazivanje podataka kada imamo u vidu samo jednu kvantitativnu varijablu. Odgovarajuće postupke za podatke dobijene na kategoričkim varijablama prikazaćemo u sledećoj glavi. Postupke koje ćemo prikazati u ovoj i narednoj glavi čitalac ove knjige može primeniti ponaosob na svaku varijablu odgovarajućeg tipa pre nego što primeni neku od statističkih procedura koje uzimaju u obzir više varijabli istovremeno.

Struktura podataka

Kao što smo to u prethodnoj glavi objasnili, prikupljene podatke njihovim unošenjem u odgovarajućem programu organizujemo u tabelu ili matricu podataka. Često je potrebno na osnovu unetih podataka, pomoću odgovarajućih komandi u programu koji koristimo za rad sa podacima, izvesti vrednosti na varijablama koje nas zanimaju u daljoj analizi i tumačenju rezultata. (U programu SPSS najveći broj operacija u preuređivanju podataka izvodi se korišćenjem komandi koje se nalaze u opciji menija TRANSFORM. To su, pre svega, komande COMPUTE, COUNT, RECODE i AUTOMATIC RECODE).[4] Na primer, ako smo u matricu podataka uneli odgovore ispitanika na pojedinačna pitanja (ili, kako se to uobičajeno zove stavke) upitnika za merenje depresivnosti CES-D tako što smo odgovore kodirali ciframa od 1 do 4 tada je potrebno u skladu sa "uputstvom za

[4] Unos i organizaciju podataka, kao i korišćenje ovih komandi u programu SPSS čitalac može naučiti sledeći video instrukcije br.1 i br.2....**

skorovanje", tj. uputstvom za izračunavanje ukupnog rezultata na upitniku izračunati ukupni rezultat (skor) na ovom upitniku kao meru na varijabli *depresivnost*.[5] Isto tako, odgovore ispitanika na testu znanja biologije u kojem je na svako pitanje ponuđeno pet odgovora možemo uneti u obliku cifara od 1 do 5, a potom ih odgovarajućim komandama bodovati u pogledu tačnosti (tačan odgovor = 1, pogrešan odgovor = 0) i izvesti ukupni rezultat. Zbir jedinica ili broj tačnih odgovora u tom slučaju možemo tretirati kao meru na kvantitativnoj varijabli *uspeh na testu znanja biologije*. U nekim slučajevima brojčane vrednosti koje smo izmerili i uneli u odgovarajućem računarskom programu predstavljaju u svom izvornom obliku mere na određenoj kvantitativnoj varijabli. Na primer, uneti podaci o visini ispitanika u santimetrima mogu predstavljati mere na kvantitativnoj varijabli *visina*. Cifre koju smo na osnovu odgovora ispitanika na pitanje o ukupnom broju članova njihovih porodica uneli mogu bez ikakvih daljih izvođenja predstavljati vrednosti na varijabli *veličina porodice*.

U svakom slučaju, podaci za jednu kvantitativnu varijablu (npr., depresivnost, uspeh na testu znanja biologije, visina , veličina porodice) predstavljaju jednu kolonu u sastavu celokupne matrice podataka. Ova kolona matematički predstavlja vektor ili kolona-matricu reda nx1 ("n puta 1") budući da sadrži n redova i jednu kolonu.[6] U opštem slučaju podaci na jednoj kvantitativnoj varijabli v mogu biti rangovi ili mere (najčešće mere intervalnog, racio ili apsolutnog tipa). Zavisno od toga o kojoj od ovih vrsta podataka je reč, za statistički opis uzorka možemo koristiti odgovarajuće statističke mere uzorka (statistike) koji će biti prikazani u ovoj glavi. Budući da se mere na kvantitativnim varijablama najčešće u psihologiji i srodnim oblastima tretiraju kao da su apsolutnog, intervalnog ili racio tipa prevashodno ćemo kroz primere prikazati postupke koji su primenljivi na takve podatke.

Rezultat, meru ili skor na kvantitativnoj varijabli za bilo koju jedinicu posmatranja označavaćemo oznakom x_i. U oznaci x_i, *i* je indeks koji može uzeti celobrojne vrednosti od 1 do n, pri čemu je n veličina uzorka ili broj jedinica posmatranja za koje smo ispitivanjem prikupili podatke na varijabli v. Dakle x_1 je rezultat jedinice poosmatranja e_1, x_2 je rezultat za jedinicu posmatranja e_2 i tako redom sve do x_n što predstavlja rezultat za jedinicu posmatranja e_n.

Podaci koji su uneti u matricu podataka, kao i vrednosti na varijablama koje su korišćenjem komandi za rad sa podacima izvedene na osnovu unetih podataka predstavljaju "sirove", tj. nesređene podatke. Ove podatke zovemo nesređenim podacima jer nisu ni na koji način uređeni za izvlačenje statističkih informacija. Na primer, ukoliko je "sirovih" podataka na samo jednoj kvantitativnoj varijabli veliki broj, njihovim pregledanjem se ne može lako ustanoviti čak ni najveći ni najmanji rezultat a kamoli struktura podataka na varijabli. U vreme kada su se statistička proračunavanja izvodila bez pomoći računara na osnovu ovih "sirovih" podataka najčešće nije bilo moguće izračunavati statističke mere. Zahvaljujući računarima u današnje vreme je moguće izračunati i najsloženije statističke mere iz tzv. sirovih podataka. Međutim, i u vreme kada se statistička računanja izvode pomoću računara sirove podatke je potrebno srediti radi

[5] CES-D (Center for Epidemiologic Studies Depression scale – CES-D) je jedan od najpoznatijih upitnika za ispitivanje depresivnosti u opštoj populaciji koji se veoma često koristi u tzv. epidemiološkim istraživanjima. Upitnik sadrži 20 stavki (ili ajtema) na koje ispitanik odgovara biranjem jednog od 4 ponuđena odgovora. Odgovor na svaku stavku boduje se dodeljivanjem od 0 do 3 boda, a ukupni rezultat, kao mera depresivnosti, dobija se sabiranjem bodova za sve stavke.

[6] Vektor se inače u opštem slučaju matematički drugačije definiše ali u ovom trenutku takva definicija za naše potrebe nije neophodna. Mi ćemo u ovom tekstu vektorima jednostavno nazivati posebne vrste matrica, matrice sa jednom kolonom ili sa jednim redom, ne upuštajući se u formalno matematičko određenja pojma vektora.

prikazivanja podataka i izvlačenja statističkih informacija iz tih podataka. Na primer, osnovnu strukturu podataka na varijabli potrebno prikazati u obliku tzv. distribucije učestalosti koja se sastoji od kolone vrednosti (ili intervala vrednosti) na varijabli i kolone sa učestalostima.

Sređivanje podataka i statistički opis uzorka u pogledu jedne kvantitativne varijable na kojoj su rezultati mere treba da bude urađen tako da pruži pregledne i jasne informacije o rasponu u kojem se rezultati kreću, o učestalosti pojedinih vrednosti ili intervala vrednosti, o vrednosti oko koje se grupiše većina rezultata, o tome koliko se vrednosti međusobno razlikuju i o obliku raspodele vrednosti. Prema tome, sređivanje i statistički opis uzorka u pogledu jedne kvantitativne varijable podrazumeva **uređivanje** podataka, određivanje **najmanjeg** i **najvećeg rezultata**, pravljenje **raspodele učestalosti** (**distribucije frekvencija**), računanje odgovarajućih statističkih mera **lokacije** (onda kada je to opravdano mera **centralne tendencije**), mera **skale** ili mera **varijabilnosti** (**raspršenja** ili **disperzije**) i mera **oblika distribucije**. Najmanji ili najveći rezultat, mere lokacije/centralne tendencije, mere skale/varijabilnosti i mere oblika distribucije koje računamo na osnovu podataka dobijenih na uzorku spadaju u statističke mere koje zovemo **statistici**.

Da bismo prikupljene podatke upotrebili za statistički opis uzorka u pogledu jedne kvantitativne varijable potrebno je podatke srediti na određeni način. To ponekad podrazumeva **sortiranje**, tj. **uređivanje podataka** u rastući ili opadajući redosled, **rangovanje** podataka koji nisu dati u obliku rangova a najčešće formiranje jedinične ili **grupisane** raspodele učestalosti. Organizovanje podataka, njihovo uređivanje, rangovanje (ako je to potrebno) i formiranje raspodele učestalosti uobičajeno se izvode u statističkim programima za analizu podataka. Bez obzira na to što će neke od ovih postupaka statistički programi obično izvesti automatski kada im se zada oodgovarajuća komanda potrebno je razumeti osnovne principe ovih postupaka. Naime, razumevanje osnovnih principa ovih postupaka preduslov je za biranje odgovarajućih opcija pri zadavanju komandi i razumevanje ispisa koji se dobija izvršavanjem ovih komandi.

1. Uređivanje podataka po veličini – sortiranje podataka

Jedan od postupaka koji se veoma često primenjuje u početnim fazama analize podataka jeste uređivanje podataka na nekoj kvantitativnoj varijabli u rastući ili opadajući niz. Operacijom sortiranja redovi u matrici podataka koji predstavljaju jedinice posmatranja premeštaju se tako da odgovaraju rastućem ili opadajućem nizu rezultata na varijabli po kojoj se vrši sortiranje. Sortiranje je operacija koja je u statističkoj analizi podataka inače neophodna za određivanje tzv. **redoslednih statistika** (engl. order statistics).

Redosledni statistici uzorka *E* koji sadrži n jedinica posmatranja sa rezultatima x_1, x_2, ..., x_n na varijabli v, jesu $X_{(1)}$, $X_{(2)}$, ...$X_{(n)}$, pri čemu je $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ (Rosenberger & Gasko, 1983). Dakle, rezultati x_1, x_2, ..., x_n uređeni po veličini, od najmanjeg do najvećeg, postaju realizovane vrednosti redoslednih statistika $X_{(1)}$, $X_{(2)}$, ...$X_{(n)}$. Treba uočiti da, na primer, početni rezultat x_1 može postati vrednost redoslednog statistika $X_{(5)}$ ako je, posle uređivanja podataka taj rezultat peti po veličini. Za *r*-ti redosledni statistik rezultata x_i koristićemo oznaku $X_{(r)}$, pri čemu r predstavlja redni broj mesta koje u podacima uređenim od najmanjeg do najvećeg zauzima rezultat x_i.

Utvrđivanje najmanjeg i najvećeg rezultata

Na osnovu sortiranih podataka moguće je videti koji je rezultat najniži, u oznaci x_{min}, a koji najviši, u oznaci x_{max}:

$$\boxed{x_{min} = \min_i x_i, \quad i = 1, \ldots, n}$$

$$\boxed{x_{max} = \max_i x_i, \quad i = 1, \ldots, n}$$

pri čemu je x_i rezultat jedinice posmatranja e_i na kvantitativnoj varijabli v. Oznake max i min sa indeksom *i* ispod ovih oznaka označavaju da se radi o rezultatima koji su najveći, odnosno najmanji od svih rezultata, tj. među svim rezultatima označenim opštom oznakom x_i .

Najniži i najviši rezultat mogu se definisati i na osnovu redoslednih statistika:

$$\boxed{x_{min} = X_{(1)}}$$

$$\boxed{x_{max} = X_{(n)}}$$

U primeru koji ćemo dati u narednoj tabeli u nastavku teksta (videti tabelu u odeljku 2. Rangovanje rezultata), $x_{min} = 4$ a $x_{max} = 10$.

2. Rangovanje rezultata

Za računanje određenih statističkih mera i za određene statističke analize podataka potrebno je mere sa intervalnog ili racio nivoa pretvoriti u rangove, tj. rangovati. Rangovanje rezultata je u najjednostavnijem slučaju – kada su svi rezultati koji se ranguju međusobno različiti – prevođenje (preslikavanje) rezultata u skup celih brojeva od 1 do n, pri čemu je n broj rezultata. Ukoliko su neki rezultati međusobno jednaki onda niz rangova sadrži i decimalne brojeve što ćemo pojasniti u nastavku teksta. Uređene podatke moguće je rangovati tako što se najvećem rezultatu dodeli rang 1, sledećem rezultatu po veličini rang 2, i tako redom. Rang određenog rezultata u takvom opadajućem nizu predstavlja **silazni rang** datog rezultata. Rangovanje je moguće sprovesti i tako što se najmanjem rezultatu dodeli rang 1, sledećem rezultatu po veličini rang 2, i tako redom. Rang određenog rezultata u takvom rastućem nizu predstavlja **uzlazni rang** datog rezultata. Uzlazni rang podatka x_i označavaćemo oznakom R_{Ai}, njegov silazni rang oznakom R_{Di} (A u oznaci potiče od engleskog ascending = uzlazni, a D od engleskog descending = silazni).

Pri rangovanju podataka često se dešava da nekoliko jedinica posmatranja imaju isti rezultat, tj. moraju da dele isti rang. Rangovi takvih jedinica posmatranja zovu se deljeni ili vezani rangovi (engl. tied ranks). Postoji nekoliko načina da se pri rangovanju rezultata rangovi dodele jedinicama posmatranja koje imaju isti rezultat. Najčešće se koristi postupak "prosečnog ranga" koji se sastoji u tome da se svim jedinicama posmatranja koje imaju isti rezultat dodeli prosečna vrednost onih rangova koji bi bili dodeljeni tim jedinicama posmatranja kada bi one imale različite rezultate. Prosečna

vrednost rangova dobija se kao količnik zbira i broja tih rangova. Na primer, prosečni rang za rangove 7, 8 i 9 je (7 + 8 + 9) / 3, tj. 7. U postupku "najnižeg ranga" jedinicama posmatranja koje imaju isti rezultat dodeljuje se najmanji od rangova koji bi bili dodeljeni tim jedinicama posmatranja kada bi one imale različite rezultate, a u postupku "najvišeg ranga" jedinicama posmatranja koje imaju isti rezultat dodeljuje se najveći od rangova koji bi bili dodeljeni tim jedinicama posmatranja kada bi one imale različite rezultate. Postupak "sekvencijalnih rangova" sastoji se u dodeljivanju rangova od 1 do s, pri čemu je s broj različitih rezultata. U postupku "sekvencijalnih rangova" jedinice posmatranja sa istim rezultatom dobijaju isti sekvencijalni rang. Različite postupke u dodeli vezanih rangova najlakše je shvatiti na primeru koji je dat u sledećoj tabeli:

Jedinica posmatranja	Rezultat	"Prosečni rang"	"Najniži rang"	"Najviši rang"	"Sekvencijalni rangovi"
1	4	1	1	1	1
2	5	2.5	2	3	2
3	5	2.5	2	3	2
4	6	4	4	4	3
5	7	5	5	5	4
6	8	7	6	8	5
7	8	7	6	8	5
8	8	7	6	8	5
9	10	9	9	9	6

U primeru koji je prikazan u ovoj tabeli vršeno je uzlazno rangovanje - rezultati su rangovani od najmanjeg ka najvećem, tj. najmanjem rezultatu dodeljen je rang 1. U koloni Rezultat vidi se da jedinice posmatranja 2 i 3 imaju isti rezultat. Isto tako, jedinice posmatranja 7, 8 i 9 imaju isti rezultat. Kada bi jedinice 2 i 3 imale različite rezultate ali veće od 4 i manje od 6 tada bi tim jedinicama sledovali rangovi 2 i 3. Dakle, u postupku "prosečnog ranga" jedinice posmatranja 2 i 3 dobijaju isti rang koji je jednak proseku rangova 2 i 3, u postupku "najnižeg ranga" obe jedinice dobijaju manji od ova dva ranga, tj. rang 2, a u postupku "najvišeg ranga" obe jedinice dobijaju veći od ova dva ranga, tj. rang 3. U postupku sekvencijalnih rangova rezultatima se dodeljuju rangovi od 1 do 6 jer ima ukupno 6 različitih rezultata, a jedinice posmatranja 2 i 3 dobijaju isti rang, rang 2, jer su njihovi rezultati drugi po veličini. Uočimo da se u svim postupcima, izuzev u postupku "sekvencijalnih rangova" rezultatu koji je sledeći po veličini u odnosu na rezultate entiteta 2 i 3 dodeljue rang 4 jer su rangovi 2 i 3 su već upotrebljeni pri dodeli vezanih rangova rezultatima entiteta 2 i 3.

<u>Korisna pravila za rangove koja važe samo ukoliko se se za dodelu vezanih rangova primenjuje postupak "prosečnog ranga":</u>

- ✓ Zbir uzlaznih ili silaznih rangova svih podataka za n jedinica posmatranja jednak je $\frac{n(n+1)}{2}$:

$$\sum_{i=1}^{n} R_{Ai} = \sum_{i=1}^{n} R_{Di} = \frac{n(n+1)}{2}$$

- ✓ Zbir uzlaznog i silaznog ranga podatka x_i jednak je n +1, pri čemu je n ukupan broj rangovanih podataka:

$$R_{Ai} + R_{Di} = n + 1$$

Sređivanjem podataka dobijenih merenjem entiteta iz uzorka u pogledu date varijable dobijaju se tzv. statističke serije strukture. Ove serije pokazuju strukturu skupa u pogledu neke varijable i mogu se formirati i za kategoričke i za kvantitativne varijable (cf. Žižić i sar., 2000). Najčešće se radi pregledanja podataka i boljeg razumevanja informacija koje postoje u nizu podataka na kvantitativnoj varijabli formira statistička serije strukture koja se uobičajeno u psihologiji i srodnim oblastima zove raspodela učestalosti ili distribucija frekvencija. Premda formiranje raspodele učestalosti sa stanovišta korisnika statističkih paketa nije neophodno za dalja računanja u ovim paketima, ovaj postupak veoma je važno izvesti pre bilo kakvih daljih statističkih analiza.[7] **Raspodela učestalosti** ili **distribucija frekvencija** za kvantitativnu varijablu predstavlja seriju strukture koja sadrži pojedinačne vrednosti varijable (ili vrednosti grupisane u razredne, tj. grupne intervale) i vrednostima (grupnim intervalima) pridružene učestalosti, tj. frekvencije pojedinih vrednosti (ili svih vrednosti unutar razrednog intervala) varijable.[8] Dakle, raspodela učestalosti predstavlja takvo sređivanje podataka koje pokazuje učestalost, tj. **frekvenciju** (u oznaci f_k) pojedinih vrednosti varijable ili učestalost svih vrednosti unutar određenog razrednog intervala na varijabli.

Raspodela učestalosti može biti jedinična ili sa razrednim, tj. grupnim intervalima. U jediničnoj raspodeli za svaku od pojedinačnih vrednosti varijable koja se pojavljuje u podacima prikazana je njena učestalost ili frekvencija. U raspodeli sa razrednim, tj. grupnim intervalima, nekoliko različitih sukcesivnih vrednosti varijable grupisane su zajedno u intervale, a frekvencija se pridružuje svakom grupnom intervalu. Frekvencija za dati interval pokazuje koliko se često u skupu podataka pojavljuje bilo koja od vrednosti koje ulaze u interval.

Jedinična raspodela učestalosti

Radi pravljenja jedinične raspodele učestalosti potrebno je naprosto utvrditi koliko se puta pojavljuju pojedinačne vrednosti varijable u podacima i upisati učestalosti

[7] Čitalac će se možda zapitati zašto se toliko pažnje u ovom tekstu poklanja pravljenju raspodele učestalosti kada to nije neophodno za dalja računanja i statističke analize. U „stara dobra vremena“ (pa i u „pradavno“ vreme kada je autor ovog teksta bio student psihologije) formiranje raspodele učestalosti bilo je nužno kako bi računanje statističkih mera i dalja statistička analiza bili uopšte izvodljivi na realno velikim skupovima podataka. Naime, potrebna računanja, koja su se tada izvodila bez pomoći računara, nisu bila realno izvodljiva (mada je to u principu moguće) iz velikog skupa nesređenih podataka. Pravljenje raspodele učestalosti bio je mukotrpan posao koji se sastojao u dugotrajnom „tabeliranju“ mera, tj. upisivanju „recki“ (nalik onom pri kartanju) za svaku meru u odgovarajući razredni, tj. grupni interval u distribuciji. (Jednostavan primer vizuelnog ishoda tog postupka zainteresovani čitalac može videti u Dragićević, 2002, str. 34). Ono što je, verujem, bio veoma koristan propratni efekat takvog posla jeste dobro upoznavanje sa strukturom podataka i grubim obrisima oblika raspodele podataka koja se postepeno vizuelno formirala pred očima onoga ko pravi raspodelu. (I danas se sećam oduševljenja koje smo moje koleginice/kolege i ja osećali na vežbama iz Statistike u psihologiji kada bi se posle obavljenog „tabuliranja“ velikog broja podataka, tokom kojeg smo se lepo zabavljali pričajući viceve i druge zanimljivosti iz života, pred nama pojavili obrisi zanimljivih oblika distribucije učestalosti). U današnje vreme, kada su svakome ko koristi statistiku na raspolaganju sofisticirani statistički paketi, formiranje raspodele učestalosti je jednostavan posao koji nije neophodan za dalja računanja te se na važnost pravljenja i pažljivog pregledanja raspodele često zaboravlja. Isto tako, za formiranje dobre i informativne raspodele učestalosti sa grupnim intervalima, kao i za grafičke prikaze ove raspodele u statističkim paketima, još uvek su neophodne intervencije i donošenje važnih odluka od strane korisnika.

[8] U ovom tekstu termine „razredni interval“ i „grupni interval“ koristićemo naizmenično i sinonimno budući da se u statističkim knjigama na našem jeziku sreću oba termina sa istim značenjem (cf. Dragićević, 2002, Žižić i sar, 2000). Taj princip ćemo primenjivati i pri korišćenju drugih statističkim termina. Na taj način čitalac će lakše moći da prati različite statističke udžbenike na našem jeziku.

pojedinačnih vrednosti u odgovarajuću kolonu. Dakle, u jednoj koloni jedinične raspodele učestalosti su (u redovima) poređane vrednosti na kvantitativnoj varijabli (od najmanje do najveće) a u drugu kolonu se za svaku od tih vrednosti u odgovarajući red upisuje učestalost te vrednosti u podacima.

Zamislimo da raspolažemo rezultatima na varijabli uspeh na određenom testu znanja biologije za uzorak od 30 ispitanika. Uspeh na ovom testu znanja iskazuje se brojem tačnih odgovora na 10 pitanja pri čemu se za odgovor na svako pitanje može dobiti nula poena (ako odgovor nije tačan) ili jedan poen (ako je odgovor tačan).[9] Uspeh na testu znanja biologije u ovom slučaju može uzeti vrednosti od 0 do 10. Dobijeni su sledeći rezultati:

7, 6, 5, 9, 6, 4, 5, 0, 8, 6, 5, 5, 10, 4, 7, 6, 1, 5, 2, 5, 2, 8, 5, 4, 5, 7, 6, 5, 6, 6

Na osnovu ovih podataka mogli bismo i bez pomoći statističkog paketa napraviti jediničnu raspodelu učestalosti koja bi izgledala ovako:

x_k	f_k
10	1
9	1
8	2
7	3
6	7
5	9
4	3
3	0
2	2
1	1
0	1
i = 1	n = 30

U prvoj koloni tabele uređene su (poređane su po veličini) moguće vrednosti varijable koje se nalaze unutar raspona podataka, označene sa x_k, a u drugoj koloni učestalosti pojavljivanja svake od ovih vrednosti u skupu podataka, označene sa f_k.
Oznakom i označili smo veličinu razrednog intervala: ona je u ovom slučaju jednaka 1 jer je u intervalu samo jedna mera, tj. svaka mera predstavlja razredni interval. Oznakom n uobičajeno označavamo veličinu uzorka i ona je u slučaju kada za sve ispitanike imamo rezultat jednaka broju rezultata odnosno zbiru učestalosti u koloni f_k. Pregledom tabele možemo uočiti da je najniži rezultat jednak 0, najviši rezultat 10, a da su najčešći rezultati 5 i 6. Uočavamo isto tako da idući od vrednosti 5 ka nižim, odnosno od vrednosti 6 ka višim rezultatima učestalosti pojedinih rezultata bivaju sve manje.

[9] Realni testovi znanja su uobičajeno znatno duži, a realni uzorci ispitanika znatno veći. Mi smo iz didaktičkih razloga celu situaciju maksimalno pojednostavili.

Prve dve kolone u tabeli u kojoj je jedinična raspodela učestalosti napravljena na osnovu ovih podataka u programu SPSS izgledale bi ovako:[10]

Uspeh na testu znanja biologije

		Frequency
Valid	0	1
	1	1
	2	2
	4	3
	5	9
	6	7
	7	3
	8	2
	9	1
	10	1
	Total	30

Prve dve kolone u tabeli iz ispisa programa SPSS su praktično iste kao tabela koju smo prethodno napravili. Međutim u ispisu iz ovog programa ne pojavljuju se u raspodeli učestalosti moguće vrednosti na varijabli kojih nema u podacima, tj. vrednosti sa frekvencijom jednakom nuli. U ovom slučaju frekvencija vrednosti 3 jednaka je nuli i te vrednosti nema u prvoj koloni. Isto tako, vrednosti u prvoj koloni tabele poređane su po veličini ali odozgo nadole! (Mada je moguće izborom odgovarajuće opcije podesiti da se vrednosti u tabeli prikazuju po veličini idući odozdo nadole, zbog posledica koje to ima na prikaz dodatnih korisnih kolona u ovoj tabeli koje nisu ovde prikazane time se ne postiže željeni efekat).

Jedinična raspodela učestalosti pogodna je za situacije u kojima se u podacima pojavljuje relativno mali broj različitih vrednosti (otprilike do 10 ili do 15) kao što je to bio slučaj u prethodnom primeru. Međutim, mnogo su češće situacije u analizama realnih podataka kada je broj različitih vrednosti u podacima znatno veći (30, 50, 100 pa i više različitih vrednosti). Jedinična raspodela učestalosti u tom slučaju ne bi bila pregledna. U takvim situacijama je mnogo bolje napraviti tzv. raspodelu učestalosti sa razrednim, tj. grupnim intervalima.

Raspodela sa grupnim intervalima

Da bi se napravila raspodela sa grupnim intervalima potrebno je odrediti veličinu razrednog, tj. grupnog intervala i broj grupnih intervala. Neka orijentaciona pravila za broj grupnih intervala, u oznaci r, predstavljena su sledećim obrascima (prema Emerson & Hoaglin, 1983): (1) $r \approx [1 + \log_2 n]$; (2) $r \approx [10 * \log_{10} n]$; (3) $r \approx [2\sqrt{n}]$. Pravilo pod (1) se u

[10] Ova tabela uobičajeno sadrži još kolona koje smo za sada izbacili iz tabele i koje ćemo prikazati u nastavku teksta.

statističkim knjigama sreće i u sledećem obliku: $r \approx [1 + 3.3*\log_{10}n]$ (cf. Žižić i sar., 2000). Ugaone zagrade u prethodnim izrazima označavaju da kao vrednost r treba uzeti samo celobrojni deo rezultata koji se dobija računanjem izraza u zagradi, a znak ≈ je matematička oznaka za "približno jednako". Oznakom n označena je veličina uzorka, tj. broj rezultata na varijabli a $\log_a(\cdot)$ predstavlja logaritam za osnovu a.

Prikazana pravila su orijentaciona i ukazuju na maksimalni broj grupnih intervala koje bi trebalo koristiti za date podatke. Primena različitih pravila neće dati iste rezultate: pravilo pod (1) će uobičajeno dati manje r od pravila pod (2) i (3). Pravilo pod (2) je korisno konsultovati kao orijentir za maksimalni broj grupnih intervala, tj. broj intervala iznad kojeg ne bi trebalo ići. Pravilo pod (3) ima smisla koristiti pre svega u situacijama kada je broj rezultata mali (do 50), a pravilo pod (1) korisno je konsultovati kao orijentir za minimalni broj grupnih intervala koje ima smisla upotrebiti.

Isto tako, postoje orijentaciona pravila za širinu grupnog intervala (prema Emerson & Hoaglin, 1983):

- ✓ Prema Skotovom pravilu širina grupnog intervala trebalo bi da bude približno jednaka vrednosti izraza $3.49*S*n^{-1/3}$, pri čemu je S standardna devijacija skupa rezultata (koja je definisana u nastavku teksta u ovoj glavi), a n veličina uzorka.
- ✓ Prema pravilu Fridmana i Diakonisa širina grupnog intervala trebalo bi da bude približno jednaka vrednosti izraza $(2*IQR)*n^{-1/3}$, pri čemu je IQR interkvartilni raspon (koji je definisan u nastavku teksta u ovoj glavi), a n veličina uzorka.

Pojedina od prikazanih pravila za broj i širinu grupnih intervala izvedena su na osnovu pretpostavki da raspodele rezultata imaju određeni teorijski oblik (na primer, oblik binomne ili normalne raspodele koje smo prikazali u glavi **).[11] Važno je uočiti da svi orijentacioni obrasci za broj grupnih intervala taj broj određuju kao funkciju veličine uzorka, tj. broja rezultata, dok orijentacioni izrazi za širinu grupnog intervala sadrže pored veličine uzorka i neku od statističkih mera varijabilnosti, tj. raspršenja rezultata. To je sasvim logično jer raspodela ne sme da bude suviše "razvučena" (veliki broj grupnih intervala sa malim učestalostima) jer postaje nepregledna, niti suviše "zbijena" (mali broj grupnih intervala sa velikim učestalostima) jer ne daje dovoljno detaljnih informacija. Naravno, najniži i najviši grupni interval u raspodeli učestalosti treba da obuhvate najniži i najviši rezultat u skupu podataka. Navedena pravila za broj grupnih intervala i širinu intervala mogu poslužiti samo kao orijentacija, a u svakom konkretnom slučaju istraživač mora sam odlučiti koji od obrazaca daje rešenje koje je najbliže optimalnom. Stoga je pravljenje grupisane raspodele učestalosti veština koja se stiče iskustvom. Dok se takvo iskustvo ne stekne, potrebno je probati na istim podacima nekoliko mogućih rešenja i odabrati ono koje je istovremeno i najpreglednije i najinformativnije. Za početak, sasvim dobru orijentaciju mogu predstavljati sledeće preporuke (prema Shavelson, 1988 i Dragićević, 2002):

- ✓ Raspodela ne bi trebalo da ima manje od 10, niti više od 20 grupnih intervala. Ipak, ako je raspon rezultata mali i ako ima malo rezultata, raspodela može imati između 5 i 10 grupnih intervala. Najprihvatljiviji broj grupnih intervala za većinu realnih podataka je između 10 i 15.
- ✓ Širina grupnog intervala, u oznaci i, bira se tako da približno odgovara broju koji se dobije kada se razlika najvećeg i najmanjeg rezultata podeli brojem grupnih

[11] Matematičke detalje dolaska do ovih pravila zainteresovani čitalac može pogledati u Emerson & Hoaglin, 1983.

intervala:

$$i \approx \frac{x_{max} - x_{min}}{r}$$

- ✓ Donja merna granica najnižeg intervala u raspodeli može se odrediti na tri načina:

1. Najniži rezultat se uzima kao donja merna granica najnižeg intervala.
Merna granica intervala je granica koja je definisana u jedinicama koje smo koristili u merenju varijable, ili, bolje rečeno, granica koja je iskazana sa onom preciznošću sa kojom su iskazani rezultati na osnovu kojih pravimo distribuciju. Na primer, ako su rezultati samo celi brojevi od 9 do 39 tada se kao merna granica bilo kojeg intervala uzima određeni celi broj od 9 do 39. U navedenom primeru donja merna granica najnižeg intervala bila bi jednaka 9. /Ako bismo odlučili da širina grupnog intervala bude jednaka 3 (i = 3) onda to znači da bismo u tom grupnom intervalu imali mere 9, 10 i 11 te bi gornja merna granica najnižeg intervala bila 11/. Ako su, pak, rezultati na varijabli iskazani kao decimalni brojevi na jednu decimalu tačnosti, tada bi donja i gornja merna granica intervala trebalo da budu decimalni brojevi sa jednom decimalom. Na primer, ako su rezultati decimalni brojevi od 8.5 do 35.5 (dakle, decimalni brojevi sa jednom decimalom) onda i merne granice treba iskazati sa jednom decimalom tačnosti). U ovom primeru bi donja merna granica najnižeg intervala mogla biti 8.5. Gornja granica intervala (ako bismo hteli da širina intervala bude 3 bila bi 11.4. Ukoliko bismo u potonjem primeru grupne intervale definisali mernim granicama iskazanim celim brojevima onda bismo imali problem da smestimo svaki rezultat u odgovarajući interval. Na primer, ako bi najniži grupni interval bio 9–11 a sledeći grupni interval 12–14 tada bismo imali problem sa raspodelom po intervalima svih rezultata od 11.1 do 11.9.
Ad hoc objašnjenje širine grupnog intervala koje smo dali u primeru u kojem su rezultati celi brojevi od 9 do 39 ne izgleda sasvim jasno i ne može se uopštiti na sve situacije. Pojam širine grupnog intervala zapravo treba definisati korišćenjem tzv. egzaktnih ili realnih granica grupnog intervala. Merne granice grupnog intervala možemo koristiti i za diskretne i za teorijski kontinuirane varijable. Egzaktne granice grupnog intervala ima smisla koristiti za teorijski kontinuirane varijable. Međutim, računsko određivanje širine grupnog intervala korišćenjem donjih i gornjih egzaktnih granica grupnog intervala možemo koristiti i za diskretne varijable. Donja egzaktna granica intervala dobija se kada se od donje merne granice intervala oduzme polovina merne jedinice, a gornja egzaktna granica intervala dobija se dodavanjem polovine merne jedinice na gornju mernu granicu intervala. Ako su mere (stoga i merne granice intervala) iskazane celim brojevima onda je polovina merne jedinice "pola celog", tj. 0.5. Ako su, pak, mere (stoga i merne granice intervala) iskazane decimalnim brojevima polovina merne jedinice zavisi od broja decimala kojima su iskazane mere: ako su mere iskazane na jednu decimalu tačnosti onda je polovina merne jedinice "pola desetog dela celog", tj. 0.05, ako su mere iskazane na dve decimale tačnosti onda je polovina merne jedinice "pola stotog dela celog", tj. 0.005 i tako redom. Na primer, ako su mere, tj. rezultati iskazani celim brojevima od 9 do 39, a varijabla je teorijski kontinuirana, donja egzaktna granica najnižeg intervala bila bi 9 - 0.5 = 8.5. Ako smo unapred odredili da širina intervala bude jednaka 3, onda bi

najniži grupni interval obuhvatao mere 9, 10 i 11 pa bi gornja merna granica najnižeg intervala bila jednaka 11. U tom slučaju gornja egzaktna granica tog intervala bila bi 11 + 0.5 = 11.5. Ako su, pak, rezultati na nekoj varijabli iskazani decimalnim brojevima sa jednom decimalom tačnosti širinu intervala ne možemo odrediti na način analogan primeru sa celim brojevima, tj. prebrojavanjem mogućih vrednosti u tom intervalu. Na primer ako su rezultati decimalni brojevi od 8.5 do 39.5 sa jednom decimalom tačnosti tada bi, ako je najniži grupni interval iskazan mernim granicama 8.5–11.4, donja egzaktna granica tog intervala bila bi 8.5 - 0.5(0.1) = 8.5 - 0.05 = 8.45. Ukoliko odlučimo da širina grupnog intervala bude 3, gornja egzaktna granica ovog intervala bila bi 8.45 + 3 = 11.45. Uočimo da je to isto kao da smo na gornju mernu granicu dodali polovinu merne jedinice, tj. jednog desetog jer je 11.4 + 0.5(0.1) = 11.4 + 0.05 = 11.45. Dakle, u opštem slučaju **širina grupnog intervala, u oznaci i, jednaka je razlici gornje egzaktne granice (G) i donje egzaktne granice (D) grupnog intervala**:

$$i = G - D$$

2. Ako je varijabla za koju se pravi raspodela učestalosti teorijski kontinuirana najniži rezultat se može uzeti i kao sredina, tj. srednje mesto najnižeg intervala, pri čemu se srednje mesto intervala, u oznaci SMI, određuje po sledećem obrascu:

$$SMI = D + \frac{G - D}{2}.$$

Pri tome su D i G, donja i gornja egzaktna granica grupnog intervala, tim redom. Radi postizanja primenljivosti ovog načina određivanja najnižeg grupnog intervala na rezultate koji su iskazani celim brojevima širina grupnog intervala treba da bude neparan broj. Na primer, ako su rezultati iskazani celim brojevima od 9 do 39 a širina intervala je 3, da bi najniži rezultat (rezultat 9) bio srednje mesto najnižeg intervala taj interval iskazan mernim granicama mora biti interval 8–10. U tom slučaju SMI = 7.5 + (10.5 – 7.5)/2 = 7.5 + 1.5 = 9. Ukoliko bismo, pak, uzeli da širina intervala bude paran broj tada srednje mesto intervala ne bi moglo bito jednako najnižem rezultatu. Kada se odredi srednje mesto najnižeg intervala donja merna granica se može dobiti i tako što se na donju egzaktnu granicu najnižeg intervala dodaje polovina jedinice kojom su iskazane mere na varijabli.

3. Ako su rezultati iskazani celim brojevima, unapred određena širina intervala množi se uzastopnim celim brojevima dok se ne dobije broj koji je jednak najnižem rezultatu ili broj koji je najbliži najmanjem rezultatu i istovremeno manji od najnižeg rezultata. Broj koji ispunjava navedene uslove uzima se kao donja merna granica najnižeg intervala. Na primer, ako su rezultati iskazani celim brojevima od 9 do 39 a širina intervala je 3, tada bi se množenjem širine grupnog intervala uzastopnim celim brojevima dobili brojevi 3, 6, 9. Budući da je 9 najniži rezultat u skupu rezltata, najniži grupni interval bio bi interval 9–11. Premda je najniži interval određen ovim načinom isti kao i najniži interval određen načinom objašnjenim pod 1 to u opštem slučaju ne mora biti tako.

U slučaju kada tri prikazana načina daju različita rešenja za određivanje najnižeg grupnog intervala treba odabrati onaj koji daje najpregledniju i najinformativniju raspodelu. Za odabir najboljeg rešenja potrebno je iskustvo koje se vremenom stiče pri radu sa podacima.

- ✓ Kada se odrede donja i gornja merna granica najnižeg intervala donja, odnosno gornja merna granica sledećeg intervala dobijaju se tako što se širina intervala dodaje na donju, odnosno gornju mernu granicu najnižeg intervala. Na isti način se na osnovu donje i gornje merne granice prethodnog intervala dobijaju donja i gornja egzaktna granica bilo kojeg narednog intervala u raspodeli. Na primer, ako je najniži grupni interval 0–4, a širina grupnog intervala 5, tada je sledeći grupni interval (0 + 5)–(4 + 5), tj. 5–9. Interval posle intervala 5–9 bio bi interval (5 + 5) – (9 + 5), tj. 10–14, i tako redom. Po istom principu se, kada se odredi donja i gornja egzaktna granica najnižeg intervala određuju egzaktne granice svakog narednog intervala.

- ✓ Svi grupni intervali bi trebalo da budu iste širine.

- ✓ Grupni intervali treba da budu tako napravljeni da svi zajedno budu iscrpni (da obuhvate sve podatke, a različiti grupni intervali uzajamno isključivi. Dakle, ne sme se desiti da neki podatak ne možemo da svrstamo ni u jedan grupni interval, niti da neki podatak možemo svrstati u više od jednog intervala.

- ✓ Pri pravljenju grupisane raspodele učestalosti treba pre svega voditi računa o tome da ta raspodela treba da bude informativna i pregledna, tj. da omogućuje što jasnije uočavanje strukture poodataka i oblika same raspodele.

Opšti format distribucije frekvencija sa mernim granicama intervala u slučaju kada su vrednosti na varijabli celobrojne (a to je u psihologiji najćešći slučaj) i kada je donja merna granica najnižeg intervala jednaka najmanjem rezultatu prikazan je u sledećoj tabeli (slovom i u toj tabeli označena je veličina grupnog intervala, k je oznaka redosleda grupnog intervala idući od najnižeg intervala za koji je $k = 1$, r je broj grupnih intervala, a f_k je oznaka frekvencije u grupnom intervalu k):

Grupni intervali	Frekvencije
$(x_{min} + (r - 1)i) – (x_{min} + ri - 1)$	f_r
----------	--
$(x_{min} + (k - 1)i)–(x_{min} + ki - 1)$	f_k
----------	--
$(x_{min} + 2i)–(x_{min} + 3i - 1)$	f_3
$(x_{min} + i)–(x_{min} + 2i - 1)$	f_2
$(x_{min})–(x_{min} + i -1)$	f_1
i =	n =

Na primer, ako je najmanji rezultat u nekom skupu mera dobijenih merenjem jedinica

uzorka u pogledu kvantitativne varijable jednak 0, najveći rezultat 48 a širina grupnog intervala 4 tada bi raspodela sadržala 13 grupnih intervala. Kolona sa grupnim intervalima u toj raspodeli izgledala bi ovako:

48–51
44–47
40–43
36–39
32–35
28–31
24–27
20–23
16–19
12–15
8–11
4–7
0–3

Uočimo da, kada su podaci na varijabli celobrojni, širina grupnog intervala govori o tome koliko je pojedinačnih mera na varijabli obuhvaćeno grupnim intervalom. Tako, interval 0–3 obuhvata mere 0, 1, 2 i 3 dok interval 36–39 obuhvata mere 36, 37, 38 i 39. Isto tako, važno je uočiti da su razlike između donjih mernih granica sukcesivnih intervala u raspodeli (4 - 0, 8 - 4, 12 - 8...), kao i razlike između gornjih mernih granica sukcesivnih intervala (7 - 3, 11 - 7, 15 - 11...) jednake širini intervala.

Međutim, kao što smo to u prethodnom delu teksta naglasili, u opštem slučaju ovakvo određenje širine grupnog intervala nije primenljivo. Ono što dodatno može da zbuni je razlika između gornje i donje merne granice intervala: ta razlika je kod celobrojnih mera za jedan manja od širine intervala. Stoga je bolje uvek određivati širinu grupnog intervala prema opštoj definiciji ovog pojma, tj. na osnovu razlike gornje i donje egzaktne granica intervala. Kolona grupnih intervala koje smo u ovom primeru prikazali koristivši merne granice intervala može se prikazati i korišćenjem egzaktnih granica intervala. U tom slučaju kolona iz prethodnog primera izgledala bi ovako:

47.5–51.5
43.5–47.5
39.5–43.5
35.5–39.5
31.5–35.5
27.5–31.5
23.5–27.5
19.5–23.5
15.5–19.5
11.5–15.5
7.5–11.5
3.5–7.5
-0.5–3.5

Uočimo da su razlike između gornje egzaktne i donje egzaktne granice grupnog intervala jednake širini intervala, u ovom slučaju 4. Na primer razlika egzaktnih granica u najnižem intervalu je 3.5 - (-0.5) = 4. Isto tako, razlika egzaktnih granica intervala 11.5–

15.5 je 15.5 - 11.5 = 4. Uočimo i da je u prikazu grupnih intervala korišćenjem egzaktnih granica gornja egzaktna granica bilo kojeg grupnog intervala jednaka donjoj egzaktnoj granici prvog intervala koji je iznad njega. Na primer, gornja egzaktna granica intervala 0–3 je 3.5. Ta vrednost je istovremeno donja egzaktna granica intervala 4–7. Time se stvara realni kontinuum koji, iako varijabla može biti kontinuirana, nije vidljiv kada se grupni intervali prikazuju u mernim granicama. Jednakost vrednosti koje predstavljaju donju egzaktnu granicu nekog intervala i gornju egzaktnu granicu narednog intervala ništa ne smeta: budući da su sve vrednosti koje imamo u podacima celobrojne nećemo imati problema u njihovom smeštanju u odgovarajući interval.

Ukoliko bi mere na datoj varijabli bile date u istom rasponu od 0 do 48, ali su merene sa jednom decimalom tačnosti, tada ne bismo smeli da pravimo grupne intervale sa celobrojnim vrednostima mernih granica. Naime, tada bismo imali rezultate koji su decimalni brojevi sa jednom decimalom, na primer, mogući rezultati bi bili 0.4, 3.9, 11.5, 22.4, 27.8, 42.3, 47.5 i tako redom. U tom slučaju mere poput 3.9 ne bismo mogli svrstati ni u grupni interval 0–3 niti u interval 4–7. Rešenje nije ni u istim grupnim intervalima koji bi bili iskazani egzaktnim granicama jer u tom slučaju ne bismo znali da li meru poput 11.5 da svrstamo u grupni interval 7.5–11.5 ili u interval 11.5–15.5. ***Dakle, grupne intervale sa mernim granicama treba iskazati u onim "jedinicama", tj. sa onom tačnošću sa kojom su dobijene mere na varijabli***. To praktično znači da ako su mere na varijabli celi brojevi onda i merne granice intervala mogu biti celi brojevi. Ako su, pak mere na varijabli date u decimalnim brojevima onda i merne granice intervala treba da budu decimalni brojevi sa onoliko decimala koliko najviše decimala ima u merama. Dakle, ako bi se mere na varijabli kretale u rasponu od 0 do 48.0, ali su iskazane decimalnim brojevima sa jednom decimalom tada bi grupni intervali u mernim granicama mogli izgledati ovako:

47.5–51.4
43.5–47.4
39.5–43.4
35.5–39.4
31.5–35.4
27.5–31.4
23.5–27.4
19.5–23.4
15.5–19.4
11.5–15.4
7.5–11.4
3.5–7.4
-0.5–3.4

(Donja merna granica najnižeg intervala je negativan broj što može delovati kao nemoguća vrednost na mnogim psihološkim varijablama. Čak i ako jeste tako, ovakva merna granica može biti odabrana kako bi se očuvala jednaka širina svih grupnih intervala). Treba uočiti da su i u ovom slučaju razlike između donjih mernih granica sukcesivnih intervala u raspodeli (3.5 - (-0.5), 7.5 - 3.5, 11.5 – 7.5...), kao i razlike između gornjih mernih granica sukcesivnih intervala (7.4 – 3.4, 11.4 – 7.4, 15.4 – 11.4...) jednake širini intervala.

Isti grupni intervali ali u egzaktnim granicama izgledali bi ovako:

47.45–51.45
43.45–47.45
39.45–43.45
35.45–39.45
31.45–35.45

27.45–31.45
23.45–27.45
19.45–23.45
15.45–19.45
11.45–15.45
7.45–11.45
3.45–7.45
-0.45–3.45

U opštem slučaju, širina grupnog intervala ne mora biti ceo broj. Ukoliko su podaci u celobrojnom obliku onda i širina grupnog intervala treba da bude ceo broj. Ako su podaci u obliku decimalnih brojeva, širina grupnog intervala može biti i decimalni broj sa onoliko decimala koliko decimala imaju i podaci.

Pored „obične“, tj. apsolutne učestalosti ili frekvencije za meru ili interval *k*, koja se najčešće označava oznakom f_k, kao važni sastavni elementi raspodele učestalosti koriste se relativna frekvencija i kumulativna relativna frekvencija.

Relativna frekvencija za vrednost ili grupni interval *k*, u oznaci p_k, računa se na sledeći način:

$$\boxed{p_k = \frac{f_k}{n}}$$

U ovom obrascu f_k je frekvencija ili učestalost za *k*-tu vrednost varijable ili *k*-ti grupni interval, a n je veličina uzorka. Relativna frekvencija može se iskazati proporcijom koja se dobija primenom obrasca ** ili procentom. Kako bismo relativnu frekvenciju iskazali procentom potrebno je p_k pomnožiti sa 100.

Kumulativna frekvencija za *k*-tu vrednost rezultata poređanih po veličini ili za *k*-ti grupni interval, u oznaci cf_k, određuje se na sledeći način:

$$\boxed{cf_k = \sum_{j<k} f_j + f_k\,, \quad k = 1, 2, \ldots r}$$

U ovom obrascu f_j je frekvencija za *j*-ti grupni interval ili *j*-tu vrednost rezultata poređanih po veličini, pri čemu je $j < k$. Dakle, kumulativna frekvencija za *k*-tu vrednost ili *k*-ti interval dobija se kao zbir učestalosti svih vrednosti manjih od *k*-te vrednosti (ili svih grupnih intervala ispod *k*-tog intervala) i učestalosti *k*-te vrednosti (ili svih mera u *k*-tom intervalu). Očigledno, kumulativna frekvencija u najvišem grupnom intervalu, grupnom intervalu r, jednaka je broju rezultata, tj. $cf_r = n$.

Relativna kumulativna frekvencija za *k*-tu vrednost rezultata poređanih po veličini ili za *k*-ti grupni interval ($k = 1, 2, \ldots, r$) u oznaci cp_k, definiše se na sledeći način:

$$\boxed{cp_k = \frac{cf_k}{n}}$$

Kao i relativna frekvencija, i relativna kumulativna frekvencija može se iskazati proporcijom ili procentom. Ako želimo da relativnu kumulativnu frekvenciju iskažemo

procentom onda cp_k dobijenu obrascem ** treba pomnožiti sa 100.

Na primer, apsolutne frekvencije u dva najniža i dva najviša grupna intervala u raspodeli izgledaju ovako:

x_k	f_k
48–51	1
44–47	1
--------	----------
4–7	38
0–3	22
i = 4	n = 249

(Isprekidanim linijama u tabeli označeno je da nedostaju grupni intervali koji su između onih koji su prikazani).

Prikazani deo ove raspodele sa dodatnim kolonama relativnih frekvencija, kumulativnih frekvencija i relativnih kumulativnih frekvencija izgledao bi ovako:

x_k	f_k	p_k	cf_k	cp_k
48–51	1	0.004	249	1
44–47	1	0.004	248	0.996
--------	----------	----------	----------	----------
4–7	38	0.153	60	0.241
0–3	22	0.088	22	0.088
i = 4	n = 249	1		

Relativna frekvencija (kolona p_k) u grupnom intervalu 0–3 dobijena je kao količnik obične frekvencije i ukupnog broja rezultata (n): 22/249 = 0.088. Istim postupkom dobijene su relativne frekvencije u ostalim grupnim intervalima. Uočimo da je zbir svih vrednosti u koloni relativnih frekvencija, kada se relativne frekvencije iskazuju proporcijama jednak 1. U glavi o teoriji verovatnoće istakli smo da se verovatnoća može, ako je broj pokušaja veliki, približno odrediti relativnom učestalošću. Dakle, ako je uzorak slučajan i dovoljno veliki, relativna frekvencija za grupni interval može poslužiti kao ocena verovatnoća da varijabla uzme neku od vrednosti u datom intervalu. I upravo kao i zbir verovatnoća za neku slučajnu varijablu, tako i zbir relativnih frekvencija na varijabli mora biti jednak 1 (ako su relativne frekvencije iskazane proporcijama) ili 100% (ako se relativne frekvencije iskazuju procentima).
Kumulativna frekvencija (kolona cf_k) u najnižem grupnom intervalu (0–3) jednaka je običnoj frekvenciji za taj razred, tj. 22, jer je zbir frekvencija ispod tog grupnog intervala jednak 0. Kumulativna frekvencija za grupni interval 4–7 jednaka je zbiru svih frekvencija ispod tog grupnog intervala i frekvencije u tom grupnom intervalu: 22 + 38 = 60. Kumulativna frekvencija za grupni interval 44–47 jednaka je zbiru svih frekvencija ispod tog intervala (iako se ne vide sve frekvencije njihov zbir je 247) i frekvencije u tom intervalu: 247 + 1 = 248. Po istom principu, kumulativna frekvencija u najvišem grupnom intervalu je 249 i jednaka je nužno ukupnom broju rezultata.
Relativna kumulativna frekvencija u najnižem intervalu ista je ista kao i relativna

frekvencija u tom intervalu, tj. 0.088. Relativna kumulativna frekvencija u intervalu 4–7 jednaka je zbiru relativne kumulativne frekvencije do tog grupnog intervalu (0.088) i relativne frekvencije u tom intervalu (0.153): 0.088 + 0.153 = 0.241. Naravno, ista vrednost se dobija kada se kumulativna frekvencija u tom grupnom intervalu podeli ukupnim brojem rezultata: 60 / 249 = 0.241. Treba uočiti da je relativna kumulativna frekvencija za najviši grupni interval jednaka 1, što je nužno. Isto tako, treba uočiti važnu analogiju između vrednosti funkcije distribucije slučajnih varijabli (što smo definisali u prikazu teorije verovatnoće) i relativnih kumulativnih frekvencija. Dakle, ako je uzorak slučajan i dovoljno veliki relativne kumulativne frekvencije mogu da posluže kao ocene vrednosti funkcije distribucije za datu varijablu, tj. ocene verovatnoće da varijabla uzme neku vrednost od najniže moguće vrednosti do vrednosti koja je jednaka gornjoj granici grupnog intervala. U prikazanom primeru verovatnoća da varijabla uzme neku vrednost manju od 7 ili jednaku 7 bila bi ocenjena kao da iznosi 0.241 ili 24.1%.

Ako je uzorak jako mali (n < 50) nema mnogo smisla da u distribuciji učestalosti, pored običnih frekvencija i kumulativnih frekvencija prikazujemo relativne frekvencije ili kumulativne relativne frekvencije. U tom slučaju, naime, male razlike u učestalostima mogu dovesti do velikih razlika u relativnim frekvencijama jer su ove frekvencije iskazane proporcijama, odnosno procentima. Da bismo to jasnije uočili prikazaćemo kompletnu jediničnu distribuciju učestalosti iz programa SPSS za zamišljeni primer u kojem smo već prikazali jediničnu distribuciju učestalosti rezultata na testu znanja biologije "dobijenih" na uzorku od 30 ispitanika. U delu teksta u kojem smo objašnjavali jediničnu distribuciju učestalosti prikazali smo samo prve dve kolone iz ove distribucije. Kompletna jedinična distribucija iz ispisa programa SPSS koja je dopunjena relativnim frekvencijama i relativnim kumulativnim frekvencijama izgleda ovako:

Uspeh na testu znanja biologije

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	3.3	3.3	3.3
	1	1	3.3	3.3	6.7
	2	2	6.7	6.7	13.3
	4	3	10.0	10.0	23.3
	5	9	30.0	30.0	53.3
	6	7	23.3	23.3	76.7
	7	3	10.0	10.0	86.7
	8	2	6.7	6.7	93.3
	9	1	3.3	3.3	96.7
	10	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

U koloni **Frequency** ove tabele su obične frekvencije, u kolonama **Percent** i **Valid Percent** su relativne frekvencije u procentima, a u koloni **Cumulative Percent** su relativne

kumulativne frekvencije. Uočimo kako se za malu promenu u frekvencijama u koloni **Frequency** znatno promene relativne frekvencije u procentima u koloni **Percent**. Kolone **Percent** i **Valid Percent** su u ovom slučaju identične ali to u opštem slučaju nije tako (u prvom primeru ispisa u kojem ove kolone ne budu identične objasnićemo razliku između njih).

Pravljenje grupisane raspodele učestalosti prikazaćemo na primeru realnih podataka koji su prikupljeni ispitivanjem slučajnog uzorka od 252 onkološka pacijenta u našoj zemlji upitnikom depresivnosti CES-D. Tri ispitanika nisu do kraja popunila ovaj upitnik tako da u stvari raspolažemo sa 249 rezultata. Naravno, iz tehničkih razloga nećemo sve pojedinačne rezultate prikazati ovde (to bi nepotrebno zauzimalo prostor) ali ćemo dati dovoljno informacija za razumevanje postupka pravljenja grupisane raspodele učestalosti na osnovu dobijenih rezultata na varijabli depresivnosti. Premda se ukupni skor depresivnosti na ovom upitniku teorijski može kretati u granicama od 0 do 60, uređivanjem rezultata ustanovljeno je da je minimalni rezultat na ispitivanom uzorku iznosio 0 a maksimalni rezultat bio je 48. (Niži maksimalni rezultat dobijen na ispitivanom uzorku nego što je maksimalno moguć rezultat na ovom upitniku ne iznenađuje budući da u ovom ispitivanju nije korišćen uzorak psihijatrijskih pacijenata sa dijagnozom depresivnih poremećaja). Jedinična distribucija sa 49 vrednosti i njima pridruženim frekvencijama bila bi u ovom slučaju potpuno nepregledna.

Primenom pravila za broj grupnih intervala za n = 249 dobijaju se sledeće orijentacione vrednosti: pravilo (1) – 8 intervala; pravilo (2) – 23 intervala; pravilo (3) – 31 interval;

Budući da je n znatno iznad 50 ishod dobijen na osnovu pravila (3) možemo odmah odbaciti. Dakle, ima smisla da broj grupnih intervala bude između 8 i 23.

Primenom pravila za širinu grupnog intervala za n = 249, S = 9.73 i IQR = 12 dobijaju se sledeće orijentacione vrednosti: Skotovo pravilo sugeriše da bi širina intervala trebalo da bude 5.4, a prema pravilu Fridmana i Diakonisa širina intervala bi trebalo da bude 3.81. /Standardnu devijaciju (S) i interkvartilni raspon (IQR) objasnićemo u nastavku teksta u ovom poglavlju/. Budući da širina grupnog intervala treba da bude ceo broj jer su rezultati celobrojni, dobijene ishode možemo tumačiti kao da sugerišu da širina grupnog intervala bude 5 ili 4. Kao što smo u prethodnom delu teksta u ovoj glavi naveli (formula **), širinu intervala približno određujemo deljenjem razlike najvećeg i najmanjeg rezultata sa brojem grupnih intervala. Na osnovu toga možemo približno odrediti broj grupnih intervala u raspodeli za datu širinu grupnog intervala. Budući da je najmanji rezultat u skupu podataka za koji pravimo grupisanu raspodelu 0 a najveći 48, razlika ova dva rezultata jednaka je 48. Prema tome, ako bismo odabrali da širina grupnog intervala bude 5 tada bismo napravili raspodelu sa približno 10 grupnih intervala, a ako odlučimo da širina grupnog intervala bude 4 u raspodeli bi bilo približno 12 intervala. Vidimo da su obe mogućnosti u skladu sa ishodom do kojeg smo došli primenom orijentacionih pravila pod (1) i pod (2) za broj grupnih intervala.

Distribucija učestalosti sa širinom intervala jednakom 5 iz ispisa programa SPSS dobijena korišćenjem komande **Recode into Different Variables** i procedure **Frequencies** izgleda ovako:[12]

[12] Korišćenje procedure Frequencies u programu SPSS čitalac može naučiti sledeći video instrukcije br.3...**

Rezultat na upitniku depresivnosti CES_D grupisano2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-4	36	14.3	14.5	14.5
	5-9	38	15.1	15.3	29.7
	10-14	66	26.2	26.5	56.2
	15-19	40	15.9	16.1	72.3
	20-24	28	11.1	11.2	83.5
	25-29	20	7.9	8.0	91.6
	30-34	8	3.2	3.2	94.8
	35-39	8	3.2	3.2	98.0
	40-44	3	1.2	1.2	99.2
	45-49	2	.8	.8	100.0
	Total	249	98.8	100.0	
Missing	System	3	1.2		
Total		252	100.0		

Distribucija učestalosti sa širinom intervala jednakom 4 u ispisu iz programa SPSS izgleda ovako:

Rezultat na upitniku depresivnosti CES_D grupisano1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-3	22	8.7	8.8	8.8
	4-7	38	15.1	15.3	24.1
	8-11	31	12.3	12.4	36.5
	12-15	58	23.0	23.3	59.8
	16-19	31	12.3	12.4	72.3
	20-23	21	8.3	8.4	80.7
	24-27	18	7.1	7.2	88.0
	28-31	13	5.2	5.2	93.2
	32-35	7	2.8	2.8	96.0
	36-39	5	2.0	2.0	98.0
	40-43	3	1.2	1.2	99.2
	44-47	1	.4	.4	99.6
	48-51	1	.4	.4	100.0
	Total	249	98.8	100.0	
Missing	System	3	1.2		

Rezultat na upitniku depresivnosti CES_D grupisano1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-3	22	8.7	8.8	8.8
	4-7	38	15.1	15.3	24.1
	8-11	31	12.3	12.4	36.5
	12-15	58	23.0	23.3	59.8
	16-19	31	12.3	12.4	72.3
	20-23	21	8.3	8.4	80.7
	24-27	18	7.1	7.2	88.0
	28-31	13	5.2	5.2	93.2
	32-35	7	2.8	2.8	96.0
	36-39	5	2.0	2.0	98.0
	40-43	3	1.2	1.2	99.2
	44-47	1	.4	.4	99.6
	48-51	1	.4	.4	100.0
	Total	249	98.8	100.0	
Missing	System	3	1.2		
Total		252	100.0		

Koja od ovih dveju varijanti distribucije depresivnosti sa grupnim intervalima je bolja? Obe distribucije su relativno pregledne. Razgledanjem distribucije sa širinom intervala 5 vidimo da su frekvencije u nekoliko najnižih grupnih intervala suviše velike, što znači da je raspodela u tom delu previše "zbijena". Pored toga, za odluku koju od ovih distribucija ćemo odabrati za dalje razgledanje i zaključivanje bitna je i sledeća informacija: prema ključu ovog upitnika rezultat jednak 16 i veći od toga upućuje na klinički relevantan nivo depresivnosti, tj. na neophodnost dalje kliničke evaluacije. Prema tome, iz potonje distribucije možemo da izvučemo važnu informaciju o broju, odnosno procentu ispitanika sa klinički relevantnim nivoom depresivnosti. Taj procenat lako dobijamo tako što od 100 oduzmemo kumulativnu relativnu frekvenciju za grupni interval 12–15 (procenat onih koji imaju rezultat manji od 16), što iznosi 59.8%. Dakle, od onih onkoloških pacijenata koji su popunili upitnik CES-D, njih 40.2% pokazuje klinički relevantan nivo depresivnosti! Naravno, to još uvek ne znači da 40.2% svih onkoloških pacijenata u Srbiji pokazuje klinički relevantan nivo depresivnosti.[13] (Statističkim postupcima koji su od pomoći pri zaključivanju o stanju u populaciji na osnovu podataka dobijenih na slučajnom uzorku posvećene su gotovo sve preostale glave ove knjige počevši od glave **). Na kraju, iz distribucije sa širinom intervala jednakom 4 jasnije uočavamo i ispitanike sa izuzetno visokim skorovima na koje bi trebalo obratiti pažnju: jedan ispitanik ima rezultat u intervalu 44–47, dok još jedan ispitanik ima rezultat u intervalu od 48–51. Pregledom

[13] Kako bismo malo ublažili težinu ove informacije navodimo i sledeći podatak: približno 20% onih koji imaju visok rezultat na ovom upitniku pokazuju brz oporavak od simptoma i u detaljnom strukturisanom intervjuu sa psihijatrom ustanovljava se da ne ispunjavaju kriterijume za dijagnozu kliničke depresije (Depressio maior).

sortiranih, tj. uređenih rezultata možemo lako ustanoviti konkretne skorove ovih ispitanika: 46 i 48. Ovi ispitanici imaju primetno veći skor od ispitanika sa sledećim najvećim rezultatom koji iznosi 42.

Dakle, na osnovu razmatranja koja smo prikazali u potonjem pasusu prednost bismo u daljem razgledanju i prikazivanju rezultata u ovom slučaju dali distribuciji sa širinom grupnog intervala jednakom 4.[14]

Podaci koji nedostaju (engl. Missing data)

Pregledom potonjih dveju distribucija učestalosti iz ispisa programa SPSS može se uočiti da se pri kraju kolone sa grupnim intervalima nalazi natpis **Missing System**. To je oznaka za nedostajanje podataka. U koloni sa frekvencijama u redu **Missing System** stoji broj 3. To znači da za 3 ispitanika ne postoji rezultat na varijabli depresivnosti jer ovi ispitanici nisu popunili upitnik CES-D. Relativne frekvencije u koloni **Percent** računate su u odnosu na ukupan broj ispitanika u uzorku (n = 252), bez obzira na to da li za sve ispitanike postoje podaci na varijabli, dok su relativne frekvencije u koloni **Valid Percent**, kao i kumulativne frekvencije u koloni **Cumulative Percent**, računate u odnosu na ukupan broj prikupljenih rezultata (n = 249).

Problemu podataka koji nedostaju treba u statističkim analizama podataka posvetiti dovoljnu pažnju. Podaci koji nedostaju mogu predstavljati pravu moru za istraživača u realnim situacijama, kada se statistički analizira nekoliko varijabli istovremeno a na svakoj od njih različitim ispitanicima nedostaju podaci. To ponekad može dovesti u pitanje i samu mogućnost primene statističke analize. Najbolje rešenje za problem nedostajanja podataka, kao i za sve probleme uopšte, jeste "učiniti sve da do tog problema ne dođe". "Učiniti sve" odnosi se prevashodno na ono što se može učiniti u fazama planiranja istraživanja i prikupljanja podataka radi svođenja broja podataka koji nedostaju na najmanju meru (npr., obezbediti valjane tehničke uslove tako da aparatura za merenje ispravno funkcioniše, obezbediti da se posmatranje ponašanja odvija u uslovima koji će omogućiti da se registruju sva ponašanja prema planu, postaviti pred ispitanika realne zahteve koje on može da ispuni, formulisati u upitnicima pitanja tako da ne provociraju kod ispitanika otpor da se na pitanja odgovori, podvući u uputstvu za ispitanika koliko je važno da odgovore na sva pitanja, pismeno zamoliti ispitanika na kraju upitnika da proveri da li je odgovorio na sva pitanja i slično). Ali, čak i kada se učini sve u fazama planiranja istraživanja i prikupljanja podataka dešava se da iz različitih razloga podaci na nekoj varijabli nedostaju za određeni broj jedinica posmatranja.

Dve ključne radnje koje treba preduzeti u vezi sa podacima koji nedostaju su:

1. Ustanoviti koliko ima jedinica posmatranja za koje nedostaju podaci na određenoj varijabli. Ovaj podatak u obliku učestalosti ili relativne učestalosti treba svakako da bude sastavni deo distribucije učestalosti, baš kao što je to prikazano na primeru distribucije iz ispisa programa SPSS;
2. Ustanoviti, ako je to moguće, postoji li neki jasan "složaj" ili "struktura" u podacima koji nedostaju, što bi moglo da sugeriše "mehanizam" koji je doveo do nedostajanja podataka ili "princip" nedostajanja podataka. Za otkrivanje "strukture", "mehanizama" ili "principa" nedostajanja podataka potrebno je da o

[14] U pojedinim starijim udžbenicima statistike (na primer, Dragićević, 2002) prednost se daje izboru neparnog broja za širinu grupnog intervala. Takav izbor je pojednostavljivao računske operacije u vreme kada su se statistička izračunavanja radila "peške", tj. bez pomoći računara. Budući da se statistička računanja u današnje vreme odvijaju gotovo isključivo korišćenjem računara, davanje prednosti neparnom broju pri izboru širine grupnog intervala smatramo nepotrebnim.

jedinicama posmatranja za koje nedostaju podaci na određenoj varijabli postoje informacije na drugim relevantnim varijablama. U pogledu strukture ili složajeva podataka koji nedostaju, odnosno mehanizama koji stoje u osnovi nedostajanja podataka, najčešće se u statističkoj literaturi navode sledeće tri mogućnosti: PSN – potpuno slučajno nedostajanje (engl. missing completely at random – MCAR), SN – slučajno nedostajanje (engl. missing at random – MAR) i NeSN – ne-slučajno nedostajanje (engl. not missing at random – NMAR). Struktura PSN podataka koji nedostaju znači da je verovatnoća da nema podatka na nekoj varijabli ista za sve jedinice posmatranja u uzorku. U tom slučaju se nedostajanje podataka na nekoj varijabli ne može dovesti u vezi ni sa jednim drugim obeležjem ispitanika (npr. da li je muško ili žensko, kakvog je obrazovnog statusa, kojeg je uzrasta i slično) niti sa samim podacima koji nedostaju. U ovoj strukturi mesta na kojima nedostaju podaci slučajno su raspoređena u celokupnoj matrici sa podacima. Jedini mehanizam koji možemo pretpostaviti da deluje je slučaj: kao da, na primer pri odgovaranju na upitnik CES-D, svaki ispitanik baca kocku i ne odgovara na upitnik ako padne neka određena strana kocke. Struktura SN zapravo ne znači potpuno slučajno već delimično sistematsko nedostajanje: u ovoj strukturi podataka koji nedostaju verovatnoća nedostajanja podataka može se dovesti u vezu sa drugim ispitivanim obeležjima jedinica posmatranja (npr. sa polom ili obrazovnim nivoom ispitanika) ali nije u vezi sa podacima koji nedostaju. U strukturi NeSN podataka koji nedostaju verovatnoća nedostajanja podataka može se dovesti u vezu sa varijablom na kojoj nedostaju podaci, tačnije sa samim podacima koji nedostaju. Dakle, nedostajanje podataka zavisi od informacija koje nam nisu na raspolaganju. Na primer, u takvoj situaciji se nalazimo ako podaci o uspehu u školi nedostaju samo za najlošije đake, ako nemamo podatke o prosečnim primanjima porodice samo za one koji su najbogatiji ili nemamo skor depresivnosti samo za one koji su najmanje depresivni. Određenje strukture NeSN deluje zbunjujuće: da bi se pokazalo da podaci koji nedostaju imaju NeSN strukturu trebalo bi dovesti u vezu podatke koji postoje na nekoj varijabli sa podacima koji nedostaju. Međutim, podatke koji nedostaju ne znamo! Sasvim pouzdano utvrđivanje strukture podataka koji nedostaju praktično nije moguće. Statističkim postupcima se mogu podržati pretpostavke o PSN i SN strukturi nedostajanja podataka, dok su za pretpostavku o NeSN strukturi potrebna konceptualna razmatranja i dobro poznavanje varijabli za koje su prikupljeni podaci. Povrh svega, tri moguće strukture podataka koji nedostaju nisu uzajamno isključive: moguće je da struktura podataka koji nedostaju u realnim situacijama bude određena kombinacija ovih mogućnosti.

Objašnjenje postupaka koji se primenjuju radi podržavanja određene pretpostavke o strukturi podataka koji nedostaju, kao i podroban prikaz postupanja sa podacima koji nedostaju izlazi izvan okvira ove glave teksta. Naime, razumevanje ovih postupaka podrazumeva statistička znanja koja će tek biti izložena u narednim glavama knjige. Stoga ćemo na ovom mestu dati samo osnovne napomene o postupanju sa podacima koji nedostaju. (Pošto savlada sadržaj koji je izložen u ovoj knjizi, zainteresovani čitalac podrobnije informacije o praktičnom tretmanu podataka koji nedostaju, prikazane na način razumljiv istraživačima u psihologiji i srodnim oblastima može naći, na primer, u Graham, Cumsille, & Elek-Fisk, 2003, Graham, 2009, Schlomer, Bauman, & Card, 2010 i Newman, 2014).

Kada na skupu svih varijabli koje se statistički analiziraju nedostaju podaci za veoma mali broj (u odnosu na ukupan broj) jedinica posmatranja (npr., manje od 5%) takve

jedinice posmatranja se mogu izostaviti iz uzorka, tj. iz daljih statističkih analiza. U situacijama kada bi izostavljanje određenoog broja jedinica posmatranja iz uzorka dovelo do problema u daljim statističkim analizama, a analiza strukture podataka koji nedostaju sugeriše da su pretpostavke o PSN (engl. MCAR) ili SN (engl. MAR) strukturi opravdane, mogu se primeniti različiti postupci simuliranja vrednosti, tj. "pripisivanja" (engl. imputation) datim jedinicama posmatranja određenih vrednosti na varijabli. Vrednosti na varijabli koje se pripisuju jedinicama posmatranja za koje ne postoje empirijski dobijeni podaci određuju se po određenim principima (i algoritmima zasnovanim na ovim principima) na osnovu raspoloživih podataka. Ukoliko, pak, analize strukture podataka koji nedostaju navode na mogućnost da je reč o NN (engl. NMAR) strukturi ovih podataka tada je potrebno pokušati sa prevođenjem NN strukture na SN strukturu. Ukoliko to nije moguće, pri zaključivanju na osnovu takvih podataka treba biti svestan toga da su su zaključci koji se donose o populaciji pristrasni i podložni velikim greškama. U svakom slučaju, tretmanu podataka koji nedostaju uvek treba posvetiti dovoljnu pažnju.

Pri analizi podataka koji nedostaju neophodno je uzeti u obzir celokupnu matricu podataka a ne samo podatke na pojedinačnim varijablama. Jedino tako se može ustanoviti koja pretpstavka o strukturi podataka koji nedostaju ima opravdanja.

U programu SPSS podaci koji nedostaju mogu se deklarisati na dva načina: ostavljanjem praznog mesta ili unošenjem neke vrednosti koja ne može predstavljati rezultat ili meru ispitanika na datoj varijabli (npr., 88 ili 999 ako su mogući rezultati na varijabli manji od 88, odnosno 999). Ostavljanje praznog mesta SPSS interpretira kao tzv. sistemsku vrednost koja nedostaje (engl. System missing value). Ukoliko se na mestu podatka koji nedostaje unese vrednost koju korisnik pri definisanju varijabli definiše kao oznaku za nedostajanje podatka, takva vrednost se zove korisnička vrednost koja nedostaje (engl. User missing value). Za analizu i tretman podataka koji nedostaju u programu SPSS za sada postoje dva specijalizovana modula **Missing Value Analysis** i **Multiple Imputation**.

Zapamtite:

- ✓ Radi pravljenja distribucije učestalosti potrebno je urediti podatke po veličini i potom odrediti najveći i najmanji rezultat, pogodan broj grupnih intervala i širinu grupnih intervala.(Određivanje širine i broja grupnih intervala ne treba uvek prepustiti automatskim opcijama statističkih programa).

- ✓ Distribucija učestalosti po pravilu treba da ima od 10 do 15 grupnih intervala, a ako je broj rezultata manji od 50 između 5 i 10 grupnih intervala. Ukoliko je broj različitih rezultata u skupu podataka mali (manji od 15) dovoljno je napraviti jediničnu raspodelu učestalosti.

- ✓ Za odabrani broj grupnih intervala, širina intervala približno se dobija deljenjem razlike najvećeg i najmanjeg rezultata sa brojem grupnih intervala.

- ✓ Širina grupnog intervala jednaka je razlici gornje i donje egzaktne granice grupnog intervala.

- ✓ Razlike između sukcesivnih donjih , kao i između sukcesivnih gornjih granica u celoj raspodeli, bilo da je reč o mernim ili egzaktnim granicama, treba da budu jednake širini intervala. Dakle, širine svih grupnih intervala u raspodeli trebalo bi da budu jednake.

- ✓ Za celobrojne podatke širina intervala treba da bude ceo broj, a za podatke iskazane decimalnim brojevima širina intervala može biti i decimalni broj sa istim brojem decimala kao što su i podaci.

- ✓ Distribucija učestalosti pored vrednosti varijable (ili grupnih intervala) i frekvencija pojedinih vrednosti (ili svih vrednosti u grupnom intervalu) treba da sadrzi sadrži relativne i kumulativne relativne frekvencije.

- ✓ Ključni kriterijumi kvaliteta raspodele učestalosti su preglednost i informativnost.

- ✓ Pri prikazivanju raspodele učestalosti, neophodno je prikazati i informaciju o učestalosti nedostajanja podataka.

4. Mere lokacije

Statističke mere lokacije, kao što i njihov naziv nagoveštava, ukazuju na određeno mesto, tj. lokaciju u raspodeli mera na kvantitativnoj varijabli. To može biti središnje mesto u raspodeli, mesto u raspodeli oko kojeg se nagomilavaju podaci, ali i bilo koje drugo mesto u raspodeli koje pruža korisne informacije o raspodeli. Statističke mere lokacije mogu služiti za opis raspodele mera u uzorku ali i u populaciji ukoliko raspolažemo merama na varijabli za sve članove populacije. Među merama lokacije u statističkom opisivanju uzorka (ili populacije) u psihologiji i srodnim oblastima najčešće se koriste kvantili, percentili i mere centralne tendencije.

Definisani su formalni uslovi koje neka statistička mera treba da zadovolji kako bi mogla biti smatrana merom lokacije. Statistička mera $l(X)$, koja opisuje distribuciju varijable, pri čemu je X slučajna varijabla sa distribucijom F, može biti mera lokacije ako ispunjava sledeće formalne uslove (prema Bickel & Lehmann, 1975):[15]

1. $l(X) \leq l(Y)$ kad god je Y stohastički veće nego X.

Y je stohastički veće nego X ako je za svaku vrednost x vrednost funkcije distribucije za Y manja od vrednosti funkcije distribucije za X. Drugim rečima, verovatnoća da varijabla Y uzme neku vrednost manju od x ili jednaku x manja je od verovatnoće da varijabla X uzme vrednost manju od x ili jednaku x. U tom slučaju su svi kvantili varijable Y veći od odgovarajućih kvantila varijable X. To naprosto znači da varijabla Y "poseduje" neki posmatrani atribut u većoj meri nego varijabla X, tj. da je distribucija varijable Y pomerena udesno (ka većim vrednostima nekog obeležja) u odnosu na distribuciju varijable X.

2. $l(b*X + a) = b*l(X) + a$, ako je $b > 0$.

Ovaj se uslov često zove lokacionom i skalnom ekvivarijantnošću mere lokacije (cf. Wilcox, 2005, str. 20). On jednostavno označava da dodavanje neke pozitivne konstante a na sve vrednosti varijable X treba da dovede do promene mere lokacije za iznos konstante a (lokaciona ekvivarijantnost), kao i da množenje svih vrednosti varijable X konstantom b treba da promeni meru lokacije za b puta (skalna ekvivarijantnost)

3. $l(-X) \leq -l(X)$

Ovaj uslov bismo mogli nazvati ekvivarijantnošću na refleksiju u odnosu na nultu tačku: on označava da ako varijablu X reflektujemo u odnosu na nultu tačku (tj. svim vrednostima na varijabli promenimo predznak) tada će i mera lokacije promeniti predznak.

Uslovima pod 2 i 3 obezbeđuje se da mera lokacije uzima vrednosti unutar raspona mogućih vrednosti varijable što je itekako smisleno: nije smisleno da mera lokacije ukazuje na lokaciju koja je izvan raspona vrednosti na varijabli. Iz prikazanih uslova slede mnoga važna matematička svojstva mera lokacije, svojstva koja su važna i sa aspekta primene ovih mera u opisu distribucije neke varijable.

[15] Opšta oznaka mere lokacije, $\mu(X)$, koja je korišćena u radu Bikela i Lemana, zamenjena je ovde oznakom $l(X)$ jer se oznakom μ u našem tekstu označava jedna konkretna mera lokacije – aritmetička sredina populacije.

Mere lokacije generalno su iskazane u mernim jedinicama, tj. u onim jedinicama u kojima su iskazani podaci na osnovu kojih se računaju.

Kvantili i percentili

Pojam kvantila teorijski smo definisali u prethodnoj glavi: kvantil p je najmanja vrednost slučajne varijable za koju važi da je verovatnoća da slučajna varijabla uzme neku vrednost manju od vrednosti kvantila p jednaka p. Kvantili kao mere lokacije mogu se koristiti i za opis uzorka na kvantitativnoj varijabli.

Kvantil uzorka, u oznaci Q_c, predstavlja vrednost na varijabli za koju važi sledeća jednakost:

$$\boxed{\mathrm{cp}_{Q_c} = c}$$

pri čemu je cp_{Q_c} kumulativna relativna frekvencija za vrednost kvantila Q_c. Dakle, kvantil uzorka predstavlja vrednost na kvantitativnoj varijabli koja je veća od određene proporcije svih rezultata dobijenih na uzorku. Kvantil je, prema tome, **mesto, tačka ili vrednost na varijabli ispod koje se nalazi određena proporcija (ili procenat) rezultata na varijabli**. Na primer, $Q_{0.33}$ predstavlja vrednost kojoj odgovara relativna kumulativna frekvencija od 0.33 ili 33%. Dakle, približno trećina ispitanika u uzorku (33%) ima rezultat koji je manji ili jednak vrednosti kvantila $Q_{0.33}$. Uočimo da kvantil uzorka predstavlja funkciju redoslednih statistika, tj. rezultata poređanih po veličini od najmanjeg do najvećeg. Dakle, za računanje bilo kojeg kvantila rezultate je neophodno sortirati po veličini, idući od najmanjeg rezultata u skupu.

Percentil (engl. Percentile), u oznaci P_P, jeste tačka, tj. vrednost na varijabli za koju važi sledeća jednakost:

$$\boxed{100 * \mathrm{cp}_{P_P} = P}$$

pri čemu je cp_{P_P} kumulativna relativna frekvencija za vrednost P_P, a P u indeksu predstavlja celobrojni procenat od 1 do 99.

Dakle, percentil P_P je vrednost na varijabli ispod koje se, na osnovu rezultata koje jedinice posmatranja imaju na varijabli v, nalazi odredjeni celobrojni postotak (P) jedinica posmatranja.[16] O kojem celobrojnom procentu jedinica posmatranja je reč označeno je brojem koji stoji umesto slova P u indeksu oznake za percentil. Prema tome, oznaka P_{30} označava da je reč o 30% jedinica posmatranja koje imaju rezultat manji ili jednak vrednosti P_{30}. Dakle, ako je relativna kumulativna frekvencija koja odgovara vrednosti x_k jednaka 30% onda je P_{30} (čita se kao "percentil trideset") jednak vrednosti x_k. Onda kada se pojam percentila ograniči na celobrojne vrednosti procenta u indeksu percentila, tada vrednosti percentila dele distribuciju rezultata poređanih po veličini na 100 jednakih delova, tj. na sto delova sa jednakom učestalošću. Ovi delovi ograničeni su na skali varijable vrednostima percentila od P_1 do P_{99}, pri čemu P_1 predstavlja gornju granicu najnižeg od sto delova, a P_{99} donju granicu najvišeg od sto delova distribucije.

[16] U statističkim udžbenicima se ne pravi uvek ovakva distinkcija između percentila i kvantila, te se percentili potpuno poistovećuju sa kvantilima, pri čemu se u indeksu kvantila koriste proporcije a u indeksu percentila procenti. Verujemo da je smisleno termin percentil ograničiti, kako smo to ovde učinili, na mere lokacije koje upućuju na lokacije u distribuciji rezultata, tj. na vrednosti na varijabli ispod kojih je celobrojan postotak rezultata.

Percentili se mogu računati kako iz jediničnih raspodela učestalosti tako i iz raspodela sa grupnim intervalima. Računanje percentila iz raspodela sa grupnim intervalima primenjivano je u vreme kada se statistička analiza podataka odvijala bez pomoći računarskih programa.[17] Čitaoci ove knjige percentile će uobičajeno računati koristeći odgovarajuće računarske programe. Algoritmi ovih programa uobičajeno percentile računaju na osnovu jedinične raspodele učestalosti.

Računanje percentila P_P iz jedinične raspodele učestalosti odvija se na sledeći način:

– Pronalazi se mera u jediničnoj raspodeli učestalosti čija je relativna kumulativna frekvencija u procentima veća od vrednosti izraza $\frac{(n+1)P}{n}$, pri čemu je n ukupan broj rezultata, a *P* procenat za traženi percentil. Drugim rečima, pregledom jedinične distribucije učestalosti pronalazi se mera x_k za koju je $100*cp_k > \frac{(n+1)P}{n}$, pri čemu je $100*cp_k$ relativna kumulativna frekvencija u procentima za meru x_k;

– Računa se vrednost izraza $\frac{(n+1)P}{n} - 100*cp_{(k-1)}$, pri čemu je $100*cp_{(k-1)}$ relativna kumulativna frekvencija za prvu meru manju od mere x_k;

– Ukoliko je vrednost izraza $\frac{(n+1)P}{n} - 100*cp_{(k-1)}$ veća ili jednaka vrednosti izraza $\frac{100}{n}$, tada je $P_P = x_k$;

– Ukoliko je vrednost izraza $\frac{(n+1)P}{n} - 100*cp_{(k-1)}$ manja od vrednosti izraza $\frac{100}{n}$, tada je $P_P = \left\{1 - \left[\frac{(n+1)P}{100} - cf_{(k-1)}\right]\right\}x_{(k-1)} + \left[\frac{(n+1)P}{100} - cf_{(k-1)}\right]x_k$ pri čemu je $cf_{(k-1)}$ kumulativna frekvencija za prvu meru manju od mere x_k.

Postupak koji je ovde prikazan odgovara algoritmu prema kojem se percentili računaju u programu SPSS.

Egzaktne vrednosti percentila ne moraju biti jednake nijednoj konkretnoj meri u distribuciji mera, ove vrednosti dobijene računski mogu se naći između dveju konkretnih mera u raspodeli.

Pojedini kvantili, poput percentila, imaju posebna imena:

Decili su percentili sa procentima u desetinama u indeksu: prvi decil je percentil 10 (P_{10}), drugi decil je percentil dvadeset (P_{20}), treći decil je percentil trideset (P_{30}) i tako redom, sve do devetog decila koji predstavlja percentil devedeset (P_{90}). Dakle, decili na izvestan način "seku" distribuciju varijable na 10 jednakih delova, tj. dele raspon vrednosti na varijabli poređanih po veličini u 10 intervala tako da su relativne učestalosti u svakom intervalu jednake i iznose 10%.

Kvintili su percentili koji dele distribuciju rezultata poređanih po veličini na pet jednakih delova prema učestalosti: prvi kvintil je percentil dvadeset (P_{20}), drugi kvintil predstavlja

[17] Zainteresovani čitalac može opis ovog postupka videti u Dragićević, 2002, str.

percentil četrdeset (P_{40}) , treći kvintil je percentil šezdeset (P_{60}), a četvrti kvintil percentil osamdeset (P_{80}).

Kvartili su percentili koji dele distribuciju rezultata poređanih po veličini na četiri jednaka dela prema učestalosti: prvi kvartil je percentil dvadeset pet (P_{25}) i uobičajeno se označava sa Q_1, drugi kvartil (Q_2) je percentil pedeset (P_{50})(drugi kvartil) i $Q_3 = P_{75}$ (treći kvartil).

Dakle, percentili dele raspodelu na 100 jednakih delova prema učestalosti, decili na 10 jednakih delova, kvintili na pet, a kvartili na četiri jednaka dela.

Uočimo da su percentili, decili, kvintili i kvartili zapravo specijalni slučajevi kvantila uzorka. Ipak, uobičajeno je da se u indeksu oznake kvantila relativna učestalost beleži u obliku proprcije (npr. $Q_{0.2}$), dok se u indeksu oznake za percentil odgovarajuća relativna frekvencija beleži u obliku procenta (P_{20}). Isto tako, treba uočiti da je u indeksu oznake kvartila oznaka redosleda kvartila (1, 2 ili 3) a ne odgovarajuća relativna frekvencija. Na primer kvantil 0.25 mogli bismo označiti na tri načina: $Q_{0.25}$, Q_1 i P_{25}.

Percentili u psihologiji i srodnim oblastima imaju široku upotrebu. Pored toga što se koriste za računanje drugih statističkih mera (na primer interkvartilnog raspona o kojem će biti reči u nastavku teksta u ovoj glavi) i za statističko opisivanje uzorka u pogledu neke kvantitativne varijable, percentili se veoma mnogo koriste pri pravljenju tzv. percentilnih normi za psihološke merne instrumente (cf. Fajgelj, **). [18]Percentilne norme se koriste za tumačenje značenja individualnog rezultata ispitanika na testovima, posebno onda kada raspodela rezultata na testu ne samo da nije normalna, već je izrazito asimetrična. Tako, na primer, ako rezultat određenog ispitanika na nekom testu odgovara vrednosti percentila 70, to onda znači da ispitanik ima veći rezultat od 70% ispitanika normativnog uzorka, tj. grupe koja je poslužila za pravljenje normi. Budući da se normativni uzorak (uzorak koji služi za pravljenje normi) po pravilu bira tako da dobro reprezentuje određenu populaciju, smatra se da ispitanik čiji rezultat odgovara percentilu sedamdeset ima veći rezultat na datom testu od približno 70% članova populacije kojoj ispitanik pripada.

Mere centralne tendencije

Podaci dobijeni merenjem kvantitativnih varijabli uobičajeno pokazuju tendenciju grupisanja oko neke vrednosti. Statističke mere lokacije koje su izraz ove tendencije spadaju u mere centralne tendencije. Mere centralne tendencije, kako njihovo ime sugeriše, trebalo bi da pokazuju srednji, prosečan, tipičan rezultat ili – ako se radi o raspodelama specifičnog tipa – rezultat koji se najčešće javlja. U glavi o osnovnim pojmovima teorije verovatnoće već smo definisali jednu od ovih statističkih mera kao parametar, tj. kao statističku meru koja se odnosi na populaciju. Naime, pri definisanju očekivane vrednosti slučajne varijable, napomenuli smo da se očekivanom vrednošću u statistici definiše aritmetička sredina populacije. Aritmetička sredina, bilo da se računa za populaciju ili uzorak, spada u mere centralne tendencije. Pri definisanju statističkih mera centralne tendencije u ovoj glavi prikazaćemo ove statističke mere isključivo kao statistike, tj. statističke mere uzorka.

Aritmetička sredina (engl. Arithmetic Mean ili samo Mean)

[18] S. Fajgelj pored termina percentilne norme koristi i termin fraktilne norme kao objedinjen naziv za percentilne, decilne i kvartilne norme (cf. Fajgelj, **, str.**). Međutim, budući da su decili i kvartili samo posebni slučajevi percentila, mi ćemo u ovom tekstu koristiti samo termin percentilne norme.

Aritmetička sredina uzorka, u oznaci M (od engleskog Mean = sredina), definiše se na sledeći način:[19]

$$M = \frac{\sum_{i=1}^{n} x_i}{n}$$

pri čemu je x_i rezultat za jedinicu posmatranja e_i (i = 1, ...n), a n je veličina uzorka, tj. broj rezultata. Dakle, aritmetička sredina predstavlja količnik zbira svih rezultata i broja rezultata. Na primer, aritmetičku sredinu rezultata 3, 4, 5, 5, 5, 5, 6, 7 dobili bismo na sledeći način:

$$M = (3 + 4 + 5 + 5 + 5 + 5 + 6 + 7)/8 = 40/8 = 5$$

Naravno, rezultati ne moraju biti sortirani, tj. poređani po veličini za računanje aritmetičke sredine. Da smo ove iste rezultate prikazali kao niz 4, 5, 5, 6, 5, 5, 7, 3, rezultat primene obrasca bio bi, zbog komutativnosti operacije sabiranja, isti:

$$M = (4 + 5 + 5 + 6 + 5 + 5 + 7 + 3)/8 = 40/8 = 5$$

Ovakvu konceptualizaciju aritmetičke sredine pojedini autori zovu „socijalističkom“ (cf. Watier, Lamontagne, & Chartie, 2011): ako pogledamo obrazac za aritmetičku sredinu vidimo da se računanje aritmetičke sredine u stvari svodi na računanje broja koji se dobija kada se suma distribucije podjednako raspodeli na n rezultata. Na taj način aritmetička sredina predstavlja jednu vrednost koja je zajednička za sve rezultate.

Istakli smo već da teorijski aritmetička sredina zapravo predstavlja očekivanu vrednost. To se iz prethodnog obrasca ne vidi sasvim jasno. Jasnije ćemo vezu između aritmetičke sredine i očekivane vrednosti iz teorije verovatnoće videti iz sledećeg obrasca za aritmetičku sredinu uzorka:

$$M = \sum_{k} x_k \frac{f_k}{n}$$

pri čemu je $\sum_{k} f_k = n$.

U ovom obrascu x_k predstavlja svaki <u>različiti</u> rezultat koji se pojavljuje u skupu rezultata na varijabli, a oznaka k ispod operatora sume označava da se sabiraju proizvodi x_k sa f_k/n za sve k, tj. za sve različite rezultate. Dakle, sabirci se dobijaju tako što se svaki različiti rezultat ponderiše se relativnom učestalošću tog rezultata. Ukoliko se prisetimo da se, pod uslovom da je broj rezultata veliki, statistički verovatnoća rezultata određuje njegovom relativnom učestalšću, tada je sličnost ovog obrasca i obrasca kojim smo definisali očekivanu vrednost za diskretne slučajne varijable (obrazac **) očigledna.[20]

Kada obrazac primenimo na podatke iz prethodnog primera uočavamo da među ukupno 8 rezultata (n = 8) postoji pet različitih rezultata (3, 4, 5, 6 i 7), te k ide od 1 do 5. Rezultati 3, 4, 6 i 7 se pojavljuju po jedanput, a rezultat 5 se pojavljuje četiri puta. Dakle, primenom prethodnog obrasca aritmetičku sredinu bismo mogli izračunati na sledeći način:

M = [3 * (1/8) + 4 * (1/8) + 5 * (4/8) + 6 * (1/8) + 7 * (1/8)] = 3 * 0.125 + 4 * 0.125 + 5 * 0.5 + 6 * 0.125 + 7 * 0.125 = 0.375 + 0.5 + 2.5 + 0.75 + 0.875 = 5.

---

[19] Veoma često se aritmetička sredina označava oznakom $\overline{X}$, a u pojedinim statističkim udžbenicima na našem jeziku i oznakom AS. Mi ćemo, jednostavnosti radi, u ovoj knjizi koristiti samo oznaku M.

[20] Bez obzira na to što varijabla može biti teorijski kontinuirana, podaci sa uzorka uvek su diskretne vrednosti te obrazac za aritmetičku sredinu uzorka liči na obrazac za očekivanu vrednost diskretne slučajne varijable.

Istovetnost obrasca ** sa obrascem ** lako je uočiti ukoliko se obrazac ** napiše u sledećem obliku:

$$M = \sum_{i=1}^{n} x_i \frac{1}{n}$$

pri čemu je x_i svaki pojedini rezultat na varijabli, a 1/n relativna frekvencija svakog rezultata. Budući da neki od rezultata mogu imati iste vrednosti pri primeni ovog obrasca svaki pojedini rezultat (čak i ako ima istu vrednost kao neki drugi rezultat) množi se sa 1/n). Ako nekoliko rezultata imaju istu vrednost tada je zbir umnožaka svakog od tih rezultata sa 1/n istovetan sa proizvodom vrednosti kojoj su jednaki ti rezultati i njihove relativne učestalosti. Na primer, ako je $x_1 = x_2 = x_3 = x_k$, tada je $x_1 * 1/n + x_2 * 1/n + x_3 * 1/n = x_k (1/n + 1/n + 1/n) = x_k * 3/n$.

Aritmetička sredina kao statistička mera ima određena matematička svojstva od kojih ćemo prikazati ona koja je potrebno poznavati kako bi se na pravi način razumela informacija koja se dobija kada se izračuna vrednost aritmetičke sredine za datu distribuciju rezultata:

- Ako je $M(v_1)$ aritmetička sredina varijable v_1, i ako se rezultati na varijabli v_1 linearno transformišu tako da se dobije varijabla $v_2 = b + av_1$ (na svaki rezultat na varijabli v_1 se doda tzv. aditivna konstanta *b* i svaki rezultat se pomnoži tzv. multiplikativnom konstantom *b*), tada je $M(v_2) = a + bM(v_1)$. Dakle, dodavanje konstante na svaki rezultat menja aritmetičku sredinu za tu konstantu, a množenje svakog rezultata konstantom menja aritmetičku sredinu za konstantu puta. Ovo svojstvo aritmetička sredina deli sa svim merama lokacije, a naziva se lokaciona i skalna ekvivarijantnost.

- Ako se od svakog rezultata u distribuciji oduzme aritmetička sredina tih rezultata, zbir tako dobijenih razlika jednak je nuli:

$$\sum_{i=1}^{n} (x_i - M) = 0$$

Ukoliko, na primer, na rezultatima iz primera u kojem smo pokazali računanje aritmetičke sredine izračunamo zbir odstupanja rezultata od aritmetičke sredine dobijamo:

(3 - 5) + (4 - 5) + (5 - 5)+ (5 - 5) + (5 - 5) + (5 - 5) + (6 - 5) + (7 - 5) = (-2) + (- 1) + 0 +0 + 0 + 0 + 1 + 2 = 0.

Uočimo da je zbir negativnih odstupanja od aritmetičke sredine jednak zbiru pozitivnih odstupanja od aritmetičke sredine. Dakle, aritmetička sredina predstavlja težište distribucije rezultata, tačku ravnoteže distribucije odstupanja rezultata u jednu i u drugu stranu od tačke koja predstavlja aritmetičku sredinu. Odavde sledi da aritmetička sredina ne može da bude manja od najnižeg, niti veća od najvišeg rezultata u distribuciji. Isto tako, na osnovu ovoga svojstva uočavamo da aritmetička sredina ne mora biti na sredini distribucije: aritmetička sredina jeste na sredini distribucije samo ako su rezultati sa jedne njene strane u distribuciji podjednako udaljeni od nje kao i rezultati sa njene druge strane. To jednostavno znači da je aritmetička sredina na sredini distribucije samo ako je distribucija simetrična.

- Zbir kvadriranih odstupanja rezultata od njihove aritmetičke sredine manji je nego od bilo koje druge vrednosti:

$$\sum_{i=1}^{n}(x_i - M)^2 = \text{minimum} \ < \sum_{i=1}^{n}(x_i - x_0)^2 \,|\, x_0 \neq M$$

(Vertikalna crta pre izraza $x_0 \neq M$ označava uslov i čita se "pod uslovom da..." ili "ako...").

Dakle, ako bismo računali zbir kvadriranih odstupanja rezultata od bilo koje vrednosti koja nije jednaka aritmetičkoj sredini tih rezultata, zbir tih kvadriranih odstupanja bio bi veći. Ovo je veoma važno svojstvo, jer pokazuje matematički princip po kojem je izvedena aritmetička sredina kao mera centralne tendencije. Reč je o tzv. **principu najmanjih kvadrata** (engl. Least squares): ako u nizu rezultata tražimo vrednost takvu da je zbir kvadriranih odstupanja rezultata od te vrednosti najmanji, tj. manji nego zbir kvadriranih odstupanja od bilo koje druge vrednosti dobićemo vrednost aritmetičke sredine. Na primer, ukoliko na rezultatima iz primera u kojem smo pokazali računanje aritmetičke sredine izračunamo zbir kvadriranih odstupanja rezultata od aritmetičke sredine dobijamo sledeći zbir:
$(3 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (6 - 5)^2 + (7 - 5)^2 = 4 + 1 + 0 + 0 + 0 + 0 + 1 + 4 = 10$.
Ukoliko izračunamo odgovarajući zbir u odnosu na vrednost različitu od aritmetičke sredine, na primer u odnosu na vrednost 6, dobijamo nužno veću vrednost zbira:
$(3 - 6)^2 + (4 - 6)^2 + (5 - 6)^2 + (5 - 6)^2 + (5 - 6)^2 + (5 - 6)^2 + (6 - 6)^2 + (7 - 6)^2 = 9 + 4 + 1 + 1 + 1 + 1 + 0 + 1 = 18$.
Princip najmanjih kvadrata mogao bi se jednostavno predstaviti na sledeći način: ukoliko bismo napravili beskonačan broj ovih zbirova kvadrata uzimajući kao vrednost u odnosu na koju posmatramo kvadrirana odstupanja bilo koju moguću vrednost u intervalu od najmanjeg do najvećeg rezultata ustanovili bismo da je zbir kvadriranih odstupanja najmanji kada kvadrirana odstupanja računamo u odnosu na vrednost aritmetičke sredine. Naravno, da je to tako znamo na osnovu matematičkog izvođenja a ne na osnovu računanja jer, kao što smo već istakli, broj takvih zbirova kvadrata je neprebrojivo beskonačan (vema zanimljiva demonstracija algebarske i geometrijske forme konceptualizacije aritmetičke sredine pomoću principa najmanjih kvadrata može se naći na linkovima koji su dati u Watier, Lamontagne, & Chartie, 2011). Ovu osobinu aritmetičke sredine možemo da posmatramo i na sledeći način: ako bismo hteli da pogodimo meru na nekoj varijabli za svakog ispitanika u uzorku a da pritom zbir kvadriranih grešaka u takvom pogađanju bude najmanji mogući, onda je najbolje da svakom ispitaniku kao "rezultat" pripišemo aritmetičku sredinu svih rezultata. (Naravno, pod uslovom da znamo kolika je aritmetička sredina svih rezultata).

Princip najmanjih kvadrata, kao osnova za definisanje aritmetičke sredine kao mere centralne tendencije, dovodi do toga da aritmetička sredina ima svojih prednosti i nedostataka u odnosu na druge mere centralne tendencije. Prednosti aritmetičke sredine nad drugim merama centralne tendencije dolaze do izražaja tek onda kada je distribucija varijable u populaciji normalna te kada su posledično empirijske distribucije rezultata dobijene na uzorcima iz populacije koja ima normalnu raspodelu. Ukoliko je distribucija varijable u populaciji normalna tada je aritmetička sredina populacije parametar koji najbolje opisuje centralnu tendenciju vrednosti na varijabli, najčešću i tipičnu vrednost. Budući da se veruje, kako smo to već istakli u glavi o teoriji verovatnoće, da je normalna raspodela adekvatan teorijski model za opis raspodele mera u populaciji na mnogim

varijablama koje se koriste u psihologiji i srodnim oblastima, ne čudi što je aritmetička sredina najčešće korišćena mera centralne tendencije. Dodatan teorijski razlog popularnosti aritmetičke sredine jeste centralna granična teorema, jedna od napoznatijih teorema iz teorije verovatnoće. Prema centralnoj graničnoj teoremi aritmetička sredina kao statistik ima, pod određenim uslovima, neka pogodna svojstva koja pojednostavljuju statističku analizu podataka (centralnu graničnu teoremu ćemo predstaviti u glavi **).

Međutim, aritmetička sredina kao mera centralne tendencije ima i svojih nedostataka. Jedan od najvećih nedostataka aritmetičke sredine je njena *nerezistentnost*, tj. osetljivost na autlajere ili iznimke u rezultatima. Iznimak ili autlajer (engl. outlier) u slučaju jedne varijable je rezultat koji znatno odstupa od glavnine podataka, dakle rezultat koji je „jako daleko“od glavnine rezultata (preciznije određenje autlajera i postupke za otkrivanje autlajera u podacima prikazaćemo u glavi **). Budući da je aritmetička sredina vrednost koja je najbliža po principu najmanjih kvadrata svim rezultatima, rezultati koji su veoma udaljeni od ostalih podataka „vuku“ aritmetičku sredinu k sebi (jer se na taj način obezbeđuje da zbir kvadriranih odstupanja bude minimalan) te se može desiti da se zbog dejstva autlajera i aritmetička sredina „udalji“ od centra, tj. glavnine podataka. Ilustrovaćemo nerezistentnost aritmetičke sredine malom izmenom u podacima iz primera koji smo koristili za prikaz njenog računanja. Zamislimo da smo pri unosu podataka pogrešno uneli samo jedan od podataka te umesto skupa rezultata:

3, 4, 5, 5, 5, 5, 6, 7

imamo sledeći skup:

3, 4, 5, 5, 5, 5, 6, 77

Dakle, dva skupa rezultata razlikuju se samo u jednom rezultatu: umesto rezultata 7 u drugom nizu imamo rezultat 77 koji je veoma daleko od glavnine podataka. Aritmetička sredina prvog skupa je 5, dok je aritmetička sredina drugog skupa podataka 13. 75 i, ne samo da ne upućuje na tipičan rezultat već je znatno udaljena od tačke oko koje se grupiše većina rezultata.

Aritmetička sredina najbolje funkcioniše kao mera centralne tendencije ako je raspodela rezultata normalna ili barem simetrična sa relativno malim brojem rezultata na krajevima raspodele. U svakom slučaju, pri korišćenju aritmetičke sredine kao mere centralne tendencije uvek treba vrednost aritmetičke sredine posmatrati uporedo sa razgledanjem same distribucije učestalosti kako bi se uočilo da li aritmetička sredina zaista dobro reprezentuje skup podataka. Kod tzv. izrazito asimetričnih raspodela (raspodela sa malim brojem rezultata na jednom kraju vrednosti na varijabli, a velikim brojem rezultata na drugom kraju) treba biti veoma uzdržan u korišćenju aritmetičke sredine kao pokazatelja centralne tendencije. Najizrazitije zloupotrebe statistike u prikazivanju standarda stnovništva u jednoj državi zasnivaju se upravo na korišćenju aritmetičke sredine plata kao pokazatelja tipičnih primanja zaposlenih: izolovana informacija o tome da prosečna plata u nekoj zemlji u kojoj je raspodela plata izrazito asimetrična (veliki broj zaposlenih ima niske plate, a manjina veoma visoke) iznosi 45 hiljada dinara predstavlja upravo zloupotrebu statistike. Naime, manjina sa visokim platama “vuče aritmetičku sredinu k sebi” i na osnovu toga se stiče utisak da većina radnika u toj zemlji pristojno zarađuje, što je daleko od istine. Naime, u skladu sa “socijalističkom” konceptualizacijom informacija o aritmetičkoj sredini nam pokazuje koliko bi svaki zaposleni imao platu kada bi suma svih novaca za plate bila raspodeljena podjednako na sve zaposlene. Onda kada je empirijska, tj. realna distribucija plata izrazito asimetrična aritmetička sredina stvara lažnu predstavu o raspodeli plata. Mnogo valjanija statistička informacija u tom slučaju bilo bi navođenje minimalne i maksimalne plate, kao i kvartila po platama. U suprotnom, dobija se karikatura aritmetičke sredine kao mere centralne tendencije, baš kao u onoj narodnoj

doskočici koja nije daleko od istine barem kada su zemlje poput naše u pitanju: *"Neko jede kupus, neko jede meso, ali u proseku jedemo sarmu"*!

Kada se koristi kao deskriptivna statistička mera, aritmetička sredina se uobičajeno zaokružuje na dve decimale.

Zapamtite:

- ✓ Aritmetičku sredinu je najbolje koristiti kao meru centralne tendencije u sledećim uslovima:
  - raspodela rezultata je normalna ili, barem, približno simetrična;
  - raspodela ne sme imati iznimke – autlajere (engl. outliers), tj. vrednosti koje ekstremno odstupaju od ostalih rezultata;
- ✓ Dodavanje konstante na svaki rezultat na varijabli menja aritmetičku sredinu za vrednost te aditivne konstante, a množenje svakog rezultata na varijabli nekom konstantom menja aritmetičku sredinu za tu multiplikativnu konstantu puta;
- ✓ Zbir odstupanja svih mera u skupu od njihove aritmetičke sredine jednak je nuli.
- ✓ Zbir kvadriranih odstupanja rezultata u skupu od aritmetičke sredine tog skupa manji je od zbira kvadriranih odstupanja tih rezultata od bilo koje druge vrednosti.

Geometrijska sredina (engl. Geometric Mean)

Geometrijska sredina uzorka, u oznaci G, definiše se na sledeći način:

$$G = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}} = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$$

Dakle, geometrijska sredina predstavlja n-ti koren iz proizvoda dobijenog množenjem n rezultata, pri čemu su svi rezultati veći od nule. Za računanje geometrijske sredine većeg broja rezultata jednostavnije je koristiti obrazac koji se dobija logaritmovanjem izraza sa obe strane jednakosti u obrascu za geometrijsku sredinu:[21]

$$\log G = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

[21] Izraz sa desne strane jednakosti dobija se na osnovu pravila da je logaritam proizvoda brojeva jednak zbiru logaritama tih brojeva i pravila po kome je $\log_a(x^r) = r*\log_a x$. U ovom slučaju r = 1/n.

Odavde se antilogaritmovanjem dobija geometrijska sredina:[22]

$$G = \text{anti}\log\left(\frac{\sum_{i=1}^{n}\log x_i}{n}\right)$$

Na osnovu skupa rezultata koji smo koristili u prikazu računanja aritmetičke sredine:

3, 4, 5, 5, 5, 5, 6, 7

izračunali bismo geometrijsku sredinu na sledeći način:
G = antilog [(log 3 + log 4 + log 5 + log 5 + log 5 + log 5 + log 6 + log 7)/8] = antilog (5.5/8) = antilog 0.6875 = $10^{0.6875}$ = 4.87.

Isti rezultat bismo dobili i primenom obrasca **:

G = $(3 * 4 * 5 * 5 * 5 * 5)^{1/8}$ = $315\,000^{0.125}$ = 4.87.
Vidimo da je za ove podatke geometrijska sredina ima nešto nižu vrednost od aritmetičke sredine. U opštem slučaju geometrijska sredina je za iste podatke manja ili jednaka aritmetičkoj sredini.
Važno je uočiti i da geometrijska sredina ima smisla kao mera samo za podatke koji su veći od nule.

U psihologiji se geometrijska sredina srazmerno retko koristi i to, pre svega, u oblasti psihofizike. Na primer, geometrijska sredina se koristi u proveravanju Fehnerovog zakona: ako je, kako tvrdi Fehnerov zakon, intenzitet draži u logaritamskom odnosu sa intenzitetom oseta onda bi stimulus koji je označen kao senzorno, tj. čulno na sredini između stimulusa S_1 i S_2 trebalo da bude jednak njihovoj geometrijskoj sredini. Geometrijska sredina se veoma često koristi u ekonomskim istraživanjima za prikazivanje dinamike pojava. Budući da je osetljiva na razlike u odnosima, a mnogi ekonomski indeksi predstavljaju odnose aktuelnih i ranijih ili baznih vrednosti, geometrijska sredina se često koristi kao srednja vrednost tzv. indeksnih brojeva (cf. Žižić i sar, 2000).

Harmonijska sredina (engl. Harmonic Mean)

Harmonijska sredina vrednosti x_i, i = 1,...,n, u oznaci H, predstavlja recipročnu vrednost aritmetičke sredine recipročnih vrednosti od x_i:

$$H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Uočimo da bi aritmetička sredina recipročnih vrednosti od od x_i bila jednaka $\frac{\sum_{i=1}^{n}\frac{1}{x_i}}{n}$, te je $\frac{1}{\frac{\sum_{i=1}^{n}\frac{1}{x_i}}{n}}$ zapravo harmonijska sredina vrednosti x_i.

[22] O antilogaritmovanju se može videti u Matematičkom pojmovniku ** pod odrednicom **Logaritamska funkcija**.

Budući da se za računanje harmonijske sredine podataka koriste njihove recipročne vrednosti, harmonijsku sredinu ima smisla računati samo za podatke koji su različiti od nule.
Na osnovu skupa rezultata koji smo koristili u prikazu računanja aritmetičke sredine:
3, 4, 5, 5, 5, 5, 6, 7
izračunali bismo harmonijsku sredinu na sledeći način:
H = 8 / [(1/3) + (1/4) + (1/5) + (1/5) + (1/5) + (1/5) + (1/6) + (1/7)] = 4.73.
Za iste podatke harmonijska sredina je manja ili jednaka geometrijskoj sredini, odnosno manja ili jednaka aritmetičkoj sredini.
U psihologiji i srodnim oblastima harmonijska sredina se relativno retko sreće. Uobičajeno se koristi za određena pomoćna računanja u okviru tzv. statističkih testova višestrukih poređenja.
Harmonijska sredina se u oblasti ekonomije koristi za računanje srednjeg indeksa za varijable koje su iskazane tzv. recipročnim pokazateljima (npr., brzina opticaja novca kao recipročna vrednost vremena potrebnog za opticaj novca, cf. Žižić i sar., 2000).
Za računanje prosečne brzine jednostavnije je koristiti harmonijsku nego aritmetičku sredinu. Na primer, ako razdaljinu od 180 km pređemo automobilom vozeći 90 km/h a zatim pređemo narednih 180 km vozeći brzinom od 60 km/h prosečnu brzinu ne možemo dobiti kao aritmetičku sredinu dve brzine, tj. kao [(90 + 60) / 2] = 75 km/h. Zapravo, ako smo 360 km prošli za 5 sati onda je prosečna brzina 360/5 = 72 km/h. Takav rezultat bismo dobili ako bismo svaku brzinu pomnožiti brojem sati vožnje pa tek onda računali aritmetičku sredinu na sledeći način: [(90 * 2 + 60 * 3) / 5] = 72 km/h.
Ako bismo, pak, na izračunali harmonijsku sredinu dveju brzina dobili bismo odmah traženi rezultat:

$$2 / (1/90 + 1/60) = 72.$$

Medijana (engl. Median)

Medijana je središnje mesto u raspodeli, vrednost iznad i ispod koje je podjednak broj (po 50%) rezultata.
Medijana uzorka, u oznaci Mdn, može se definisati na sledeći način:

a) ako je broj rezultata u uzorku neparan, $Mdn = X_{(m)}$; $m = \frac{n+1}{2}$;

b) ako je broj rezultata u uzorku paran, $Mdn = \frac{1}{2}(X_{(m)} + X_{(m+1)})$; $m = [\frac{n+1}{2}]$

U gornjim obrascima za medijanu $X_{(m)}$ i $X_{(m+1)}$ predstavljaju redosledne statistike, tj. rezultate koji su po veličini na mestima m i m + 1 idući od najmanjeg ka najvećem. Zagrada u izrazu $m = [\frac{n+1}{2}]$ označava činjenicu da se kao vrednost m u tom slučaju koristi samo celobrojni deo tog izraza.

Jednostavnije rečeno, iz n sortiranih vrednosti na varijabli, tj. n mera uređenih po veličini mesto na kojem je medijana, u oznaci mMdn, određuje se po obrascu:

$$\boxed{mMdn = \frac{n+1}{2}}$$

Prema tome, ako je broj rezultata neparan medijana odgovara rezultatu koji je na m-tom mestu po veličini. Ako je, pak, broj rezultata paran m je decimalan broj pa medijana odgovara sredini između dva rezultata: onom koji je najbliži s leve strane mestu koje odgovara decimalnoj vrednosti m i onom rezultatu koji je tom mestu najbliži sa desne strane. Medijana je, dakle, centralno mesto u raspodeli, tj. vrednost ispod i iznad koje se nalazi jednak broj rezultata.

Ukoliko se prisetimo definicija kvantila i percentila lako ćemo uočiti da je medijana isto što i kvantil 0.5 ($Q_{0.5}$), percentil 50 (P_{50}) ili drugi kvartil (Q_2).

Iz niza sortiranih rezultata

3, 4, 5, 5, 5, 5, 6, 7

medijanu bismo, budući da imamo 8, tj. paran broj rezultata, mogli odrediti na sledeće alternativne načine:

a) Prvo bismo izračunali m, m = [(8 + 1)/2] = 4. Budući da je m = 4, četrvti rezultat po veličini je 5, tj. $X_{(4)} = 5$. Rezultat koji je na mestu m + 1, tj. peti rezultat po veličini, takođe je 5. Dakle, $X_{(5)} = 5$. Medijana je, prema tome, aritmetička sredina četvrtog i petog rezultata, tj. aritmetička sredina vrednosti redoslednih statistika $X_{(4)}$ i $X_{(5)}$: Mdn = (5 + 5)/2 = 5.
b) Prvo bismo odredili mesto na kojem je medijana: mMdn = (8 + 1) / 2 = 4.5. Očigledno, ako su rezultati sortirani od najmanjeg do najvećeg, medijana je na mestu 4.5, tj. na sredini između četvrtog i petog rezultata. Dakle, opet bi bilo: Mdn = (5 + 5)/2 = 5.

Pogledajmo sada šta će se desiti sa medijanom, ako zamislimo, kao što smo to uradili pri prikazu aritmetičke sredine, da smo pri unosu podataka pogrešno uneli samo jedan od podataka te umesto skupa rezultata:

3, 4, 5, 5, 5, 5, 6, 7

imamo sledeći skup:

3, 4, 5, 5, 5, 5, 6, 77.

Očigledno u potonjem nizu postoji jedan iznimak, tj. rezultat 77, koji ekstremno odstupa od glavnine rezultata. Medijana je opet aritmetička sredina četvrtog i petog rezultata, tj. medijana je opet jednaka 5. Prema tome, medijana, za razliku od aritmetičke sredine nije mnogo osetljiva na mali broj iznimaka u skupu rezultata. Stoga za medijanu kažemo da spada u tzv. rezistentne mere lokacije. Zbog toga je za izrazito asimetrične raspodele (raspodela sa malim brojem rezultata na jednom kraju vrednosti na varijabli, a velikim brojem rezultata na drugom kraju) medijana bolja mera centralne tendencije nego aritmetička sredina. Međutim, ni aritmetička sredina, ni medijana neće biti dobra mera "centralne tendencije" ukoliko je distribucija učestalosti takva da se veliki broj rezultata nagomilava oko dveju ili više međusobno relativno udaljenih vrednosti na varijabli. Takve raspodele se zovu bimodalne, odnosno polimodalne. Na primer, za sledeći skup rezultata možemo da kažemo da ima tzv. bimodalnu raspodelu, i to raspodelu koja je poznata pod imenom „raspodela U tipa" (budući da idealizovani grafički prikaz takve raspodele liči na latinično slovo U):

1, 1, 1, 1, 1, 2, 3, 4, 5, 5, 5, 5, 5

Takve raspodele često se javljaju kada posmatramo odgovore ispitanika na pojedine stavke pri ispitivanja nekih stavova. Naime, ispitanici, zavisno od povoljnosti svog stava biraju ili odgovor 0, ne slažu se uopšte sa tvrdnjom izrečenom u stavci ili odgovor 5, tj. sasvim se slažu sa tvrdnjom sadržanom u stavci.
Ukoliko na osnovu podataka u ovom primeru izračunamo aritmetičku sredinu i medijanu dobićemo sledeće vrednosti:
$M = 39/13 = 3$; $m = (13 + 1)/2 = 7$, $Mdn = X_{(7)} = 3$.
Dakle, i aritmetička sredina i medijana sugerišu da rezultati teže da se grupišu oko vrednosti 3 ili da je relativno najčešći odgovor 3 ili da „tipičan“ ispitanik bira odgovor 3, što je očigledno potpuno nesaglasno sa opštom strukturom, tj. distribucijom rezultata. Na ovaj primer bi se baš dobro mogla primeniti ona doskočica o aritmetičkoj sredini i medijani koje upućuju na to da svi jedu sarmu dok u realnosti neki jedu uglavnom kupus a drugi uglavnom meso! Za ovakve i slične „bimodalne“ raspodele, ili raspodele u kojima postoji čak i više od dve relativno udaljene lokacije oko kojih se nagomilavaju rezultati (tzv. „polimodalne“ raspodele najbolje je kao meru lokacije upotrebiti mod ili modalnu vrednost.

Za razliku od aritmetičke sredine, koja predstavlja meru saglasnu sa principom najmanjih kvadrata, medijana predstavlja meru koja je saglasna sa principom najmanjih apsolutnih odstupanja. Naime, već smo istakli da je za neki skup rezultata aritmetička sredina ona vrednost u odnosu na koju je zbir kvadriranih odstupanja rezultata manji od zbira kvadriranih odstupanja rezultata u odnosu na bilo koju drugu vrednost. S druge strane, medijana za neki skup rezultata je vrednost u odnosu na koju je zbir apsolutnih odstupanja rezultata manji nego zbir apsolutnih odstupanja od bilo koje druge vrednosti. Preciznije to možemo napisati na sledeći način:

$$\sum_{i=1}^{n} |x_i - \text{Mdn}| = \text{minimum} \quad < \sum_{i=1}^{n} |x_i - x_0| \,|\, x_0 \neq \text{Mdn}$$

(Vertikalna crta pre izraza $x_0 \neq \text{Mdn}$ označava uslov i čita se “pod uslovom da...” ili “ako ...”).[23]

Pored “lokacione i skalne ekvivarijantnosti” koju, kao mera lokacije, deli sa aritmetičkom sredinom, medijana ima opštije matematičko svojstvo koje bismo mogli nazvati “ekvivarijantnošću na monotone transformacije”: medijana posle monotone transformacije sirovih rezultata (bilo koje transformacije koje čuvaju redosled rezultata po veličini) jednostavno se može rekonstruisati na osnovu medijane sirovih rezultata i prirode monotone transformacije. Na primer, ako sve sirove rezultate logaritmujemo medijana logaritmovanih rezultata jednaka je logaritmu medijane sirovih rezultata. Aritmetičku sredinu, međutim, ne možemo na ovaj način rekonstruisati posle nelinearnih monotonih transformacija.

Kada se koristi kao deskriptivna statistička mera, medijana se uobičajeno zaokružuje na jednu decimalu.

[23] Kada je broj rezultata paran, tada je zapravo zbir apsolutnih odstupanja minimalan ne samo za vrednost medijane već i za sve vrednosti koje se nalaze između redoslednih statistika $X_{(m)}$ i $X_{(m+1)}$ čija aritmetička sredina predstavlja medijanu u tom slučaju.

Zapamtite:

- ✓ Medijanu možemo koristiti kao meru centralne tendencije ako rezultati predstavljaju rangove i intervalne ili racio mere;
- ✓ Medijana spada u tzv. **postojane**, tj. **rezistentne statistike** jer nije osetljiva na mali broj iznimaka, tj. autlajera;
- ✓ Za raspodele koje su simetrične ali imaju "gušće" krajeve od normalne raspodele (veću učestalost rezultata na krajevima raspodele nego što je to slučaj kod normalne raspodele) i za asimetrične raspodele smisaonije je kao meru centralne tendencije koristiti medijanu nego aritmetičku sredinu;
- ✓ Medijanu nema smisla koristiti kao meru centralne tendencije za raspodele "U tipa", tj. raspodele u kojima su rezultati grupisani na krajevima, a ne u sredini.

Mod (engl. Mode)

Mod, modus ili modalna vrednost je mera koja se najčešće javlja u nizu mera, tj. rezultat sa najvećom frekvencijom. U istoj raspodeli može biti više modalnih vrednosti. Raspodele učestalosti koje imaju jedan mod zovu se unimodalnim, one koje imaju dva moda bimodalnim, a raspodele sa više modalnih vrednosti polimodalnim raspodelama. Mod se može jednostavno odrediti pregledom jedinične raspodele učestalosti: vrednost moda jednaka je meri ili rezultatu sa najvećom učestalošću. Na primer, za niz rezultata 3, 4, 5, 5, 5, 5, 6, 7 vrednost moda jednaka je 5. Ukoliko, pak, dva, odnosno više rezultata imaju najveće ali jednake učestalosti tada mod ima dve, odnosno više vrednosti. Na primer, za niz rezultata 1, 1, 1, 1, 1, 2, 3, 4, 5, 5, 5, 5, 5 koji ima tzv. U raspodelu, postoje dve vrednosti moda: 1 i 5.
Kada se koristi kao deskriptivna statistička mera za opis uzorka u poogledu jedne kvantitativne varijable, mod se uobičajeno iskazuje u istom numeričkom obliku koji ima najučestalija mera.

Odnosi između aritmetičke sredine, medijane i moda za pojedine vrste distribucija

Ukoliko je distribucija rezultata na nekoj varijabli normalna tada znamo da je aritmetička sredina vrednost koja deli raspodelu na dva jednaka dela. Na grafiku normalne raspodele koji smo prikazali u prethodnoj glavi ovo se vidi sasvim jasno: ukoliko na grafiku normalne raspodele povučemo ordinatu iznad aritmetičke sredine ta ordinata deli celokupnu površinu ispod normalne krive na dva simetrična dela. Prema tome, aritmetička sredina normalne raspodele predstavlja i vrednost kvantila 0.5. Budući da je medijana isto što i kvantil 0.5 očigledno je da su aritmetička sredina i medijana normalne raspodele jednake. Isto tako, kod normalne raspodele visina ordinate, koja govori o gustini, ili, uslovno rečeno, o učestalosti rezultata, najveća je onda kada se ordinata podigne iznad aritmetičke sredine. To znači da je i mod normalne raspodele jednak aritmetičkoj sredini. Jednakost aritmetičke sredine i medijane ne postoji samo kod normalne raspodele već i kod svih simetričnih raspodela. Međutim, jednakost moda sa aritmetičkom sredinom i medijanom ne može se proširiti na sve simetrične raspodele, već samo na unimodalne

simetrične raspodele, tj. simetrične raspodele koje imaju samo jedan mod. Na primer, jedinična raspodela niza rezultata 1, 1, 1, 1, 1, 2, 3, 4, 5, 5, 5, 5, 5 izgledala bi ovako:

x_k	f_k
5	5
4	1
3	1
2	1
1	5
i = 1	n = 13

Očigledno je da je ova raspodela simetrična oko vrednosti 3, jer su učestalosti vrednosti manjih od 3 i vrednosti većih od 3 simetrične. Kao što smo to već prikazali, aritmetička sredina i medijana ove raspodele su međusobno jednake i imaju vrednost 3. Međutim, već na osnovu prvog pogleda na distribuciju učestalosti jasno je da 3 nije najučestaliji rezultat. Prema tome, iako je distribucija simetrična, budući da je ovde reč o bimodalnoj raspodeli, mod nije jednak aritmetičkoj sredini, te stoga ni medijani.
Postojanje jednakosti moda, aritmetičke sredine i medijane kod unimdoalnih simetričnih raspodela ne znači da iz jednakosti moda, medijane i aritmetičke sredine automatski možemo da zaključimo da je distribucija iz koje su izračunate ove statističke mere simetrična. Na primer, ako za niz podataka 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 3 napravimo jediničnu raspodelu učestalosti, ona će izgledati ovako:

x_k	f_k
3	1
2	2
1	6
0	4
i = 1	n = 13

Aritmetička sredina, medijana i mod ove raspodele jednaki su 1. Dakle, aritmetička sredina, medijana i mod mogu biti jednaki iako distribucija rezultata nije simetrična. Kada su, pak, u pitanju teorijske kontinuirane raspodele, najčešće je slučaj da jednakost aritmetičke sredine, medijane i moda upućuje na simetričnost raspodele, iako i ovde ima izuzetaka od ovog pravila (cf. von Hippel, 2005). Svakako, pri analizi podataka jednakost aritmetičke sredine, medijane i moda može se uzeti kao signal koji nagoveštava mogućnost da je raspodela rezultata simetrična, ali se na osnovu takve jednakosti ne sme zaključiti da je raspodela nužno simetrična.

Zapamtite:

- ✓ Ako je distribucija rezultata na nekoj varijabli **simetrična** onda su **aritmetička sredina** i **medijana** tih rezultata **jednake**. Međutim, medijana i aritmetička sredina mogu biti jednake i ako distribucija varijable nije simetrična.
- ✓ Ako je distribucija rezultata na nekoj varijabli **unimodalna i simetrična** onda su **aritmetička sredina, medijana** i **mod** tih rezultata **jednaki**.
  Budući da je normalna raspodela unimodalna i simetrična aritmetička sredina, medijana i mod normalne raspodele su jednaki.
  Međutim, medijana, aritmetička sredina i mod mogu biti jednaki i ako distribucija varijable nije simetrična.

Pored prikazanih mera lokacije, definisan je, uglavnom u poslednjih pet decenija, određeni broj ovih mera koje spadaju u tzv. robustne mere lokacije. U takve mere spadaju postrižena aritmetička sredina (engl. Trimmed mean), vinzorizovana aritmetička sredina (engl. Winsorized mean) i M-ocenitelj lokacije (engl. M-estimator of location). O ovim merama i o situacijama kada je njihovo korišćenje opravdano biće reči u glavi**.

Izbor mere centralne tendencije

Od svih mera centralne tendencije u psihologiji se, kao i u mnogim drugim oblastima, najčešće koristi aritmetička sredina. Međutim, kao što smo već istakli, **aritmetička sredina je samo u određenim uslovima najbolja mera centralne tendencije**. Ukoliko aritmetičku sredinu koristimo kao deskriptivnu statističku meru koja treba da nam pruži jednostavnu informaciju o prosečnoj, "tipičnoj" ili "najčešćoj" meri jedinica posmatranja u uzorku u pogledu određene varijable onda ćemo na osnovu aritmetičke sredine dobiti valjanu informaciju o centralnoj tendenciji samo ukoliko je distribucija učestalosti na toj varijabli određenog oblika, tj. ukoliko je distribucija učestalosti normalna. Ukoliko je, pak, distribucija rezultata izrazito asimetrična, aritmetička sredina nam može dati potpuno iskrivljenu sliku centralne tendencije rezultata, tj. potpuno pogrešnu predstavu o tome koji je rezultat "tipičan" ili "prosečan". Pogledajmo kako to izgleda na jednom jednostavnom primeru:

Rezultati uzorka I: 2; 3; 3; 3; 4; 4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 6.
Rezultati uzorka II: 2; 3; 3; 3; 4; 4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 36.

Očigledno je da "tipičan" ispitanik u oba uzorka ima meru 4 ili meru veoma blizu toj meri. Aritmetička sredina rezultata uzorka I jednaka je 4, a uzorka II jednaka je 5.88! Ova dva uzorka razlikuju se samo u jednom rezultatu: najveći rezultat u uzorku 1 je 6, a u uzorku 2 je 36. Da li su ova dva uzorka baš sasvim različita u pogledu "prosečnog" ili "tipičnog" člana uzorka kako bismo to mogli zaključiti iz njihovih aritmetičkih sredina? Da li tipičan ispitanik u drugom uzorku ima rezultat veoma blizu 6? Moglo bi se reći da aritmetička sredina o drugom uzorku ne daje valjanu informaciju o *centralnoj tendenciji* i "tipičnosti" već upravo informaciju koju od nje ne tražimo, tj. informaciju o *netipičnosti*! Distribucija rezultata u uzorku 2 je nesimetrična zbog prisustva jednog rezultata koji izrazito odstupa od svih ostalih rezultata, tzv. **iznimka** ili **autlajera**.

S druge strane, medijana i mod za oba uzorka jednaki su 4. Prema tome, u ovom slučaju medijana ili mod bi pružili valjaniju informaciju od aritmetičke sredine o centralnoj tendenciji rezultata u drugom uzorku.

Praktično posmatrano, pri analizi podataka najbolje je pri ekplorisanju podataka izračunati sve mere centralne tendencije koje nam stoje na raspolaganju. Pažljiva analiza nesaglasnih informacija koje slede iz različitih mera centralne tendencije može biti veoma korisna u razumevanju distribucije rezultata. Za sažet statistički opis centralne tendencije rezultata treba odabrati onu meru koja je najbliža "tački nagomilavanja" rezultata. Ukoliko, pak, nijedna od mera centralne tendencije uzeta sama po sebi ne daje valjanu informaciju o distribuciji rezultata onda je bolje navesti nekoliko mera lokacije koje će adekvatnije opisati datu raspodelu. Često se za takve situacije preporučuje tzv. petobrojni sažetak (engl. Five-Number Summary, cf. Moore & McCabe, 1998) koji se sastoji u navođenju najmanjeg rezultata, prvog kvartila, medijane, trećeg kvartila i najvećeg rezultata, tim redom. Na primer, pri davanju informacija o visini plata u nekoj zemlji to je svakako adekvatniji postupak nego navođenje "prosečne plate", tj. aritmetičke sredine plata.[24]

U ovoj glavi smo o svojstvima pojedinih mera centralne tendencije govorili prevashodno sa stanovišta njihovog korišćenja u deskriptivne svrhe – za sažet statistički opis distribucije rezultata. O svojstvima ovih mera, kada se one koriste u inferencijalne svrhe – za zaključivanje o karakteristikama populacije na osnovu rezultata dobijenih na uzorku biće reči u glavi **.

Zapamtite:

- ✓ Ako je distribucija rezultata na nekoj varijabli izrazito **asimetrična** ali **unimodalna** bolje je **umesto aritmetičke sredine** kao deskriptivnu meru centralne tendencije koristiti **medijanu.** Za raspodele "U tipa", kao i za polimodalne raspodele najbolje je od ovde prikazanih mera centralne tendencije koristiti modalne vrednosti raspodele.
- ✓ Ukoliko nijedna pojedinačna mera centralne tendencije ne daje valjanu informaciju o "centralnoj tendenciji" distribucije rezultata najbolje je za statistički opis distribucije koristiti "petobrojni sažetak" koji čine najmanji rezultat, prvi kvartil, medijana, treći kvartil i najveći rezultat.

5. Mere skale

Mere lokacije ukazuju samo na određene ograničene aspekte distribucije rezultata na kvantitativnoj varijabli. Na primer, percentil 20 (P_{20}) ukazuje na mesto u raspodeli ispod kojeg je 20% rezultata. Aritmetička sredina, medijana ili mod ukazuju na vrednost koja je "prosečna", "tipična", središnja ili najčešća u raspodeli rezultata. Međutim, dve suštinski ili, čak drastično, različite distribucije rezultata mogu imati iste mere lokacije. Na primer, sledeći skupovi podataka imaju identične aritmetičke sredine i medijane:

Skup I: 5 5 5 5 5 5 5 5
Skup II: 3, 4, 5, 5, 5, 5, 6, 7
Skup III: 3, 3, 3, 5, 5,7,7,7

Uz to skup I i skup II imaju i jednake modalne vrednosti. Pretpostavićemo da su sva tri skupa podataka dobijena merenjem iste varijable na kojoj su moguće vrednosti u intervalu od 1 do 7. Ipak, i površan pogled na ova tri skupa podataka dovoljan je da se uoči njihova

[24] Upravo u vreme kada je pisan ovaj tekst, u jednoj od najkvalitetnijih dnevnih novina izašao je još jedan napis o „prosečnoj plati" u Srbiji u kojem je navedena samo aritmetička sredina. Ko zna da li će ikada ta besmislena praksa biti korigovana. Ova praksa, naravno, jedino nije besmislena sa stanovišta vlasti, te se možda u tome i krije tajna njenog održavanja.

suštinska različitost. I dok se priroda različitosti skupa I i skupa II u odnosu na skup III može naslutiti iz razlika među samim merama centralne tendencije (u skupu III postoje dva moda koja su bitno različita od aritmetičke sredine i medijane) o različitosti između skupa I i skupa II nema nikakvih naznaka u pomenutim merama centralne tendencije. Kako bi se o distribuciji rezultata na kvantitativnoj varijabli mogla dobiti potpunija informacija mere centralne tendencije potrebno je dopuniti merama skale. Najpoznatije i najčešće korišćene mere skale su mere varijabilnosti, disperzije ili raspršenja.[25]

Postoje određeni formalni uslovi koje statistička mera treba da ispuni da bi se mogla smatrati merom skale (modifikovano na osnovu Bickel & Lehmann, 1976):

1. Mera skale $\tau(X)$, pri čemu je X slučajna 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.

Mera skale može biti mera disperzije, varijabilnosti ili raspršenja ako ispunjava sledeći dodatni uslov:

3. $\tau(F) \leq \tau(G)$, pri čemu su F i G simetrične distribucije, kad god je G raspršenije od F. Ako je F distribucija varijable X, a G distribucija varijable Y, G je raspršenije od F onda kada je $|Y - l_Y|$ stohastički veće od $|X - l_X|$, pri čemu su l_X i l_Y središnje vrednosti, mere centralne tendencije, tj. centri simetrije distribucija, u odnosu na koje se posmatraju odstupanja vrednosti na varijabli. Dakle, ako je distribucija apsolutnih odstupanja vrednosti na varijabli Y od središnje vrednosti pomerena udesno (ka većim vrednostima) u odnosu na distribuciju apsolutnih odstupanja vrednosti na varijabli X od središnje vrednosti tada je distribucija G (distribucija varijable Y) stohastički veća od distribucije F (distribcije varijable Y). Pojednostavljeno rečeno, mera disperzije bi trebalo da bude veća za raspršenije raspodele (raspodele sa izraženijim razlikama među vrednostima varijable) nego za manje raspršene raspodele.

Mere varijabilnosti (disperzije ili raspršenja)

Ako za tri prethodno prikazana skupa podataka izračunamo jednu od najpoznatijih mera varijabilnosti – standardnu devijaciju (koju ćemo objasniti u nastavku teksta)

[25] Termini varijabilnost (engl. Variability) i disperzija, tj. raspršenje (engl. Dispersion) koriste se u ovoj knjizi sinonimno. U teorijskim statističkim radovima Bikela i Lemana termin *dispersion* ograničava se na označavanje varijabilnosti simetričnih raspodela, dok se za asimetrične raspodele koristi termin *spread* (rasprostiranje)(cf. Bickel & Lehmann, 1976). U ovom tekstu nećemo praviti takve terminološke distinkcije, a najčešće ćemo za različitost rezultata na varijabli koristiti termin varijabilnost.

dobićemo sledeće vrednosti: za skup I standardna devijacija iznosi 0, za skup II ona uzima vrednost 1.20, a za skup III vrednost standardne devijacije je 1.48. Očigledno, mera varijabilnosti nam ukazuje na to da je svaki od ovih skupova unekoliko specifičan. To je informacija koju nismo mogli dobiti samo na osnovu mera centralne tendencije. Onda kada smo pored mera centralne tendencije izračunali i neku meru varijabilnosti lakše uočavamo razlike u distribucijama rezultata u ova tri skupa. Dakle, za statistički opis uzorka na osnovu kvantitativnih podataka dobijenih merenjem određene varijable nije dovoljno navesti samo meru lokacije, odnosno meru centralne tendencije već je potrebno navesti i odgovarajuću meru varijabilnosti.

Mere varijabilnosti imaju dvostruku ulogu jer sadrže u sebi informaciju o stepenu različitosti rezultata ali i informaciju o reprezentativnosti pojedinih mera centralne tendencije. S jedne strane, mere varijabilnosti izražavaju stepen variranja ili raspršenja u nizovima kvantitativnih podataka: pod određenim uslovima, što je vrednost određene mere varijabilnosti veća to je i varijabilnost skupa podataka iz kojeg je ta mera izračunata veća. Izraz "pod određenim uslovima" znači da je zaključivanje o varijabilnosti skupa podataka na osnovu mere varijabilnosti zavisno od konteksta podataka: merne skale koja je korišćena u merenju varijable, tj. mernih jedinica, veličine mere centralne tendencije i drugih kontekstualnih uslova. U prethodno prikazanom primeru, odabrana mera varijabilnosti pokazuje da među podacima u skupu I nema varijabilnosti, a da je varijabilnost podataka u skupu III veća nego u skupu II. (Zaključivanje o varijabilnosti skupa podataka neposredno na osnovu veličine standardnih devijacija u ovom primeru omogućeno je eksplicitnom pretpostavkom da je ista varijabla merena u sva tri slučaja korišćenjem iste skale, tj. istog instrumenta, a aritmetičke sredine su jednake u sva tri slučaja). S druge strane, mere varijabilnosti mogu pružiti informaciju o tome koliko je neka od mera centralne tendencije koju koristimo dobar predstavnik svih mera u nizu. Kada su sve mere u nizu jednake, kao što je to bio slučaj sa podacima iz skupa I u prethodnom primeru, tada su mere centralne tendencije dobar predstavnik svih ovih mera. Kako to najčešće nije slučaj, potrebno je znati koliko neka mera centralne tendencije dobro predstavlja niz iz kojega je izračunata.To se postiže tako što se uz svaku meru centralne tendencije po pravilu navodi i odgovarajuća mera varijabilnosti.

Najčešće korišćene mere varijabilnosti mogu se svrstati u dve široke kategorije (prema Žižić i sar., 2000): jednoj kategoriji pripadaju apsolutne, a drugoj relativne mere varijabilnosti. Apsolutne mere varijabilnosti iskazane su u mernim jedinicama, tj. u jedinicama u kojima su iskazani podaci na osnovu kojih se računaju, dok relativne mere varijabilnosti predstavljaju relativne odnose i mogu biti iskazane proporcijama (procentima) ili brojevima koji prikazuju odnos dveju ili više statističkih mera (najčešće mera skale i mera lokacije)

Apsolutne mere varijabilnosti

Raspon (engl. Range)

Raspon ili interval varijacije, u oznaci R, predstavlja najgrublju meru varijabilnosti i definiše se razlikom najvećeg i najmanjeg rezultata:

$$R = x_{max} - x_{min}$$

Raspon se može definisati i preko redoslednih statistika, kao razlika poslednjeg i prvog podatka u nizu podataka sortiranih po veličini od najmanjeg do najvećeg:

$$R = X_{(n)} - X_{(1)}$$

Dakle, ova mera zavisi samo od dva krajnja rezultata u raspodeli. Stoga je raspon veoma osetljiv na prisutvo iznimaka. Ipak, u nekim oblastima, posebno u ekonomiji, daje veoma korisne informacije, kao što su one o rasponu cena na tržištu ili rasponu ličnih dohodaka. Kada se koristi kao deskriptivna statistička mera, raspon se uobičajeno zaokružuje na onoliko decimala koliko decimala imaju najveći i najmanji rezultat.

Prosečno odstupanje (Engl. Mean absolute deviation)

Prosečno odstupanje, u oznaci PO, predstavlja aritmetičku sredinu apsolutnih odstupanja niza mera od njihove aritmetičke sredine:

$$PO = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

Oznaka x_i je oznaka rezultata za jedinicu posmatranja *i*, M je aritmetička sredina podataka, n je broj podataka, a $|\cdot|$ je oznaka za apsolutnu vrednost. Dakle, prosečno odstupanje je aritmetučka sredina svih odstupanja rezultata od njihove aritmetičke sredine, pri čemu se predznak negativnih odstupanja ne uzima u obzir, tj. sva odstupanja se tretiraju kao pozitivna.

Prosečno odstupanje se u psihologiji i srodnim oblastima veoma retko koristi te stoga nećemo ovu meru razmatrati podrobnije.

Standardna devijacija (engl. Standard deviation)

Standardna devijacija je jedna od najčešće korišćenih mera varijabilnosti u psihologiji i srodnim oblastima. To je prevashodno posledica veoma raširenog verovanja da se veliki broj varijabli koje se sreću u ovim oblastima normalno raspodeljuju u populaciji. U glavi o osnovnim pojmovima teorije verovatnoće definisali smo standardnu devijaciju populacije. Standardna devijacija uzorka, u oznaci S, definiše se na sledeći način:[26]

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n-1}}$$

Pri tome, x_i je rezultat za jedinicu posmatranja *i*, M je aritmetička sredina podataka, a n je broj podataka ili (ako posedujemo rezultate za sve jedinice posmatranja) veličina uzorka. Pogledom na obrazac uočavamo da standardna devijacija u osnovi predstavlja kvadratni koren iz određene vrste proseka kvadriranih odstupanja rezultata od njihove aritmetičke sredine. Veoma je važno uočiti u brojiocu potkorenog izraza u obrascu za standardnu devijaciju već poznatu strukturu: zbir kvadriranih odstupanja rezultata od njihove aritmetičke sredine. Ovu strukturu smo već pominjali pri objašnjavanju principa najmanjih kvadrata kao matematičkog principa za definisanje aritmetičke sredine. Iz određenih

[26] Standardna devijacija uzorka često se označava i oznakom SD. Mi ćemo u ovoj knjizi koristiti isključivo oznaku S.

statističkih razloga, pri računanju standardne devijacije uzorka najčešće se zbir kvadriranih odstupanja rezultata od aritmetičke sredine deli sa n – 1, a ne sa n. U izuzetnim slučajevima, pri računanju standardne devijacije uzorka, moguće je koristiti obrazac sa n umesto sa n – 1 u imeniocu potkorenog izraza. To je opravdano samo onda kada standardnu devijaciju uzorka koristimo isključivo kao meru varijabilnosti datog uzorka. Međutim, najčešće standardnu devijaciju uzorka koristimo i da ocenimo varijabilnost populacije u pogledu ispitivanog obeležja. U tom slučaju, iz razloga koje ćemo objasniti u glavi **, standardnu devijaciju uzorka računamo primenom obrasca sa n – 1 u imeniocu potkorenog izraza. Kada je broj rezultata veliki, tj. kada je uzorak veći od 150 praktično isti rezultati se dobijaju bez obzira na to da li se standardna devijacija računa deljenjem zbira kvadriranih odstupanja sa n ili sa n – 1.

Računanje standardne devijacije pokazaćemo na primeru niza podataka koje smo koristili za demonstriranje računanja aritmetičke sredine: 3, 4, 5, 5, 5, 5, 6, 7.

- Prvi korak u računanju standardne devijacije jeste računanje aritmetičke sredine: aritmetička sredina ovog niza podataka iznosi 5. Dakle, M = 5;
- Potom se računa zbir kvadriranih odstupanja rezultata od aritmetičke sredine: $(3 – 5)^2 + (4 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (6 – 5)^2 + (7 – 5)^2 = (-2)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 2^2 = 4 + 1 + 0 + 0 + 0 + 0 + 1 + 4 = 10$;
- Zbir kvadriranih odstupanja deli se brojem rezultata umanjenim za 1: 10 / (8 - 1) = 1.4285714;
- Iz broja koji se dobija deljenjem u prethodnom koraku računa se pozitivni kvadratni koren: $+\sqrt{1.4285714} = 1.20$.

Dakle, S = 1.20 (Čitaocu ove knjige savetujemo da samostalno ponovi ovaj postupak, bez obzira na to što će standardnu devijaciju u principu računati primenom nekog od statističkih programa, jer verujemo da se samostalnim obavljanjem procedure računanja standardne devijacije primenom definicionog obrasca stiče bolje razumevanje prirode ove mere).

Standardna devijacija ima određena matematička svojstva koja je korisno znati kako bi ova mera varijabilnosti bila adekvatno korišćena i tumačena:

- Ako je $S(v_1)$ standardna devijacija varijable v_1, i ako se rezultati na varijabli v_1 linearno transformišu tako da se dobije varijabla $v_2 = a + bv_1$ (na svaki rezultat na varijabli v_1 se doda tzv. aditivna konstanta a i svaki rezultat se pomnoži tzv. multiplikativnom konstantom b), tada je $S(v_2) = |b| S(v_1)$. Dakle, dodavanje konstante na svaki rezultat ne menja standardnu devijaciju, a množenje svakog rezultata konstantom menja standardnu devijaciju za apsolutnu vrednost konstante puta. Ova svojstva standardna devijacija deli sa svim merama skale, a nazivaju se lokaciona invarijantnost i skalna ekvivarijantnost, tim redom.
- Standardna devijacija je iskazana u mernim jedinicama, tj. u jedinicama u kojima su iskazani podaci iz kojih se računa.
- Standardna devijacija pokazuje koliko je, po principu najmanjih kvadrata, aritmetička sredina "prosečno" udaljena od svakog podatka u skupu podataka iz kojih je izračunata. Prema tome, standardna devijacija ukazuje na to koliko dobro aritmetička sredina predstavlja sve mere iz kojih je izračunata. Ako fiksiramo, tj. držimo konstantnim broj mera, jasno je da što je aritmetička sredina bliža merama koje predstavlja to će suma kvadrata biti manja, te će stoga i standardna devijacija biti manja. Manja suma kvadrata i, posledično, manja standardna devijacija govore o boljoj

reprezentovanosti niza mera njihovom aritmetičkom sredinom. Naravno, treba imati u vidu da ovde nije reč o bilo kakvoj reprezentovanosti, već o reprezentovanosti niza mera jednom vrednošću po principu najmanjih kvadrata.

- Standardna devijacija je i mera koja pokazuje koliko je, prosečno gledano, svaki podatak udaljen od svakog drugog. Prema tome, standardna devijacija predstavlja i pokazatelj međusobne nesličnosti rezultata.[27] Naime, standardna devijacija se može definisati i na još jedan, u udžbenicima neuobičajen ali intuitivno veoma smislen način: posmatranjem međusobnih distanci svih rezultata, tj. bez uzimanja u razmatranje odstupanja rezultata od njihove aritmetičke sredine (Gordon, 1986). Na ovaj način, pri definisanju standardne devijacije uzima se u obzir koliko se svaki rezultat razlikuje od svakog drugog. Standardna devijacija posmatrana na ovaj način može se definisati sledećim izrazom:

$$S = \sqrt{\frac{\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)^2}{2n(n-1)}}$$

pri čemu su x_i i x_j bilo koja dva rezultata.[28] Uočimo da brojilac potkorene veličine predstavlja zapravo zbir svih elemenata matrice kvadriranih distanci svakog rezultata od svakog drugog. Imenilac potkorene veličine jednak je dvostrukom broju poređenja svakog rezultata sa svakim drugim umanjenim za broj poređenja svakog rezultata sa samim sobom.[29] Broj poređenja je umanjen za broj poređenja rezultata sa samim sobom jer je distanca rezultata od samog sebe jednaka nuli pa ne doprinosi ništa brojiocu. Budući da ista kvadrirana distanca između bilo koja dva rezultata doprinosi brojiocu dva puta (jednom kao razlika $x_i - x_j$ a drugi put kao distanca $x_j - x_i$) onda je i broj poređenja u imeniocu pomnožen sa 2. Standardna devijacija izračunata na ovaj način jednaka je standardnoj devijaciji po obrascu **, dakle sa n - 1 u imeniocu potkorenog izraza.
Za skup rezultata koje smo već koristili za demonstriranje računanja standardne devijacije

3, 4, 5, 5, 5, 5, 6, 7

matrica kvadriranih distanci izgledala bi ovako:

[27] Na sličan način u sledećem poglavlju definisaćemo meru „varijabilnosti“, raznolikosti ili raznovrsnosti kategoričkih podataka. U slučaju kategoričkih podataka nećemo uzimati u obzir veličinu razlika između rezultata – jer to nije ni moguće – već samo učestalost tih razlika.

[28] Ovaj obrazac sledi na osnovu sledeće teoreme: Ako su $x_1, x_2, ..., x_n$ podaci, tada je $\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)^2 = 2n\sum_{i=1}^{n}(x_i - M)^2$. Dokaz ove teoreme može se naći u Jones & Scariano, 2014.

[29] Matematički broj poređenja svakog rezultata sa svakim drugim rezultatom predstavlja broj varijacija bez ponavljanja, tj. broj permutacija 2 elementa od ukupno n elemenata (pogledati tačku 4 u odrednici **Osnovni pojmovi i pravila kombinatorike** u Matematičkom pojmovniku u Dodatku **.

	3	4	5	5	5	5	6	7
3	0	1	4	4	4	4	9	16
4	1	0	0	1	1	1	4	9
5	4	1	0	0	0	0	1	4
5	4	1	0	0	0	0	1	4
5	4	1	0	0	0	0	1	4
5	4	1	0	0	0	0	1	4
6	9	4	1	1	1	1	0	1
7	16	9	4	4	4	4	1	0

Zbir svih elemenata ove matrice distanci koji je u obrascu predstavljen izrazom u brojiocu potkorene veličine jednak je 160. Prema tome standardna devijacija ovog skupa rezultata jednaka je:

$$S = \sqrt{\frac{160}{2*8*7}} = 1.1952 \approx 1.20$$

Uočimo da je zbir distanci iznad glavne dijagonale (u kojoj su distance rezultata sa samima sobom, te su nužno jednake nuli) jednak 80 i, istovremeno, jednak zbiru distanci ispod glavne dijagonale. Dakle, distanca svakog para rezultata se pojavljuje dva puta (jednom iznad a drugi put ispod glavne dijagonale) te je stoga potrebno zbir svih distanci u matrici podeliti dvostrukim brojem parova rezultata koji se porede. Pri tome, u ukupan broj parova koji se porede ne ulaze parovi rezultata sa samima sobom te je broj elemenata u matrici umanjen za broj rezultata. Dakle, broj parova u imeniocu potkorenog izraza je 56, što je 64 (koliko je elemenata u matrici distanci) umanjeno za 8 (broj rezultata), a to se upravo i dobija kao proizvod 8*7.

Uočimo, dakle, da standardna devijacija uzorka ne pokazuje samo koliko su jedinice posmatranja u pogledu nekog obeležja „prosečno“ udaljene od aritmetičke sredine već istovremeno i koliko se, „prosečno gledano“, jedinice posmatranja razlikuju međusobno, tj. svaka od svake druge. Na taj način, standardna devijacija je mera varijabilnosti i nesličnosti, tj. raznolikosti rezultata. Pojam raznolikosti (u smislu međusobne nesličnosti jedinica posmatranja) veoma je koristan jer se može generalizovati i na kategoričke podatke, tj. podatke za koje nema smisla računati aritmetičku sredinu, te i sama definicija varijabilnosti koja uključuje odstupanja rezultata od aritmetičke sredine nema smisla. (O raznolikosti kategoričkih podataka biće reči u narednoj glavi knjige).

- Budući da se radi o meri u čijem definisanju ključnu ulogu igra zbir kvadriranih odstupanja pojedinačnih rezultata od njihove aritmetičke sredine, tj. princip najmanjih kvadrata, standardna devijacija je veoma osetljiva na postojanje iznimaka u podacima. Pogledajmo kako će se promeniti standardna devijacija, ako zamislimo da smo pri unosu podataka, kao što smo to uradili pri prikazu aritmetičke sredine i medijane, pogrešno uneli samo jedan od podataka, te umesto skupa rezultata:

3, 4, 5, 5, 5, 5, 6, 7

imamo sledeći skup:

3, 4, 5, 5, 5, 5, 6, 77.

Dakle, u potonjem skupu podataka postoji jedan iznimak, tj. rezultat 77, koji ekstremno odstupa od glavnine rezultata. Standardna devijacija prvog niza, kako smo to već prikazali, jednaka je 1.20 i sugeriše nam informaciju o tome da je glavnina kvadriranih odstupanja rezultata od njihove aritmetičke sredine relativno mala (ako pogledamo pojedinačna kvadrirana odstupanja u primeru u kojem smo to izračunali videćemo da je većina kvadriranih odstupanja jednaka 0 ili 1). Rekli bismo da je to relativno ispravna informacija. Ukoliko, pak, izračunamo standardnu devijaciju drugog skupa podataka (skupa sa iznimkom) dobijamo sledeći zbir kvadriranih odstupanja:
$(3 – 5)^2 + (4 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (5 – 5)^2 + (6 – 5)^2 + (77 – 5)^2 = (-2)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 72^2 = 4 + 1 + 0 + 0 + 0 + 0 + 1 + 5184 = 5190$.
Standardna devijacija u ovom slučaju jednaka je 25.57! Da li standardna devijacija u ovom slučaju ispravno sugeriše kolika je veličina glavnine odstupanja? Očigledno ne, jer informacija koju ona sugeriše jeste da je glavnina tih odstupanja veoma velika, što naprosto nije tačno: glavnina ovih odstupanja je i u ovom skupu podataka mala, a samo jedno odstupanje je izuzetno veliko. Istina, može se reći da standardna devijacija matematički ispravno prikazuje informaciju o svim odstupanjima. Međutim, informacija koju onaj ko koristi standardnu devijaciju očekuje jeste najčešće informacija o veličini glavnine odstupanja. Mogli bismo malo slobodnijim i na izvestan način politički formulisanim rečima reći da se, u određenim uslovima, "demokratičnost" standardne devijacije (podjednako vođenje računa o "interesima" svih, pa i onih rezultata koji su ekstremni) pretvara u "potpadanje pod nesrazmeran uticaj" ovih ekstremnih rezultata.

Standardna devijacija se uobičajeno koristi kao mera varijabilnosti uz aritmetičku sredinu kao meru centralne tendencije. To nije slučajno. Kao što smo već istakli, obe ove mere zasnovane su na principu najmanjih kvadrata. Stoga je, poput aritmetičke sredine kao mere centralne tendencije, standardna devijacija najpodesnija mera varijabilnosti ukoliko skup rezultata ima normalnu raspodelu ili ukoliko je računamo iz podataka za uzorak iz populacije u kojoj se, po pretpostavci, varijabla kojom se bavimo normalno raspodeljuje. U svakom slučaju, za podatke koji sadrže iznimke standardna devijaciju ne treba koristiti.

Kada se koristi kao deskriptivna statistička mera, standardna devijacija se uobičajeno zaokružuje na dve decimale.

Varijansa (engl. Variance)

Budući da nije iskazana u mernim jedinicama u kojima su iskazani podaci iz kojih se računa, varijansa ne spada u apsolutne mere varijabilnosti. Ipak, objasnićemo varijansu kao meru varijabilosti ovde zbog važnosti ove statističke mere u statistici i zbog njenog bliskog matematičkog odnosa prema standardnoj devijaciji.

Varijansa uzorka, u oznaci S^2, predstavlja na izvestan način prosečno kvadrirano odstupanje skupa rezltata mera od njihove aritmetičke sredine:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - M)^2}{n - 1}$$

Kao što se iz definicionog obrasca za varijansu uzorka može videti, prosečno kvadrirano odstupanje računamo na uzorku, kao i pri računanju standardne devijacije, deljenjem zbira kvadriranih odstupanja sa brojem tih odstupanja umanjenim za 1.[30] (Kao što smo to već obećali pri objašnjavanju standardne devijacije, statističke razloge zbog kojih u imeniocu varijanse uzorka koristimo n – 1 izložićemo u glavi **). Upređivanjem obrazaca za standardnu devijaciju i varijansu uzorka možemo odmah uočiti da je, budući da je kvadrat korena nekog broja jednak tom broju, varijansa matematički jednaka kvadratu standardne devijacije. Varijansa, prema tome, nije iskazana u mernim jedinicama, već u kvadriranim mernim jedinicama, te ne spada u apsolutne mere varijabilnosti.

Obrazac za varijansu uzorka može se napisati i u obliku analognom onome u kojem smo pri izlaganju osnovnih pojmova teorije verovatnoće definisali varijansu diskrene slučajne varijable:

$$S^2 = \sum_{k} (x_k - M)^2 \frac{f_k}{n - 1}$$

Pri tome, sa x_k su označene sve <u>različite</u> vrednosti koje se pojavljuju u skupu podataka, a f_k su učestalosti tih vrednosti. Budući da su rezultati dobijeni na uzorku uvek diskretni, bez obzira na to da li je varijabla teorijski kontinuirana ili diskretna, obrazac za varijansu uzorka zapravo je analogan matematičkoj definiciji varijanse za diskretnu slučajnu varijablu. Ako na trenutak zanemarimo to što iz statističkih razloga u obrascu za varijansu uzorka u imeniocu količnika na desnoj strani stoji n – 1 umesto n, količnikom fk/n zapravo ocenjujemo verovatnoće različitih vrednosti varijable

Varijansa se vrlo retko koristi kao mera za statistički opis uzorka u pogledu neke kvantitativne varijable. Ova mera se, međutim, veoma često koristi u daljim statističkim analizama podataka dobijenim na uzorku. Stoga je korisno poznavati sledeća matematička svojstva varijanse:

- Ako je $S^2(v_1)$ varijansa varijable v_1, i ako se rezultati na varijabli v_1 linearno transformišu tako da se dobije varijabla $v_2 = a + bv_1$ (na svaki rezultat na varijabli v_1 se doda tzv. aditivna konstanta *a* i svaki rezultat se pomnoži tzv. multiplikativnom konstantom *b*), tada je $S^2(v_2) = b^2S^2(v_1)$. Dakle, varijansa je lokaciono invarijantna (ne menja se pri dodavanju konstante na svaki rezultat) i skalno ekvivarijantna (množenje svakog rezultata konstantom menja varijansu za kvadrat konstante puta).
- Varijansa je mera koja, slično kao i standardna devijacija, pokazuje koliko je aritmetička sredina "prosečno" udaljena od svakog podatka u skupu podataka iz kojih je izračunata. Međutim, ova "prosečna" udaljenost nije u slučaju varijanse iskazana mernim jedinicima samih podataka već kao kvadrirana udaljenost, te je za tumačenje udaljenosti aritmetičke sredine od podataka jednostavnije uzeti u razmatranje standardnu devijaciju.
- Varijansa je mera koja pokazuje i koliko su, prosečno gledano, velike kvadrirane udaljenosti (distance) svakog rezultata od svakog drugog. Prema

[30]Naravno, u retkim slučajevima, kada nas zanima isključivo varijabilnost uzorka za koji računamo varijansu, prosečno kvadrirano odstupanje, tj. varijansu možemo računati deljenjem zbira kvadriranih odstupanja sa brojem tih odstupanja.

tome, kao i standardna devijacija, varijansa predstavlja i pokazatelj međusobne nesličnosti rezultata. Samo je u slučaju varijanse ova nesličnost iskazana kvadriranim jedinicama a ne jedinicama kojima su iskazani sirovi rezultati.

- Budući da je, kao i standardna devijacija, zasnovana na principu najmanjih kvadrata, varijansa je izrazito osetljiva na postojanje iznimaka u podacima. Promena vrednosti varijanse u prisustvu autlajera još je drastičnija nego što je promena vrednosti standardne devijacije. Ako uporedimo vrednosti varijanse za dva skupa podataka iz primera kojim smo ilustrovali osetljivost aritmetičke sredine i standardne devijacije na autlajere, videćemo da je varijansa za skup podataka bez autlajera, tj. skup 3, 4, 5, 5, 5, 5, 6, 7 jednaka 1.43, a za skup podataka sa jednim autlajerom, tj. skup 3, 4, 5, 5, 5, 5, 6, 77 jednaka 653.93. Varijansa se u ovom slučaju povećala približno za 457 puta, dok se standardna devijacija povećala približno za 21.5 puta. Povećanje varijanse za onoliko puta koliko iznosi odgovarajuće kvadrirano povećanje za standardnu devijaciju je razumljivo jer je posledica činjenice da je varijansa matematički jednaka kvadratu standardne devijacije.
- Varijansa pod određenim uslovima ima svostvo aditivnosti: pod određenim uslovima (koje nećemo ovde obrazlagati) varijansa zbira dveju ili više varijabli jednaka je zbiru varijansi tih varijabli.
- Varijansa se može razlagati na sastavne komponente prema pretpostavljenim izvorima koji stoje iza pojedinih komponenti varijanse. Naime, sume kvadrata koje figuriraju u brojiocu varijanse, a koje iskazuju varijabilitet na varijabli, imaju svostvo aditivnosti tako da se ukupna suma kvadrata za neku varijablu može razložiti na aditivne komponente koje su posledica delovanja slučajnih i sistematskih faktora. Na ovom razlaganju ukupnih suma kvadrata počiva čitav niz statističkih postupaka koji se objedinjeno zovu *analiza varijanse*.

Zbog jednostavne matematičke veze koja postoji između standardne devijacije i varijanse, računanje kombinovane varijanse nećemo pokazivati detaljnije na primeru. Za primer na kojem smo prikazali računanje kobinovane standardne devijacije, kombinovanu varijansu bila bi jednaka kvadratu kombinovane standardne devijacije, tj. 28.06

Kada se koristi kao deskriptivna statistička mera, varijansa se uobičajeno zaokružuje na dve decimale.

Interkvartilni raspon (engl. Interquartile range)

Interkvartilni raspon ili interkvartilna razlika, u oznaci IQR, predstavlja razliku trećeg i prvog kvartila, ili, što je isto, razliku percentila 75 i percentila 25:

$IQR = Q_3 - Q_1 \equiv P_{75} - P_{25}$
(Oznaka ≡ predstavlja oznaku identiteta i čita se "identično jednako").

Kao što se iz definicije ove mere vidi, i ona, slično kao raspon, zavisi samo od dve vrednosti. Međutim, pošto vrednosti od kojih zavisi interkvartilni raspon nisu na krajevima distribucije, ova mera nije osetljiva na iznimke ukoliko broj iznimaka nije jako veliki. Na primer, ova mera bi mogla biti osetljiva na iznimke ukoliko je njihov broj na jednom kraju raspodele veći od četvrtine ukupnog broja podataka, što se veoma retko dešava. Stoga se interkvartilni raspon koristi u pojedinim postupcima za otkrivanje iznimaka, o čemu će biti

reči u ovoj glavi pri objašnjenju kutijastog dijagrama kao grafičkog postupka za eksplorisanje kvantitativnih podataka. Poređenjem interkvartilnog raspona sa rasponom moguće je doći do korisnih informacija o distribuciji podataka: ukoliko je raspon dramatično veći od interkvartilnog raspona to može ukazivati na postojanje iznimaka u raspodeli.
Za razumevanje određenih obrazaca u statistici korisno je znati da je interkvartilni raspon za teorijsku raspodelu koju zovemo standardizovanom normalnom raspodelom (koju smo definisali u glavi o osnovnim pojmovima teorije verovatnoće) jednak 1.349. Naime, percentil 75 za tu raspodelu jednak je 0.6745, a percentil 25 jednak je -0.6745. Prema tome: IQR = 0.6745 – (-0.6745) = 1.349. U opštem slučaju, kada je distribucija varijable normalna, interkvartilni raspon je približno jednak standardnoj devijaciji varijable pomnoženoj sa 1.349:

$$\mathrm{IQR} \approx 1.349\sigma$$

Interkvartilni raspon najsmisaonije je upotrebiti kao meru varijabilnosti ukoliko se kao mera centralne tendencije koristi medijana. Dakle, ovu meru možemo koristiti i za statistički opis unimodalnih asimetričnih raspodela rezultata.

Kada se koristi kao deskriptivna statistička mera, interkvartilni raspon se uobičajeno zaokružuje na jednu decimalu.

Kvartilna devijacija (engl. Quartile deviation ili Semi-interquartile range)

Kvartilna devijacija, u oznaci Q, predstavlja polovinu interkvartilnog raspona:

$$Q = \frac{Q_3 - Q_1}{2} \equiv \frac{P_{75} - P_{25}}{2}$$

Budući da je direktno izvedena iz interkvartilnog raspona deljenjem skalarom, tj. brojem 2, u pogledu korišćenja kvartilne devijacije važe iste preporuke koje smo dali za interkvartilni raspon.
Kako je interkvartilni raspon za standardizovanu normalnu raspodelu jednak 1.349, kvartilna devijacija za ovu raspodelu iznosi 0.6745. Budući da je standardna devijacija standardizovane normalne raspodele jednaka 1, kvartilna devijacija za standardizovanu normalnu raspodelu iznosi 67.5% standardne devijacije te raspodele.

Kada se koristi kao deskriptivna statistička mera, kvartilna devijacija se uobičajeno zaokružuje na jednu decimalu.

U situacijama kada u skupu podataka postoje autlajeri mogu se, umesto prikazanih mera varijabilnosti koristiti tzv. robustne mere skale, kao što su vinzorizovana standardna devijacija, medijansko apsolutno odstupanje i druge. Neke od robustnih mera skale prikazaćemo u glavi **.

Relativne mere varijabilnosti

Za razliku od apsolutnih mera varijabilnosti, relativne mere predstavljaju odnose i stoga se njihovih korišćenjem mogu međusobno porediti varijabilnosti skupova podataka koji nisu iskazani istim mernim jedinicama, tj. nisu dobijeni korišćenjem iste merne skale.

Koeficijent varijacije (engl. Coefficient of variation)

Koeficijent varijacije, u oznaci CV, predstavlja količnik standardne devijacije (S) i aritmetičke sredine (M):

$$CV = \frac{S}{M}$$

Ovaj koeficijent se često pojavljuje i u obliku u kojem je količnik standardne devijacije i aritmetičke sredine pomnožen sa 100.

Budući da je reč o relativnoj meri varijabilnosti, koeficijent varijacije se može koristiti za poređenje različitih uzoraka prema varijabilnosti u pogledu iste varijable, ili za poređenje varijabilnosti istog uzorka u pogledu različitih varijabli. Na primer, ako bismo želeli da vidimo da li su učenici nekog odeljenja varijabilniji, tj. da li se međusobno više razlikuju u pogledu uspeha na testu znanja biologije ili u pogledu uspeha na testu znanja istorije mogli bismo u te svrhe upotrebiti koeficijent varijacije. U situacijama kada podaci na kojima računamo koeficijent varijacije potiču sa racio nivoa merenja, ima smisla računati i koliko puta je određeni varijabilitet veći u odnosu na neki drugi, a ne samo porediti različite varijabilitete po veličini ("koji je veći a koji manji"). U osnovi koeficijent varijacije najviše ima smisla koristiti kada podaci poticu sa racio nivoa merenja.

Računanje koeficijenta varijacije pokazaćemo na istom primeru na kojem smo prikazali računanje zajedničke standardne devijacije. Statistički pokazatelji uspeha na testu znanja biologije za tri odeljenja istog razreda u određenoj školi prikazani su u sledećoj tabeli:

Odeljenje	VII1	VII2	VII3
Aritmetička sredina	$M_1 = 14.00$	$M_2 = 17.00$	$M_3 = 15.00$
Standardna devijacija	S1 = 5.00	S2 = 6.00	S3 = 4.00
Broj učenika koji su polagali test	$n_1 = 25$	$n_1 = 30$	$n_1 = 20$

Na osnovu prikazanih pokazatelja, dobili bismo sledeće vrednosti koeficijenata varijacije:

VII1 CV = 5 / 14 = 0.36
VII2 CV = 6 / 17 = 0.35
VII3 CV = 4 / 15 = 0.26

Dakle, u odeljenju VII3 uočljivo je manja varijabilnost u pogledu uspeha na testu znanja biologije nego u preostala dva odeljenja.

Budući da je zasnovan na računanju aritmetičke sredine i standardne devijacije koeficijent varijacije najviše ima smisla koristiti za rezultate koji imaju normalnu raspodelu ili za rezultate dobijene na uzorcima iz populacije u kojoj se varijabla po pretpostavci normalno distribuira.

Kada se koristi kao deskriptivna statistička mera, koeficijent varijacije se uobičajeno zaokružuje na dve decimale.

Koeficijent interkvartilne varijacije (engl. Quartile coefficient of dispersion, Coefficient of quartile variation)

Koeficijent interkvartilne varijacije uzorka, u oznaci V_Q, definisan je na sledeći način:

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

I ovaj koeficijent, kao i koeficijent varijacije, često se pojavljuje i u obliku u kojem je količnik razlike i zbira kvartila pomnožen sa 100. Koeficijent interkvartilne varijacije može uzeti vrednost u segmentu od 0 do 1 (ili od 0% do 100% ako je pomnožen sa 100), pri čemu veća vrednost ukazuje na izraženiju varijabilnost. Budući da je zasnovan na prvom i trećem kvartilu, ovaj koeficijent ima smisla koristiti i za distribucije koje nisu normalne, te i za asimetrične raspodele.

Kada se koristi kao deskriptivna statistička mera, koeficijent interkvartilne varijacije se uobičajeno zaokružuje na jednu decimalu.

6. Mere oblika distribucije

Pojam "oblika distribucije" odnosi se u osnovi na vizuelni oblik koji distribucija neke varijable ima kada se ta distribucija prikaže grafički. Grafički prikaz distribucije određene varijable može davati vizuelni oblik koji je simetričan ili asimetričan u odnosu na određenu liniju. Ta linija se uobičajeno kod statističkih distribucija kvantitativnih varijabli odnosi na vertikalnu liniju koja se diže iz tačke na apscici koja odgovara aritmetičkoj sredini. Na primer, ukoliko na grafičkom prikazu normalne funkcije gustine, tj. na prikazu koji zovemo Gausova kriva, dignemo ordinatu iz tačke na apscisi gde se nalazi aritmetička sredina (medijana i mod) možemo lako uočiti simetričnost Gausove krive u odnosu na ovu ordinatu: levi deo krive je naprosto "slika u ogledalu" desnog dela ove krive. Postoji, naravno, beskonačan broj, kako teorijskih, tako i empirijskih raspodela koje nemaju ovo svojstvo simetričnosti. To su tzv. asimetrične raspodele. Isto tako, grafički prikazi pojedinih raspodela su izduženiji ili spljošteniji u odnosu na Gausovu krivu, ili, pak imaju veće ili manje nagomilavanje rezultata na krajevima raspodele nego što je to slučaj kod normalne raspodele.

Dakle, dva ključna svojstva koja uzimamo u obzir kada razmatramo oblik neke distribucije učestalosti jesu simetričnost i izdignutost ("kurtotičnost") krajeva raspodele. Oba svojstva je jednostavnije razumeti ako se posmatraju u odnosu na normalnu raspodelu:

a) Distribucije koje su u pogledu simetrije poput normalne raspodele, tj. kod kojih "leva" strana njihovog grafičkog prikaza predstavlja "sliku u ogledalu" desne strane tog grafičkog prikaza u odnosu na ordinatu koja se diže iz "središnje vrednosti" su simetrične. Kada je reč o empirijskim raspodelama simetričnost znači da su učestalosti rezultata manjih i većih od aritmetičke sredine jednake ali su jednake i udaljenosti u odnosu na aritmetičku sredinu ovih dveju kategorija rezultata. Tačnije rečeno, za svaki rezultat koji je veći od aritmetičke sredine postoji odgovarajući rezultat manji od aritmetičke sredine koji je podjednako udaljen od nje.

b) Normalna raspodela se u razmatranju izdignutosti krajeva raspodele tretira kao "mezokurtična" raspodela, tj. raspodela sa određenom visinom najviše ordinate i sa određenom izdignutošću krajeva raspodele. Zašto je izdignutost krajeva raspodele važna? Zato što se u raspodelama sa izdignutijim krajevima (leptokurtičnim raspodelama) relativno veći broj rezultata nalazi na krajevima raspodele nego u raspodelama sa manje izdignutim krajevima (platikurtičnim raspodelama). Dakle, leptokurtične raspodele imaju veće nagomilavanje rezultata na krajevima nego što je to slučaj kod normalne raspodele, dok je u platikurtičnim raspodelama nagomilavanje rezultata na krajevima raspodele manje nego kod normalne raspodele.

U delu teksta pod naslovom "Odnosi između aritmetičke sredine, medijane i moda za pojedine vrste distribucija" istakli smo da se često na osnovu jednakosti aritmetičke sredine, medijane i moda može zaključiti da je reč o simetričnoj raspodeli. Neke informacije o obliku raspodele možemo u određenim situacijama dobiti iz statističkih mera koje smo do sada prikazali u ovoj glavi. Međutim, te informacije mogu biti nepouzdane: kako smo to već pokazali u delu teksta pod pomenutim naslovom, aritmetička sredina, medijana i mod mogu biti jednaki a da raspodela rezultata bude asimetrična. Kao što smo to u glavi posvećenoj osnovnim pojmovima teorije verovatnoće istakli, za dobijanje specifične informacije o obliku raspodele definisane su dve statističke mere: skjunis i kurtozis. Skjunis daje informaciju o simetričnosti a kurtozis o izdignutosti krajeva raspodele, tj. verovatnoći (ili relativnoj učestalosti) vrednosti na krajevima raspodele u odnosu na verovatnoću (ili relativnu učestalost) vrednosti u sredini raspodele. U glavi ** smo skjunis i kurtozis definisali kao parametre oblika, tj. kao mere koje govore o obliku distribucije verovatnoća, odnosno fnkcije gustine slučajne varijable. U ovoj glavi ćemo definisati skjunis i kurtozis kao statistike, tj. kao statističke mere koje ukazuju na oblik distribucije rezultata na varijabli dobijenih na uzorku.

Skjunis (engl. Skewness)[31]

Skjunis uzorka ili koeficijent asimetrije distribucije učestalosti dobijene na uzorku, u oznaci Sk, računa se primenom sledećeg obrasca:[32]

$$Sk=\frac{\sum_{i=1}^{n}(x_i-M)^3}{S^3}\frac{n}{(n-1)(n-2)}$$

Oznakom M u obrascu označena je aritmetička sredina uzorka, oznakom S standardna devijacija, a n označava veličinu uzorka, tj. broj podataka na osnovu kojih se računa skjunis. Iz obrasca za skjunis vidimo da ovaj koeficijent u osnovi predstavlja odnos zbira kubnih odstupanja rezultata od njihove aritmetičke sredine i standardne devijacije dignute na treći stepen (ovaj ključni deo obrasca za skjunis uzorka ima istu formalnu strukturu kao i ključni deo obrasca ** za skjunis kao parametar distribucije diskretne slučajne varijable. (O statističkim razlozima zbog kojih umesto n u obrascu figuriraju složeni izrazi koji su funkcija od n biće reči u glavi **. Za sada je dovoljno uočiti da se ceo količnik koji sadrži n približava vrednosti 1/n kako n raste, te da je ovaj količnik praktično jednak 1/n za $n \geq 50$). Deljenje standardnom devijacijom dignutom ne treći stepen čini ovaj koeficijent relativnom merom i omogućuje upoređivanje njegove vrednosti za distribucije sa različitim mernim skalama. Predznak skjunisa, budući da su S i n po definiciji pozitivni brojevi, zavisi isključivo od zbira kubnih odstupanja u brojiocu obrasca za skjunis. Ukoliko su rezultati manji od aritmetičke sredine češći i(li) udaljeniji od nje nego rezultati veći od aritmetičke sredine, zbir negativnih kubnih odstupanja će biti veći od zbira pozitivnih kubnih odstupanja pa će skjunis imati negativni predznak, tj. biće manji od nule. Ukoliko su rezultati veći od aritmetičke sredine češći i(li) udaljeniji od nje nego rezultati manji od

[31] Skew u engleskom znači kos (u geometrijskom smislu) ili iskošen. Stoga se u pojedinim statističkim knjigama na našem jeziku za skjunis koriste i termini kooeficijent iskošenosti ili koeficijent zakrivljenosti.
[32] Različiti statistički programi koriste unekoliko različite verzije obrasca za računanje skjunisa. U ovom tekstu prikazan je obrazac za računanje koeficijenta asimetrije za distribuciju uzorka koji se koristi i u statističkom paketu SPSS.

aritmetičke sredine, zbir pozitivnih kubnih odstupanja će biti veći od zbira negativnih kubnih odstupanja pa će skjunis imati pozitivni predznak, tj. biće veći od nule. Ukoliko je distribucija rezultata simetrična učestalost i udaljenost rezultata manjih od aritmetičke sredine biće isti kao učestalost i udaljenost rezultata većih od aritmetičke sredine. U tom slučaju negativna i pozitivna kubna odstupanja rezultata od njihove aritmetičke sredine će se međusobno "poništiti" pa će zbir u brojiocu biti jednak nuli. Tada će i skjunis biti jednak nuli.

Pri tumačenju vrednosti skjunisa mogu se koristiti sledeća orijentaciona pravila:

- ✓ Ako je skjunis jednak 0, raspodela rezultata na varijabli je simetrična;
- ✓ Ako je skjunis manji od 0 ali nije manji od -0.5, distribucija rezultata je blago negativno asimetrična;
- ✓ Ako je skjunis između -0.5 i -1, distribucija rezultata je umereno negativno asimetrična;
- ✓ Ako je skjunis manji od -1, distribucija rezultata je znatno negativno asimetrična;
- ✓ Ako je skjunis veći od 0 ali nije veći od 0.5, distribucija rezultata je blago pozitivno asimetrična;
- ✓ Ako je skjunis između 0.5 i 1, distribucija rezultata je umereno pozitivno asimetrična;
- ✓ Ako je skjunis veći od 1, distribucija rezultata je znatno pozitivno asimetrična;

Dakle, negativan skjunis ukazuje na negativnu, a pozitivan skjunis na pozitivnu asimetričnost raspodele. U empirijskim raspodelama koje se sreću u psihologiji i srodnim oblastima asimetrične raspodele su najčešće takve da je mali broj znatno udaljenih rezultata s jedne strane aritmetičke sredine, a znatno veći broj gusto zbijenih rezultata sa druge strane aritmetičke sredine. Negativno asimetrična raspodela najčešće ima "rep" na levu stranu, tj. relativno mali broj rezultata manjih od aritmetičke sredine koji su znatno udaljeni od nje. S druge strane, pozitivno asimetrična raspodela najčešće ima "rep" na desnu stranu, tj. relativno mali broj rezultata većih od aritmetičke sredine koji su znatno udaljeni od nje.

Računanje skjunisa demonstriraćemo na primerima dva skupa podataka koje smo koristili pri demonstriranju računanja aritmetičke sredine i standardne devijacije. Za skup podataka

3, 4, 5, 5, 5, 5, 6, 7

skjunis bismo izračunali na sledeći način (M = 5.00, S = 1.19523, n = 8):

$Sk = (3 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (4 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (5 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (5 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (5 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (6 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] + (7 - 5)^3 / 1.19523^3 * [8 / (7 * 6)] = (-0.8924) + (-0.1116) + 0 + 0 + 0 + 0 + 0.1116 + 0.8924 = 0.$

Uočimo iz ovog računa kako doprinos vrednosti skjunisa zavisi od udaljenosti rezultata od aritmetičke sredine. Isto tako, budući da je distribucija simetrična, uočimo da su sabirci levo od nula identični po apslutnim vrednostima sabircima desno od četiri nule.

Za skup podataka u kojem smo rezultat 7 zamenili iznimkom 77, tj. za skup podataka

3, 4, 5, 5, 5, 5, 6, 77.

skjunis bismo izračunali ovako (M = 13.75, S = 25.57203, n = 8):

$Sk = (3 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (4 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (5 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (5 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (5 - 13.75)^3 /$

$25.57203^3 * [8 / (7 * 6)] + (5 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (6 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] + (77 - 13.75)^3 / 25.57203^3 * [8 / (7 * 6)] = (-0.0142) + (-0.0106) + (-0.0076) (-0.0076) + (-0.0076) + (-0.0076) + (-0.0053) + 2.8822 = 2.822.$
(Zagrade su stavljene i na pojedinim mestima na kojima to nije nužno kako bi čitalac lakše pratio obrazac).
Uočimo da je sada ogroman doprinos vrednosti skjunisa dao iznimak. Na prvi pogled može se učiniti da promena vrednosti skjunisa u odnosu na promenu aritmetičke sredine i standardne devijacije nije tako velika pod dejstvom iznimka. Međutim, pri tumačenju vrednosti skjunisa treba uzeti u obzir da nije reč o meri koja je iskazana na skali podataka već o relativnoj meri koja je dobijena deljenjem zbira kubnih odstupanja sa trećim stepenom standardne devijacije. Promena skjunisa sa 0 na 2.822 je stoga zapravo drastična promena. Dakle, distribucija ovih podataka je izrazito pozitivno asimetrična zbog prisustva jednog iznimka na desnoj strani raspodele, tj. jednog izrazito visokog rezultata.

Primer koji smo prikazali služio je samo za jednostavnije uočavanje matematičke prirode skjunisa. Pri korišćenju skjunisa treba voditi računa da uzorak na kojem računamo skjunis bude dovoljno veliki: poželjno je da bude veći od 50. Tumačenje skjunisa za uzorak manji od 50 je veoma rizičan poduhvat. Objašnjenje ove preporuke daćemo u glavi ** pri izlaganju oocenjivanja parametara.

Kada se koristi kao deskriptivna statistička mera, skjunis se uobičajeno zaokružuje na tri decimale.

Kurtozis (engl. Kurtosis)[33]

Kurtozis uzorka, u oznaci Ku, računa se primenom sledećeg obrasca:[34]

$$\mathrm{Ku}=\frac{\sum_{i=1}^{n}(x_i-M)^4}{S^4}\frac{n(n+1)}{(n-1)(n-2)(n-3)}-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Ako zanemarimo na trenutak izraze u obrascu za kurtozis u kojima figurira n, tj. veličina uzorka, uočavamo da je kurtozis uzorka u osnovi količnik zbira odstupanja rezultata od njihove aritmetičke sredine, pri čemu su odstupanja dignuta na četvrti stepen, i standardne devijacije rezultata dignute na četvrti stepen. (Ovaj ključni segment obrasca za kurtozis uzorka ima istu formalnu strukturu kao ključni segment obrasca ** za kurtozis kao parametar distribucije diskretne slučajne varijable. O statističkim razlozima zbog kojih umesto n u obrascu figuriraju složeni izrazi koji su funkcija od n biće reči u glavi **. Za

[33] Od grčkog *kurtosis* što znači konveksnost, izbočina ili ispupčenje.
[34] U statističkim tekstovima na našem jeziku često se sreću i nazivi „koeficijent izduženosti“ ili „koeficijent spljoštenosti“, a u tekstovima na engleskom jeziku „peakedness coefficient“ ili „flatness coefficient“, što je posledica veoma čestog, i u osnovi pogrešnog, tumačenja ove statističke mere. Naime, često se smatra da skjunis iznad određene vrednosti (0 ili 3, zavisno od oblika obrasca za skjunis) ukazuje na izduženu raspodelu, tj. raspodelu sa višim „vrhom“ (dužom najvišom ordinatom) u odnosu na normalnu raspodelu, dok se za skjunis manji od te vrednosti smatra da ukazuje na spljošteniju raspodelu (raspodelu sa nižim vrhom, tj. kraćom najvišom ordinatom) od normalne. Mi ćemo u ovom tekstu koristiti isključivo termin kurtozis budući da bi eventualni termin na našem jeziku usklađen sa značenjem ove mere morao biti složen i posledično nepraktičan (na primer „koeficijent razvučenosti i izdignutosti krajeva raspodele“).

sada je dovoljno uočiti da se izraz koji sadrži n a nalazi se levo od znaka minusa približava vrednosti 1/n, a izraz desno od znaka minusa približava vrednosti 3 kako se n povećava. Izraz koji sadrži n a nalazi se levo od znaka minusa praktično je jednak izrazu 1/n za n ≥ 50, dok se izraz desno od znaka minusa razlikuje za manje od 0.05 od vrednosti 3 za n ≥ 200). Deljenje standardnom devijacijom dignutom na četvrti stepen ima istu funkciju koju ima deljenje kubom standardne devijacije kod skjunisa, tj. definisanje ove mere kao relativne mere. Budući da je standardna devijacija po definiciji pozitivna i da negativna odstupanja dignuta na četvrti stepen imaju pozitivan predznak prvi količnik u obrascu sa zbirom u brojiocu uvek će biti pozitivan. Količnik u ovom obrascu u kojem figurira n a nalazi se desno od znaka minusa ima funkciju da obezbedi da vrednost kurtozisa za normalnu raspodelu bude jednaka nuli (vrednost prvog količnika, onog sa sumom u brojiocu, kada se podeli sa n, za normalnu raspodelu iznosi 3). Ovaj oblik kurtozisa, za razliku od oblika kurtozisa čija vrednost za normalnu raspodelu iznosi 3, mogao bi se zvati "kalibrisani kurtozis" (u tekstovima na engleskom jeziku on se uobičajeno zove "excess kurtosis" što bi doslovno bilo "prekomerni kurtozis"). U ovoj knjizi nećemo praviti ove terminološke distinkcije: pri objašnjenju analiza empirijskih podataka uvek ćemo koristiti "kalibrisani kurtozis" koji se računa prema obrascu ** i zvaćemo ga prosto kurtozis.[35].

U udžbenicima o primeni statistike kurtozis se najčešće tretira kao da jednoznačno govori o izduženosti, odnosno spljoštenosti raspodele:

- ako je kurtozis veći od nule smatra se da to govori o leptokurtičnoj distribuciji, tj. distribuciji koja je izduženija (ima viši vrh) od normalne distribucije;[36]
- ako je kurtozis jednak nuli reč je o mezokurtičnoj distribuciji kod koje je izduženost distribucije ista kao i kod normalne raspodele;[37]
- ako je kurtozis manji od nule smatra se da to govori o platikurtičnoj distribuciji, tj. distribuciji koja je spljošenija (ima niži vrh) od normalne.[38]

Međutim, eksperimentisanje sa dodavanjem rezultata u sredinu i na krajeve distribucije pokazuje da nije mnogo mudro samo na osnovu vrednosti kurtozisa zaključivati o obliku distribucije (cf. Chissom, 1970). Naime, ako se u normalnu distribuciju rezultata dodaju rezultati u centru distribucije, tj. oko vrednosti aritmetičke sredine tada se kurtozis raspodele zaista povećava od nule naviše. Ako se, pak, u normalnoj distribuciji rezultata ravnomerno dupliraju svi postojeći rezultati tada se kurtozis smanjuje i čak postaje negativan, iako se vrh raspodele očigledno povećava u odnosu na normalnu. Dakle, distribucija je i dalje leptokurtična (ima viši vrh u odnosu na normalnu) a kurtozis može da postane čak negativan. Isto tako, kurtozis sa savršeno simetričnu bimodalnu raspodelu (raspodelu sa modalnim vrednostima na krajevima raspodele, i bez ikakvih vrednosti između) iznosi -2, što je istovremeno i najniža moguća vrednost kurtozisa. S druge strane, kurtozis nije ograničen u pogledu najviše vrednosti (dokaz o najnižoj i najvećoj mogućoj vrednosti kurtozisa može se naći u Darlington, 1970). Očigledno, o obliku distribucije se ne može ništa precizno zaključiti samo na osnovu vrednosti kurtozisa. Generalno, za razliku od pojma lokacije, skale (varijabilnosti) i skjunisa (asimetričnosti) distribucije, pojam „kurtotičnosti" distribucije nije jednoznačan i prilično ga je teško interpretirati (o matematičkoj pozadini ove nejednoznačnosti i složenosti pojma kurtozisa može se videti u Dodge & Rousson, 1999). Poruka koju o kurtozisu treba zapamtiti jeste da u tumačenju

[35] Odluka o ovoj terminološkoj neiznijansiranosti doneta je na osnovu činjenice da se u ispisima iz analiza podataka u programu SPSS vrednosti „kalibrisanog kurtozisa" uvek pojavljuju uz naziv „Kurtosis". U tim ispisima se nigde ne pojavljuje precizan naziv „Excess kurtosis".
[36] Naziv leptokurtičan potiče od grčkog leptos što znači tanak.
[37] Naziv mezokurtičan potiče od grčkog mesos što znači srednji.
[38] Naziv platikurtčan potiče od grčkog platys što znači širok, ravan.

njegove vrednosti pri analizama realnih podataka treba biti veoma oprezan. Načelno, veoma visoke pozitivne i veoma niske negativne vrednosti kurtozisa upućuju svakako na veći stepen nagomilavanja podataka na krajevima raspodele, dok vrednosti veoma blizu nule sugerišu da se radi o raspodeli sa nagomilanim rezultatima u sredini raspodele i malim brojem rezultata na krajevima raspodele. Na kraju, kurtozis blizak vrednosti -2 može se uzeti kao signal nagomilavanja većine rezultata na oba kraja raspodele.

Iz ključnog dela obrasca za kurtozis (prvi izraz posle znaka jednakosti) može se jasno uočiti da na vrednost kurtozisa najveći uticaj imaju rezultati na krajevima raspodele (najniži i najviši rezultati): odstupanja ovih rezultata od aritmetičke sredine dignuta na četvrti stepen, budući da su ti rezultati udaljeniji od aritmetičke sredine od rezultata u sredini raspodele, doprinosiće mnogo više ukupnoj sumi u brojiocu nego rezultati u blizini aritmetičke sredine (precizna matematička razrada uticaja krajeva i središnjih delova distribucije na vrednost kurtozisa može se naći u Westfall, 2014). Dakle, što je više rezultata na krajevima raspodele i što su oni dalje od aritmetičke sredine to će kurtozis biti, pod ostalim jednakim uslovima, veći. Ovo je veoma važno uočiti radi pravilnog tumačenja vrednosti kurtozisa. Naime, ako na vrednost kurtozisa više utiču rezultati na krajevima raspodele nego oni iz sredine, ne može se očekivati da kurtozis neposredno i nedvosmisleno ukazuje na "izduženost" tj. visinu "vrha" raspodele. Istina, postoji izvesna veza između "izduženosti" distribucije i veličine kurtozisa (cf. Wheeler, 2011a) ali ova veza niti je direktna niti savršena, te se na osnovu veličine kurtozisa ne može izvući nedvosmislena informacija o "izduženosti" raspodele.

Pri tumačenju vrednosti kurtozisa dobijenih po obrascu ** mogu se koristiti sledeća orijentaciona pravila ali samo ako je reč o unimodalnoj raspodeli kod koje je modalna vrednost blizu sredine raspodele:

- ✓ Ako je kurtozis jednak nuli distribucija varijable je normalna;
- ✓ Ako je kurtozis veći od nule distribucija varijable ima razvučenije i(li) izdignutije krajeve nego što je to slučaj sa normalnom raspodelom;
- ✓ Ako je kurtozis manji od nule distribucija varijable ima manje razvučene i(li) manje izdignute krajeve od normalne raspodele.

Dakle, nizak kurtozis (niži od nule) ukazuje na to da takva unimodalna raspodela nema rezultate koji su izrazito udaljeni od aritmetičke sredine, dok visok kurtozis (znatno veći od nule) ukazuje na raspodelu u kojoj ima rezultata veoma udaljenih od aritmetičke sredine.

Računanje kurtozisa demonstriraćemo na primerima dva skupa podataka koje smo koristili radi demonstriranju računanja skjunisa. Za skup podataka bez iznimka kurtozis bismo izračunali na sledeći način (M = 5.00, S = 1.19523, n = 8):

$Ku = \{(3 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (4 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (6 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)] + (7 - 5)^4 / 1.19523^4 * [8 * 9 / (7 * 6 * 5)]\} - 3 * 7^2 / (6 * 5)$
$= (2.6880 + 0.16780 + 0 + 0 + 0 + 0 + 0.16780 + 2.6880) - 4.90 = 5.7120 – 4.90 = 0.812.$

Uočimo iz vrednosti u zagradi u poslednjem redu kako doprinos pojedinih rezultata kurtozisu zavisi od njihove udaljenosti od aritmetičke sredine. Isto tako, budući da je distribucija simetrična, uočimo da su sabirci u toj zagradi levo od četiri nule identični vrednostima sabiraka desno od tih nula. S druge strane, rezultati jednaki aritmetičkoj sredini uopšte ne doprinose vrednosti kurtozisa. Vrednost koja se oduzima od kurtozisa kako bi on za normalnu raspodelu bio jednak 0 nije jednaka 3, već je nešto veća. To je zbog toga što je uzorak veoma mali. U vrednost koja se oduzima uključena je, pored broja 3, i korekcija za veličinu uzorka koja je sadržana u vrednostima koje su u zagradi levo od

znaka minusa. Korekcija za veličinu uzorka za ovako male uzorke nije jednaka 1/n već je veća, ta je stoga nužno od dobijene vrednosti sa leve strane znaka minusa oduzeti broj veći od 3.

Za skup podataka u kojem smo rezultat 7 zamenili iznimkom 77, tj. za skup podataka

3, 4, 5, 5, 5, 5, 6, 77

skjunis bismo izračunali ovako (M = 13.75, S = 25.57203, n = 8):

$Sk = \{(3 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (4 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (5 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (6 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] + (77 - 13.75)^4 / 25.57203^4 * [8 * 9 / (7 * 6 * 5)] - 3 * 7^2 / (6 * 5) = (0.0107 + 0.072 + 0.0047 + 0.0047 + 0.0047 + 0.0047 + 0.0029 + 12.8320) - 4.90 = 7.972.$

Uočimo da je ogroman doprinos zbiru u zagradi u poslednjem i pretposlednjem redu (12.8320) dao jedan iznimak. Kao i kod skjunisa promena kurtozisa zbog jednog iznimka je ogromna, zapravo još izrazitija nego promena skjunisa. Dakle, u skladu sa vrednošću kurtozisa, distribucija ovih podataka ima u odnosu na normalnu raspodelu izrazito razvučen desni kraj zbog prisustva jednog iznimka.

Primer koji smo prikazali, kao što je to bilo i kod skjunisa, služio je samo za jednostavnije uočavanje matematičke prirode kurtozisa kao statističke mere. Kurtozis nema mnogo smisla računati za uzorak manji od 50, a za smisleno tumačenje kurtozisa poželjno je da uzorak bude veći od 100 ili, čak, 150. Objašnjenje ove preporuke daćemo u glavi ** pri izlaganju oocenjivanja parametara.

Kada se koristi kao deskriptivna statistička mera, kurtozis se uobičajeno zaokružuje na tri decimale.

Osetljivost skjunisa i kurtozisa na iznimke

Vrednost skjunisa i kurtozisa pod snažnim su uticajem iznimaka u podacima. Uticaj iznimaka na skjunis i kurtozis demonstrirali smo u prethodnim primerima. Međutim, važno je uočiti da će se skjunis najviše promeniti ukoliko iznimci postoje samo na jednoj strani raspodele, dok podjednak broj jednako udaljenih iznimaka sa jedne i druge strane aritmetičke sredine neće promeniti vrednost skjunisa. Jednostavnije rečeno, iznimci u simetričnoj raspodeli neće uticati na vrednost skjunisa. S druge strane, vrednost kurtozisa će se povećati pod uticajem iznimaka čak i kada su oni simetrično smešteni sa jedne i druge strane aritmetičke sredine.

Primer statističkog opisa uzorka u pogledu kvantitativne varijable na stvarnim podacima iz istraživanja

U delu teksta o pravljenju grupisane raspodele učestalosti prikazali smo raspodelu rezultata na Upitniku depresivnosti CES-D za slučajni uzorak od 252 onkološka pacijenta u našoj zemlji. Za statistički opis ovog uzorka u pogledu depresivnosti potrebno je, pored grupisane raspodele učestalosti, izračunati odgovarajuće mere lokacije (percentile i mere centralne tendencije), varijabilnosti i oblika raspodele. U ovom slučaju te mere izračunate

su uključivanjem određenih opcija u proceduri **Frequencies** u programu SPSS.[39]Pored statističkih mera prikazaćemo ponovo grupisanu raspodelu učestalosti sa širinom razrednog intervala jednakom 4, koju smo pri objašnjavanju pravljenja raspodele dabrali kao najadekvatniju, radi jednostavnijeg praćenja komentara ispisa.

Ispis iz programa SPSS izgleda ovako:

Statistics

CES_D_skala depresivnosti-ukupni rezultat

N	Valid	249
	Missing	3
Mean		15.01
Median		14.00
Mode		12
Std. Deviation		9.726
Skewness		.765
Std. Error of Skewness		.154
Kurtosis		.412
Std. Error of Kurtosis		.307
Minimum		0
Maximum		48
Percentiles	25	8.00
	75	20.00

[39] Korišćenje procedura za statistički opis uzorka u pogledu kvantitativne varijable u programu SPSS čitalac može naučiti sledeći video instrukcije br.3...**

Rezultat na upitniku depresivnosti CES_D grupisano1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-3	22	8.7	8.8	8.8
	4-7	38	15.1	15.3	24.1
	8-11	31	12.3	12.4	36.5
	12-15	58	23.0	23.3	59.8
	16-19	31	12.3	12.4	72.3
	20-23	21	8.3	8.4	80.7
	24-27	18	7.1	7.2	88.0
	28-31	13	5.2	5.2	93.2
	32-35	7	2.8	2.8	96.0
	36-39	5	2.0	2.0	98.0
	40-43	3	1.2	1.2	99.2
	44-47	1	.4	.4	99.6
	48-51	1	.4	.4	100.0
	Total	249	98.8	100.0	
Missing	System	3	1.2		
Total		252	100.0		

Iz ispisa (iz reda **Missing** u grupisanoj raspodeli) vidimo da postoji zanemarljiv procenat podataka koji nedostaju (1.2%). Aritmetička sredina (**Mean**) jednaka 15.01, medijana (Median) je nešto niža i iznosi 14, a mod (**Mode**) je 12. Na osnovu nejednakosti aritmetičke sredine i medijane uviđamo da distribucija nije simetrična. (Međutim, da su kojim slučajem aritmetička sredina, medijana i mod bili jednaki ne bismo stopostotno smeli da zaključimo da je distribucija simetrična jer ove tri mere mogu biti jednake i kod posebnih vrsta asimetričnih raspodela). Pregledanjem grupisane raspodele uočavamo da se distribucija znatno dalje proteže (ima više razreda) od razrednog intervala u kojem je aritmetička sredina (razreda 12–15) ka višim nego ka nižim rezultatima. Isto tako, učestalosti u razrednim intervalima sa rezultatima većim od aritmetičke sredine znatno su niže nego u intervalima sa rezultatima manjim od aritmetičke sredine. Isto tako, na osnovu vrednosti skjunisa (**Skewness**) od 0.765 zaključujemo da je raspodela umereno pozitivno asimetrična, dok na osnovu vrednosti kurtozisa (**Kurtosis**) zaključujemo da je distribucija razvučenija od normalne raspodele, a pregledom raspodele vidimo da je učestalost visokih rezultata znatno udaljenih od aritmetičke sredine relativno visoka. Znači, u poređenju sa normalnom krivom ova distribucija ima nešto izdignutiji desni "rep" (onaj ka višim vrednostima).

Standardna devijacija (**Std. Deviation**) je dosta visoka i iznosi 9.73, što je, pre svega, posledica postojanja malog broja veoma visokih rezultata (u intervalima 40–43, 44–

47 i 48–51). Očigledno, aritmetička sredina ne reprezentuje baš najbolje sve mere u raspodeli.

S obzirom na asimetričnost raspodele, bolji statistički opis ove raspodele dobili bismo korišćenjem petobrojnog sažetka: najniži rezultat (**Minimum**) je 0, percentil 25 (**Percentiles 25**) jednak je 8, medijana iznosi 14, percentil 75 (**Percentiles 75**) jednak je 20, a najviši rezultat (**Maximum**) je 48. Dakle, 25% ispitanika ima rezultat veći od 20 što je veće od granične vrednosti 16 koja ukazuje na početak klinički relevantnog nivoa depresivnosti. Pet ispitanika imaju rezultat iznad 40 što je veoma zabrinjavajuće budući da ovde nije reč o psihijatrijskoj populaciji.

Dakle, na osnovu prikazanog statističkg opisa mogli bismo reći da je ovde reč o uzorku onkoloških pacijenata sa znatno izraženom depresivnošću. Na osnovu podataka koje smo prikazali još uvek ništa ne zaključujemo o depresivnosti u populaciji onkoloških pacijenata iz koje je ovaj uzorak. O zaključivanju o stanju u populaciji na osnovu statističkih mera dobijenih na slučajnom uzorku biće reči u glavi **. Tada će biti jasnije i kakve informacije u sebi sadrže statistici **Std. Error of Skewness** i **Std. Error of Kurtosis** koje za sada nećemo uopšte uzeti u razmatranje.

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

Chissom, B. S. (1970). Interpretation of the Kurtosis Statistic. *The American Statistician*, *24*(4), 19–22.

Darlington, R. B. (1970). Is Kurtosis Really "Peakedness?".*The American Statistician*, *24*(2), 19–22.

Dodge, Y, & Rousson, V. (1999). The Complications of the Fourth Central Moment. *The American Statistician*, *53*(3), 267–269.

Gordon, T. (1986). Is the standard deviation tied to the mean? *Teaching Statistics*, *8*(2), 40–42.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the RealWorld. *Annual Review of Psychology* , *60*, 549–576. doi: 10.1146/annurev.psych.58.110405.085530

Jones, D. L., & Scariano, S. M. (2014). Measuring the variability of data from other values in the set. *Teaching statistics*, 36(3), 93–96.

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. Organizational Research Methods, 17(4) 372–411. DOI: 10.1177/1094428114548590

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology*, *57*(1), 1–10. DOI: 10.1037/a0018082

Tukey, J. W. (1969). Analyzing data:sanctification or detective work? American Psychologist, 24, 83–91.

von Hippel, P. T. (2005). Mean, median, and skew: correcting a textbook rule. *Journal of Statistics Education*, *13*(2). www.amstat.org/publications/jse/v13n2/vonhippel.html

Westfall, P. H. (2014). Kurtosis as peakedness, 1905 – 2014. *American Statistician, 68*(3), 191–195. doi:10.1080/00031305.2014.917055.

7. Grafičko prikazivanje podataka na jednoj kvantitativnoj varijabli

Zašto je važno grafičko prikazivanje podataka

Ako bismo parafrazirali staru kinesku poslovicu o tome da "slika vredi koliko i hiljadu reči", mogli bismo reći da statistički grafički prikazi ponekad vrede na hiljade brojeva. Iako većinu sažetih statističkih opisa skupa podataka možemo predstaviti i tabelarno i grafički, grafički prikaz podataka može imati neke prednosti nad tabelarnim prikazom. Grafički predstavljeni podaci postaju pregledniji i bolje se razumeju. Jedan pogled na grafik ponekad više govori od dugotrajnog pregledanja kolona sa cifarskim podacima. Grafičkim prikazom možemo složenu strukturu podataka pretvoriti u sliku koju je lako razumeti zahvaljujući urođenoj sposobnosti čoveka da opaža vizuelne složajeve. Osim toga, grafici ponekad omogućuju potpunije razumevanje globalne strukture podataka nego tabele. Grafici omogućuju da uočimo pojedina suptilna svojstva podataka koja bismo iz njihovog tabelarnog prikaza, ma koliko bio savršen, teže uočili. Takođe, pomoću grafika možemo lakše i brže međusobno upoređivati razlike u složajevima različitih struktura podataka što bi bez grafičkog prikaza bilo veoma teško, sporo ili pokatkad nemoguće. Grafici mogu biti dragoceni u upoređivanju strukture empirijskih podataka sa teorijskim modelom. Premda se u statističkim tabelama nalaze precizne brojčane vrednosti koje sa grafika nije lako uočiti često se fini složajevi i pravilnosti u podacima ne mogu uočiti iz tabela, barem ne tako lako kao iz grafika. Osim što se lakše opažaju, grafički prikazi su ponekada i ubedljiviji (pa i zanimljiviji) u prenošenju ključnih informacija od tabelarnih prikaza. Grafici mogu u većoj meri od tabela, svojom zanimljivošću, podstaći na analitičko razmišljanje o podacima. Danas je, zahvaljujući specijalizovanim softverima, postalo praktično uobičajeno eksplorisati i grafički prikazivati velike baze podataka koje postoje na Internetu u nastojanjima da se otkriju zanimljivi složajevi u njima /čak postoji čitava oblast primene statistike koja se zove "rudarenje po podacima" (engl. data mining)/. U psihologiji se već dovoljno zna o ljudskoj percepciji i te se zakonitosti mogu mudro upotrebiti u dizajniranju statističkih grafičkih prikaza. Prema istraživanjima sprovedenim u SAD postoji veoma snažna pozitivna veza između "tvrdoće", tj. "naučne strogosti" oblasti i stepena korišćenja statističkih grafičkih prikaza: naučne oblasti sa većom "naučnom strogošću" (fizika, hemija…) znatno više koriste grafičke statističke prikaze podataka od "mekših" nauka (sociologija, ekonomija, psihologija). Ista takva veza postoji između "naučne strogosti" oblasti i korišćenja statističkih grafičkih prikaza i unutar psihologije: ovakvi prikazi znatno su češći u časopisima iz "naučno strožih" oblasti psihologije (na primer, bihejvioralne neuronauke i eksperimentalna psihologija) nego u časopisima iz tzv. "mekših" oblasti (na primer, savetodavna, obrazovna, klinička psihologija). Kada je reč o korišćenju statističkih tabela, situacija je potpuno obrnuta! Što je još zanimljivije, ova razlika nije posledica razlike u prostoru u časopisima koji su posvećeni prikazu podataka – ovaj prostor je približno jednak u svim oblastima – niti stepena kvantifikacije ili korišćenja statističkih procedura (Smith, Best, Stubbs, Bastiani Archibald, & Roberson-Nay, R., 2002). Može se, dakle, reći da su statistički grafički prikazi korisni:

a) za sažeto prikazivanje podataka;
b) kao pomoćno dijagnostičko sredstvo pri statističkoj analizi podataka i
c) kao moćno sredstvo u tumačenju dobijenih rezultata i dolaženja do novih ideja za dalja istraživanja.

U tom smislu, Vajner i Veleman ističu da su savremeni statistički grafički prikazi podataka "aktivni učesnici u procesu naučnog otkrića" i "dinamički partneri i pre vodiči u budućnost nego što su statički spomenici prošlih otkrića" (Wainer & Velleman, 2001, str. 305).

U samom istraživačkom procesu statistički grafički prikazi mogu poslužiti na sledeće načine (prema Cox, 1978):

a) kao pomoćno sredstvo u eksploraciji podataka: grafičko prikazivanje nam služi da uočimo složajeve, pravilnosti, neregularnosti, saobraznost raspodele podataka sa nekim teorijskim modelom;
b) kao ključno sredstvo u odgovaranju na neko specifično istraživačko pitanje koje je precizno formulisano u skladu sa određenim teorijskim modelom;
c) kao sredstvo za predstavljanje zaključaka do kojih se došlo istraživanjem.

Na žalost, u ovoj knjizi nećemo koristiti statističke grafičke prikaze podataka onoliko koliko bismo to želeli, pogotovu ne one u bojama, jer bi to poskupelo njeno izdavanje. Nastojaćemo da na odgovarajućim mestima ukažemo na osnovne principe koje treba poštovati u statističkim grafičkim prikazima podataka i koristićemo grafičke prikaze uglavnom u situacijama kada bi njihovo zamenjivanje brojčanim (tabelarnim) prikazivanjem veoma otežalo prenošenje željene poruke. Budući da će korisnici ove knjige svakako statističke grafičke prikaze podataka praviti korišćenjem grafičkih ili statističkih paketa preporučujemo stav apriornog neprihvatanja podrazumevane (engl. default) opcije u takvim paketima. Dobar i smislen grafik podrazumeva gotovo uvek editovanje ili intervenisanje na grafiku koji se automatski dobije iz statističkog paketa.

Ono što smo u prethodnim pasusima rekli o grafičkom prikazivanju podataka nikako ne treba shvatiti kao zalaganje za zamenjivanje tabelarnog (brojčanog) prikazivanja podataka grafičkim prikazima. Svaki od ovih prikaza ima svoje prednosti i nedostatke. Tabelarno prikazivanje je u mnogim situacijama jedino moguće. Osim toga, u situacijama kada je neophodno prikazati precizne brojčane vrednosti i sačuvati ih za buduće analize i poređenja sa drugim istraživanjima tabelarno prikazivanje nije smisaono zamenjivati grafičkim.

Opšta uputstva za grafičko prikazivanje podataka

Pre objašnjavanja konkretnih statističkih grafičkih prikaza podataka navodimo opšta uputstva koja treba imati na umu pri statističkom grafičkom prikazivanju podataka, a posebno kada se ti grafički prikazi koriste za publikovanje dobijenih rezultata (prema Cox, 1978, str.8):

1. Ose treba da budu jasno označene imenima varijabli i sa oznakama jedinice mere
2. Prekidi skala treba da budu korišćeni za tzv. "lažne" početke, tj. za početak distribucije kada ona prirodno ne počinje od nule;
3. Poređenje povezanih dijagrama treba da bude pojednostavljeno korišćenjem identične skale merenja i poravnavanjem dijagrama jednog uz drugog.
4. Skale treba da budu aranžirane tako da sistematska ili približno linearna veza bude prikazana pod uglom od 45° u odnosu na X osu.
5. Legenda treba da učini grafik što jasniji sam po sebi nezavisno od osnovnog teksta van grafika
6. Tumačenje ne treba da bude prejudicirano tehnikom prezentacije, na primer postavljanjem podebljanih izglađenih krivih na dijagramu raspršenja na kojem su tačke bledo prikazane.

U ovoj glavi prikazaćemo samo one statističke grafičke postupke koji se često koriste u prikazivanju raspodele rezultata na jednoj kvantitativnoj varijabli ili u eksplorisanju podataka sa takve varijable.

Grafički prikazi podataka na teorijski kontinuiranoj varijabli

Tri osnovna grafika za prikazivanje raspodele rezultata na jednoj teorijski kontinuiranoj kvantitativnoj varijabli su: histogram (poligon stubaca), poligon učestalosti i grafik kumulativnih frekvencija. Dakle, ova tri postupka ima smisla koristiti za varijable koje su teorijski kontinuirane, kakve su primerice uzrast, visina, inteligencija, ekstraverzija, depresivnost. Naravno, empirijski podaci i za teorijski kontinuirane varijable su nužno diskretni, budući da kontinuum postoji samo u teoriji.

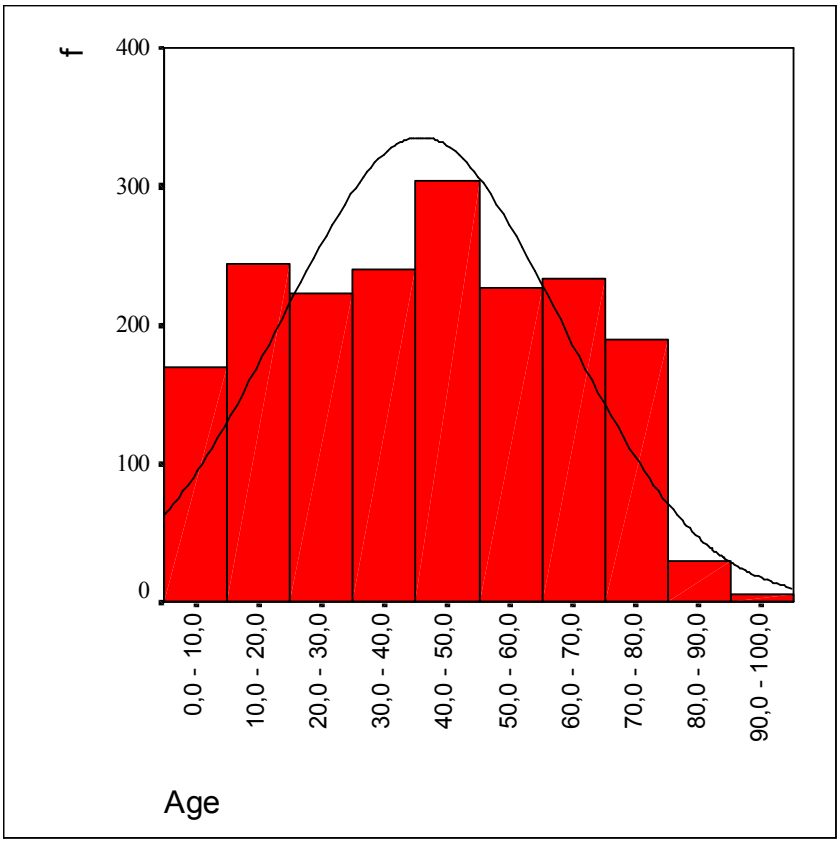
Histogram ili poligon stubaca

Za grafički prikaz podataka na jednoj kvantitativnoj varijabli najčešće se koristi histogram (poligon stubaca). Naziv ovog grafika potiče od grčkih reči histos (izduženi vertikalni oblici, tkački razboj) i gram (nešto što je napisano). (U stvari, logičnije bi bilo da se ovaj grafik zove histograf a ne histogram). Pri konstruisanju histograma u koordinatnom sistemu na apscisu se nanose donje i gornje egzaktne granice grupnih intervala, a na ordinatu frekvencije. Iznad segmenta [D, G] na apscisi, koji je definisan donjom (D) i gornjom (G) egzaktnom granicom razreda, konstruiše se pravougaonik čija širina odgovara veličini grupnog intervala a čija visina frekvenciji za dati grupni interval. (Određivanje egzaktnih granica i veličinu grupnog intervala objasnili smo u delu teksta u ovoj glavi pod naslovom "Raspodela sa grupnim intervalima") Na taj način, budući da su svi intervali jednake veličine poređenjem površine stubaca, tj. pravougaonika možemo porediti učestalosti rezultata u različitim intervalima. Dakle, frekvencije ili relativne frekvencije rezultata u određenom grupnom intervalu predstavljene su na histogramu površinom stubića iznad tog intervala.

U programu SPSS, preko histograma je moguće superponirati crtež normalne krive sa parametrima (μ i σ) koji odgovaraju aritmetičkoj sredini i standardnoj devijaciji rezultata prikazanih histogramom. Na taj način vizuelno se može uočiti stepen odstupanja distribucije rezultata od normalne raspodele.

.

Primer : Histogram za varijablu starost izgledao bi ovako:



Iz grafika treba uočiti dve bitne stvari:

- Apscisa (rezultati ispitanika na varijabli su u intervalima razreda) i ordinata (učestalost mera po razredima – **f**) su označene odgovarajućim nazivima.To je obavezno u svakom grafičkom prikazu podataka!
- Prikazani su samo rezultati ispitanika za koje imamo podatke, frekvencija podataka koji nedostaju ne unosi se nigde.

Poligon učestalosti

Pri konstruisanju poligona učestalosti u koordinatnom sistemu, na apscisu se nanose donje i gornje egzaktne granice, kao i srednja mesta grupnih intervala. (Određivannje srednjeg mesta grupnog intervala objasnili smo u delu teksta u ovoj glavi pod naslovom "Raspodela sa grupnim intervalima"). Na ordinati poligona učestalosti su frekvencije. Poligon učestalosti dobijamo tako što pravim linijama povežemo tačke čije su koordinate srednje mesto intervala i frekvencija rezultata u datom grupnom intervalu. Dakle, tačke koje povezujemo linijama predstavljaju ortogonalne projekcije vrednosti sa apscise i ordinate. Preporučljivo je dodati sa obe strane grupisane raspodele još po jedan grupni interval kako bi površina koju zatvara poligon predstavljala ukupan broj rezultata u raspodeli (cf. Dragićević, 2002).

Kriva kumulativnih procenata ili ogiva

Ovaj grafik dobijamo tako što pravim linijama povežemo tačke čije su koordinate gornja egzaktna granica grupnog intervala i relativna kumulativna frekvencija za dati grupni interval.

Grafički prikazi podataka na teorijski diskretnoj kvantitativnoj varijabli

Za grafičko prikazivanje podataka na teorijski diskretnoj kvantitativnoj varijabli najčešće se koristi štapićasti dijagram.

Štapićasti dijagram (engl. Barchart)

Kod štapićastog dijagrama nema prave apscise u matematičkom smislu, budući da je reč o teorijski diskretnim varijablama, već se na horizontalnoj liniji nanose grupni intervali u *mernim* granicama – ako se grafički predstavlja grupisana raspodela, ili pojedinačne mere – ako se grafički predstavlja jedinična raspodela. Između različitih grupnih intervala, odnosno različitih pojedinačnih mera treba da postoji određeni razmak kako bi bilo jasno da je reč o diskretnoj varijabli. Iznad intervala, odnosno pojedinačne mere ucrtava se vertikalni štapić ili stubac tako da njegova visina odgovara ukupnoj frekvenciji svih mera obuhvaćenih intervalom, odnosno frekvenciji pojedinačnih mera. Poželjno je da svi razmaci između na štapićastom dijagramu budu jednaki kako bi se lakše uočio oblik distribucije.Visinom štapića moguće je, umesto običnih frekvencija predstaviti i relativne frekvencije. Bez obzira na to što liči na histogram štapićasti dijagram nikako ne treba mešati sa histogramom: histogram pretpostavlja postojanje apscise u matematičkom smislu, tj. kontinuirane apscisne ose, dok na štapićastom dijagramu ta kontinuirana osa realno ne postoji već se na horizontalnoj liniji nalik apscisnoj osi nanose grupni intervali, odnosno pojedinačne mere. Pored toga, stubići na štapićastom dijagramu mogu biti manje ili više razmaknuti, dok su stubići na histogramu nužno spojeni, tj. njihove osnove se nastavljaju jedna na drugu čineći tako kontinuum. Ukupna površina svih stubića na histogramu na taj način predstavlja ukupnu učestalost svih mera.

Štapićastim dijagramom moguće je grafički predstaviti i raspored učestalosti po kategorijama kategoričke varijable, o čemu će biti reči u narednoj glavi. Prilikom prikazivanja štapićastog dijagrama za kategoričke varijable daćemo i dodatna uputstva kojih se treba držati pri korišćenju ovog dijagrama.

Ponekada se umesto štapića ucrtavaju tačke na kraju ordinate za obične ili za relativne frekvencije i potom se te tačke spajaju u tzv. linijski dijagram. Smatramo da takvi dijagrami nisu baš najpodesniji za grafičko predstavljanje teorijski diskretnih varijabli jer posmatraču neopravdano sugerišu da postoji kontinuiranost u vrednostima na varijabli.

Grafički prikazi za eksploratornu analizu podataka na kvantitativnoj varijabli

Ueksploratornoj analizi podataka najčešće se koriste stablogram i kutijasti dijagram.

Stablogram

Stablogram ili prikaz stabljika sa lišćem (engl. stem and leaf plot) nije pravi grafički prikaz. On se sastoji iz dva skupa brojeva razdvojenih vertikalnom linijom. Svaki rezultat se deli na stabljiku (vodeću cifru) i list (prateću cifru). Sa leve strane su "vodeće

cifre", a sa desne strane svake vodeće cifre su preostale ("prateće") cifre određenih mera (čija je to vodeća cifra). Tako npr., ako su rezutati na nekoj varijabli 11, 13, 23, 23, 23, 24, 26, 29, 31, 33 prikaz će izgledati ovako:

Steam	Leaf
1	13
2	333469
3	13

Rezultati prikazani na stablogramu očitavaju se tako što se kombinuje vodeća cifra u koloni **Steam** sa pojedinačnim ciframa u koloni **Leaf**. Na primer, u drugom redu (sa vodećom cifrom 2) prikazani su sledeći rezultati: 23, 23, 23, 24, 26 i 29.

Primer:

Stablogram za varijablu starost (n = 1865):

Age Stem-and-Leaf Plot
Frequency Stem & Leaf

65,00 0 . 0011112223334444
105,00 0 . 555556666667777788999999
113,00 1 . 000001111122223333344444444
131,00 1 . 55555666666777777788888889999999
116,00 2 . 0000000111122223333334444444
106,00 2 . 55555556666777777888889999
117,00 3 . 000000011111122222233334444
123,00 3 . 5555555666666777777888899999
132,00 4 . 0000011111112222223333333444444
172,00 4 . 555555566666666677777777778888888889999999
132,00 5 . 000000000111111222222222333333444
95,00 5 . 5555556666677788888999
95,00 6 . 00011112222333333444444
138,00 6 . 555556666666677777778888999999999
123,00 7 . 000000011111222222233333334444
66,00 7 . 5555556666777789
26,00 8 . 001224&
4,00 8 . &
5,00 9 . &
1,00 9 . &

Stem width: 10
Each leaf: 4 case(s)

& denotes fractional leaves.

Uočimo da je u ovom prikazu sadržano ono što imamo u raspodeli učestalosti (kolona **Frequency**) kao i ono što vidimo iz histograma (listovi grade oblik raspodele). Na dnu prikaza objašnjeno je kolika je širina "stabljike" (kolika je razlika sukcesivnih vrednosti u koloni **Stem**) i koliko rezultata je prikazano jednom cifrom u desnom odeljku **leaf**. U ovom slučaju rezultati u koloni **Stem** razlikuju se za jednu deseticu (**Stem width: 10**), a svaka cifra u delu za „lišće" stoji za četiri ispitanika (**Each leaf: 4 case(s))**. Ukoliko je broj rezultata koji imaju istu vodeću cifru veliki tada se za svaku vodeću cifru može koristiti više od jednog reda. U prikazanom primeru svaka vodeća cifra zauzima dva reda u prikazu.

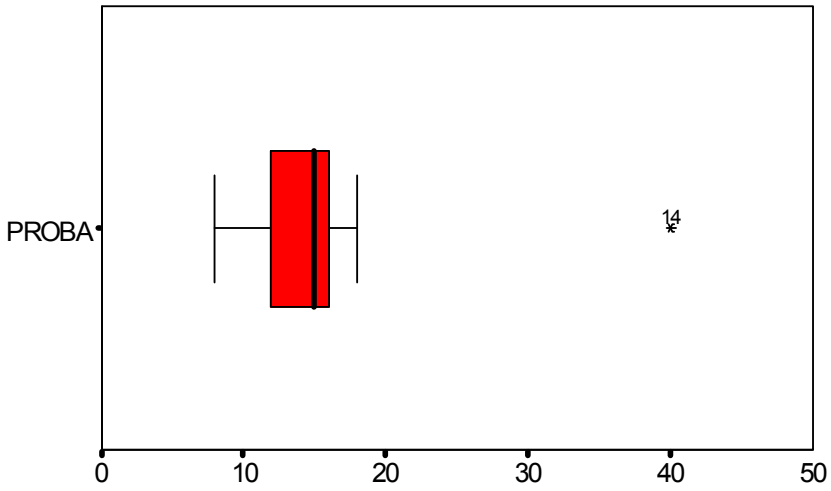
Stablogram je najpogodnije koristiti u situacijama kada broj rezultata nije izrazito veliki (na primer, broj rezultata nije veći od 200). Za prikaz raspodela sa velikim brojem rezultata pogodnije je koristiti kutijasti dijagram.

Kutijasti dijagram (engl. Boxplot ili Box and whisker plot)

Kutijasti dijagram prikazuje distribuciju kvantitativne varijable kroz tri komponente:

1. Središnja linija kutije prikazuje centralnu tendencu, najčešće Medijanu. Na osnovu položaja ove linije unutar kutije može se zaključiti da li je distribucija simetrična (linija je na sredini kutije) ili asimetrična (linija je bliže levoj ili desnoj ivici kutije).
2. Kutija, čije ivice približno odgovaraju trećem i prvom kvartilu (utoliko više što je uzorak veći). Lokacija rezultata na kojima će biti gornja i donja ivica kutije se računaju tako da se na lokaciju medijane [(n+1)/2] doda jedinica i dobijeni rezultat podeli sa 2. Ivice se zatim ucrtavaju tako da odgovaraju k-tom najmanjem i k-tom najvećem rezultatu gde je k lokacija (zaokružena na ceo broj) ivica dobijena na opisani način. H-opseg je raspon između gornje i donje ivice kutije.
3. Kanapi ili brkovi, tj. linije sa jedne i druge strane kutije ucrtavaju se do najniže i najviše okolinske vrednosti. Donja okolinska vrednost je najmanji rezultat koji je veći ili jednak donjoj ogradi, a gornja okolinska vrednost je najveći rezultat koji je manji ili jednak gornjoj ogradi. Donja ograda se računa tako što se od donje ivice oduzme vrednost koja je jednaka 1.5*H-opsega, a gornja ograda se dobija tako što se na gornju ivicu doda 1.5*H-opseg.
4. Zvezdicama se ucrtavaju rezultati koji su veći od gornje i manji od donje okolinske vrednosti. Tako se vizuelno uočavaju vangranične vrednosti –autlajeri (eng. outliers)

Primer Za sledeći niz brojeva (rezultati na varijabli PROBA): 8,10,12,12, 14,14,15,15,15,15, 16,17,18,40 ovaj grafik bi izgledao ovako:



Lokacija medijane je (14+1)/2 pa je medijana na sredini između sedmog i osmog rezultata i iznosi 15. Medijana je na grafiku prikazana kao linija koja preseca kutiju. Iz činjenice da ova linija u prikazanom primeru nije podjednako udaljena od obe ivice kutije možemo zaključiti da raspodela nije simetrična. Lokacija ivica kutije je (7.5+1)/2 pa je gornja ivica kutije četvrti rezultat počev od najvišeg (to je ovde skor 16) a donja ivica četvrti rezultat počev od najnižeg (u ovom slučaju 12). Gornja i donja ivica su veoma blizu prvog (ovde je on 12) i trećeg kvartila (koji je ovde 16.25) pa se može reći da kutija predstavlja interkvartilni raspon. H-opseg je 16 - 12 = 4 pa je donja ograda 12 - 1.5*4 = 6 a gornja 16 + 1.5*4 = 22. Donja okolinska vrednost je 8, a gornja 18. Rezultat 40 je prikazan zvezdicom i označen svojim rednim brojem u nizu prikazanih rezultata. Očigledno se radi o autlajeru, tj. iznimku.

Bickel, P. J., & Lehmann, E. L. (1975). Descriptive statistics for nonparametric models II. Location. *The Annals of statistics*, *3*(5), 1045–1069.

Box, D. R. (1978). Some remarks on the role in statistics of graphical methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 27*(1), 4–9.

Emerson, J., & Hoaglin, D. C. (1983). Stem-and-leaf displays. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 7–32). New York: John Wiley & Sons, Inc.

Fajgelj, S. (20**). *Psihometrija, Metod i teorija psihološkog merenja, ** izdanje*, Beograd: Centar za primenjenu psihologiju.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka &W. F. Velicer (Eds.), *ResearchMethods in Psychology: Vol. 2. Handbook of psychology,* pp. 87–114). New York, NY: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons, Inc.

Marmolejo-Ramos, F., & Matsunaga, M. (2009). Getting the most from your curves: Exploring and reporting data using informative graphical techniques. *Tutorials in Quantitative Methods for Psychology*, *5*(2), 40–50.

Moore, D. S., & McCabe, G. P. (1998). *Introduction to the practice of statistics*, *Third edition*. New York: W. H. Freeman and Company.

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, 17*(4) 372–411.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297–338). New York: John Wiley & Sons, Inc.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology*, *57* (1), 1–10. DOI: 10.1037/a0018082

Smith, L. D., Best, L. A., Stubbs, D, A., Bastiani Archibald, A., & Roberson-Nay, R. (2002). Constructing Knowledge, The Role of Graphs and Tables in Hard and Soft Psychology. *American Psychologist*, *57*(10), 749–761. DOI: 10.1037//0003-066X.57.10.749

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: mapping the pathways of science. *Annual Review of Psychology, 52*, 305–335.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32,* 191–241.

Watier, N. N., Lamontagne, C., & Chartie, S (2011). What does the mean mean? *Journal of Statistics Education*, *19*(2). www.amstat.org/publications/jse/v19n2/watier.pdf

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Amsterdam: Elsevier Inc.