**Sveučilište u Mostaru**

**Fakultet prirodoslovno-matematičkih i odgojnih znanosti  
Studij Informatike**

Gabrijel Batista

**DESKRIPTIVNA STATISTIKA**

Završni rad

Mostar, 2019.

**Sveučilište u Mostaru**

**Fakultet prirodoslovno-matematičkih i odgojnih znanosti  
Studij Informatike**

Gabrijel Batista

**DESKRIPTIVNA STATISTIKA**

Završni rad

Mentor: dr.sc. Sanja Tipurić-Spužević  
Student: Gabrijel Batista

Mostar, listopad 2019.

**SAŽETAK**

Deskriptivna statistika je dio statistike koji se bavi organizacijom i prikazivanjem podataka koji su prikupljeni provođenjem nekih istraživanja.

Prikupljeni podaci mogu se grupirati u skupine ovisno o njihovim karakteristikama. Kako bi proces organizacije i prikazivanja tih podataka bio jednostavniji, te skupine podataka mogu biti predstavljene uz pomoć jedne odgovarajuće središnje vrijednosti. Te središnje vrijednosti nazivamo mjerama centralne tendencije, a to su aritmetička sredina, centralna vrijednost (medijan), dominantna vrijednost (mod), geometrijska sredina i harmonijska sredina.

Pod deskriptivnu statistiku ubrajaju se i mjere disperzije, odnosno mjere raspršenja. Najčešće korištene mjere disperzije su varijanca, raspon rezultata, standardna devijacija i koeficijent varijabilnosti. Ove mjere računaje se kako bi se dobile informacije o međusobnom razlikovanju rezultata unutar neke skupine.

Prikupljeni podaci mogu se prikazati uz pomoć tehnike tabličnog prikaza podataka koji podrazumjeva prikazivanje podataka putem podatkovnih tablica i tehnike grafičkog prikaza podataka koji podatke prikazuje slikovito.

**SADRŽAJ**

[1. Uvod 4](#_Toc24299101)

[2. Mjere centralne tendencije 5](#_Toc24299102)

2.1. Aritmetička sredina....................................................................................5

2.2. Medijan......................................................................................................6

2.3. Mod............................................................................................................8

2.4. Geometrijska sredina..................................................................................9

2.5. Harmonijska sredina.................................................................................10

[3. Organiziranje podataka i prikazivanje rezultata 12](#_Toc24299103)

3.1. Grupiranje podataka.................................................................................13

3.2. Tablični prikaz podataka..........................................................................16

3.2. Grafički prikaz podataka..........................................................................18

3.2.1. Kružni dijagram............................................................................19

3.2.2. Stupčani dijagram.........................................................................20

3.3.3. Poligon frekvencija.......................................................................21

[4. Mjere disperzije 23](#_Toc24299104)

4.1. Raspon rezultata.......................................................................................23

4.2. Varijanca..................................................................................................24

4.3. Standardna devijacija...............................................................................25

4.4 Koeficijent varijabilnosti..........................................................................27

[5. Zaključak 28](#_Toc24299105)

[6. Literatura 29](#_Toc24299106)

1. Uvod

Statistika je matematička disciplina koja se bavi analizom podataka, odnosno

proučavanjem načina za prikupljanje, sažimanje i prikazivanje zaključaka iz nekih podataka. Prema osnovnoj podjeli dijeli se na inferencijalnu i deskrpitivnu statistiku.

U ovom završnom radu fokus je na deskriptivnoj statistici. Ovaj dio matematičke statistike se bavi uzorkom, dok izravno o populaciji govori vrlo malo, odnosno bavi se prikazivanjem osnovnih statističkih vrijednosti uz pomoć grafičkog i tabličnog prikaza, te mjerama disperzije i centralne tendencije. Deskriptivna statistika je obično prvi korak pri statističkoj analizi podataka, a svrha ove grane statistike je da uz pomoć brojeva opiše značenje podataka.

Zadatak statističke analize je izvođenje zaključaka o osnovnom skupu na temelju podataka iz uzorka. Faze statističke analize su statističko posmatranje, što se odnosi na plansko prikupljanje podataka, sređivanje podataka, tj. tablično i grafičko prikazivanje, te obrada i analiza rezultata, odnosno matematička obrada podataka i njihovo tumačenje.

Tehnike despriptivne statistike kao što su grafičko, tablično i numeričko prikazivanje podataka koriste se za jednostavni prikaz informacija koje su dobivene iz uzorka.

U nastavku će biti opisani načini, odnosno tehnike prikupljanja podataka, organizacije podataka, prikazivanja rezultata, te načini korištenja mjera centralne tendencije i mjera disperzije.

1. Mjere centralne tendencije

Računanje centralne tendencije jedna je od najčešće primjenjivanih metoda u statistici čiji je cilj okupljanje podataka oko jedne središnje vrijednosti koja predstavlja skupinu rezultata u slučaju kada rezultati imaju tendenciju grupiranja oko neke vrijednosti. Koristi se kako bi se određeni skup podataka mogao što jasnije prikazati. Pri računanju centralne tendencije cijeli skup podataka mijenja jedna određena vrijednost za koju se smatra da taj skup vrlo precizno predstavlja. Zbog toga je potrebna velika pažnja prilikom odabira jedne od mjera centralne tendencije koja će predstavljati taj skup.

Mjere centralne tendencije (mjere prosjeka) su:

* aritmetička sredina,
* medijan (centralna vrijednost),
* mod (dominantna vrijednost),
* geometrijska sredina i
* harmonijska sredina.

Ove mjere predstavljaju prosječne, najčešće i središnje vrijednosti uzorka. U nekim slučajevima moguće je koristiti više mjera centralne tendencije za određene skupove, a njihova usporedba može biti vrlo informativna kada je u pitanju oblik distribucije podataka.

* 1. Aritmetička sredina

Aritmetička sredina je brojevna karakteristika koja je načešće korištena i smatra se najtočnijim pokazateljem prave vrijednosti mjerenja od svih mjera centralne tendencije.

Aritmetička sredina je jedina mjera centralne tendencije koja se može koristiti pri složenijim obradama podataka. Koristi se kao srednja vrijednost za homogene statističke skupove s niskom razinom varijabilnosti. Među mjerama centralne, aritmetička sredina najčešće je korištena u praksi, koristi se u skoro svakoj akademskoj oblasti. Da bi korištenje aritmetičke sredine bilo opravdano moraju postojati pravi mjerni podaci koji su točno određeni bar na intervalnoj skali, mora biti izmjeren dovoljan broj podataka, sva mjerenja moraju biti vršena u istim uvjetima, te distribucija podataka mora biti simetrična, odnosno normalna. Za označavanje aritmetičke sredine obično se koristi oznaka , iako se često kao oznaka koristi i veliko latinično slovo M.

Za ne grupirane podatke računa se tako što se zbroj svih vrijednosti elemenata promatranog uzorka podijele sa ukupnim brojem tih elemenata. Ukoliko ukupan broj elemenata označimo malim latiničnim slovom n, te zbroj svih vrijednosti elemenata skupa označimo oznakom , aritmetička sredina računa se preko matematičke formule (1).

(1)

Na primjer, ukoliko je zadan skup brojeva A={1, 4, 4, 22, 8, 9} i želimo izračunati njegovu aritmetičku sredinu, to bismo uradili na sljedeći način: . Dakle, aritmetička sredina promatranog skupa je 8.

Odstupanje vrijednosti ispod aritmetčke sredine uvijek je jednako odstupanju vrijednosti iznad aritmetičke sredine, odnosno zbroj odstupanja svih rezultata od aritmetičke sredine uvijek je jednak nuli. Zbog toga se aritmetička sredina naziva još i težištem rezultata.

Budući da na vrijednost aritmetičke sredine utječu svi elementi određenog skupa, vrijednost aritmetičke sredine značajno se mijenja ukoliko postoje elementi s ekstremno velikim ili ekstremno malim vrijednostima.

* 1. Medijan

Medijan, odnosno centralna vrijednost, kao i aritmetička sredina, spada u mjere centralne tendencije. Opisan je činjenicom da je barem pola podataka veće ili jednake vrijednosti kao i medijan, te da je istovremeno barem pola podataka manje ili jednake vrijednosti kao i medijan.

Za označavanje medijana obično se koristi veliko latinično slovo C, ali u nekim slučajevima koristi se i oznaka . Vrijednost medijana se računa ovisno o tome da li je broj podataka paran ili neparan.

Ukoliko je broj podataka u promatranom uređenom skupu neparan, tada postoji podatak čija je vrijednost na srednjoj poziciji u uređenom skupu podataka i ta vrijednost predstavlja medijan. Ako ukupan broj podataka promatranog skupa označimo malim latiničnim slovon n, pri pronalaženju srednje pozicije u skupu podataka za koju se obično kao oznaka koristi malo latinično slovo r, odnosno pozicije medijana, koristi se matematička formula (2), što znači da je medijan zapravo: =, gdje predstavlja vrijednost elementa koji se nalazi na srednjoj poziciji.

(2)

Na primjer, ukoliko je zadan skup brojeva B={4, 1, 5, 12, 67, 45, 67} i želimo odrediti njegov medijan, najprije trebamo poredati članove skupa po veličini s obzirom na njihove vrijednosti kako bi dobili uređeni skup, od najvećeg ka najmanjem ili od najmanjeg ka najvećem, sasvim je svejedno. Ukoliko za ovaj primjer poredamo elemente skupa krenuvši sa najmanjim, dobit ćemo skup B=(1, 4, 5, 12, 45, 67, 67). Zatim je potrebno pronaći srednju poziciju tog promatranog skupa, a to možemo učiniti na sljedeći način: . Dakle, srednja pozicija skupa, odnosno pozicija medijana je 4. U zadanom skupu na toj poziciji nalazi se broj 12, što znači da je to vrijednost medijana.

No, ukoliko je broj podataka u promatranom uređenom skupu paran, tada ne postoji jedinstveni podatak čija je vrijednost na srednjoj poziciji u uređenom skupu podataka jer se na srednjoj poziciji nalaze dva podatka. U tom slučaju medijan se definira kao aritmetička sredina tih dviju vrijednosti, odnosno broj koji se nalazi na polovini intervala između ovih dviju vrijednosti. Ukoliko te dvije vrijednosti označimo sa i , za računanje medijana u ovom slučaju koristili bi matematičku formulu (3).

.) (3)

Na primjer, ukoliko je zadan skup brojeva Y={18, 7, 3, 24, 66, 55} i želimo odrediti njegov medijan, najprije trebamo poredati članove skupa po veličini s obzirom na njihove vrijednosti kako bi dobili uređeni skup (kao u prethodnom primjeru). Ukoliko za ovaj primjer poredamo elemente skupa od najmanjeg ka najvećem, dobit ćemo urešđeni skup Y=(3, 7, 18, 24, 55, 66). Zatim je potrebno pronaći dvije srednje pozicije promatranog skupa. Prvu srednju poziciju možemo pronaći na sljedeći način: , a drugu srednju poziciju na sljedeći način: . Sljedeći korak je računanje aritmetičke sredine dviju vrijednosti koje se nalaze na tim pozicijama. U ovom slučaju to su brojevi 18 i 24, a njihova aritmetička sredina, koja je ujedno predstavlja i vrijednost medijana je: .

Za razliku od aritmetičke sredine, ekstremne vrijednosti ne utječu na vrijednost medijana, odnosno bitan je ukupan broj rezultata, a ne njihove vrijednosti. Zbog toga se centralna vrijednost većinom računa u slučajevima kada se u određenom skupu podataka nalaze ekstremne vrijednosti.

* 1. Mod

Mod, odnosno dominantna vrijednost, je još jedna mjera centralne tendencije. Obično se označava velikim latiničnim slovom D. Kao što samo ime kaže to je vrijednost koja dominira s obzirom na učestalost pojavljivanja u određenom skupu. Definiran je kao rezultat s najvećom frekvencijom, odnosno vrijednost koja se najčešće pojavljuje u nekom skupu rezultata. Kao kod centralne vrijednosti, bitan je ukupan broj rezultata, dok pojedinačne vrijednosti rezultata iz određenog skupa ne utječu na dominantnu vrijednost. Zbog toga se najčešće koristi u slučajevima kada se u skupu nalazi veliki broj rezultata među kojim se mogu pojaviti ekstremne vrijednosti, te postoji jedna vrijednost koja dominira tim skupom. Za dominantnu vrijednost se smatra da je najslabija među mjerama centralne tendencije, no usprkos tome može biti korisna u određenim slučajevima.

Na primjer, ukoliko je zadan skup brojeva T={2, 4, 4, 1, 7, 6, 7, 7, 3, 9}, tada broj 7 predstavlja dominantnu vrijednost jer se najčešće (3 puta) pojavljuje u promatranom skupu, tj. D=7. Za razliku od aritmetičke sredine i medijana, mod je vrijednost koja se uvijek pojavljuje među elementima skupa (osim u slučaju kada dva rezultata imaju jednaku frekvenciju koja je ujedno i najveća, tada se dominantna vrijednost ne mora nužno pojavljivati među elementima skupa).

Ukoliko se dva rezultata pojavlju jednak broj puta, odnosno ukoliko dva rezultata imaju istu frekvenciju koja je ujedno i najveća frekvencija, tada dominantnu vrijednost predstavlja aritmetička sredina tih dviju vrijednosti.

Na primjer, ako je zadan skup brojeva Z={1, 5, 12, 9, 76, 23, 11, 15, 11, 23}, tada dominantnu vrijednost predstavlja aritmetička sredina dvaju rezultata s najvećim frekvencijama, u ovom slučaju to su rezultati 23 i 11 jer se oba ova rezultata pojavljuju dva puta(svi ostali se pojavljuju samo po jednom). Dakle, dominantna vrijednost promatranog skupa je: .

U slučaju kada radimo s nezbrojivim vrijednostima, odnosno vrijednostima na nominalnoj mjernoj skali, tada za takvu distribuciju kažemo da je bimodalna.

Na primjer, ukoliko želimo provjeriti frekvencije doktora znanosti u različitim državama, te se ispostavi da u Francuskoj i Njemačkoj najviše doktora znanosti, tada nema smisla računati aritmetičku sredinu broja doktora znanosti u Francuskoj i broja doktora znanosti u Njemačkoj pa kažemo da postoje dvije dominantne vrijednosti.

U slučaju da se više od dva rezultata pojavljuju jednak broj puta, odnosno da više od dva rezultata imaju istu frekvenciju koja je ujedno i najveća, tada kažemo da dominantna vrijednost ne postoji jer nijedan rezultat ne dominira promatranim skupom. Isto vrijedi i za skupove kod kojih su svi elementi različitih vrijednosti.

Na primjer, neka su zadani skupovi brojeva P={2, 4, 14, 14, 2, 2, 15, 15, 15, 14} i R={1, 14, 5, 9, 11, 7, 65, 68, 45} za oba ova skupa kažemo da nemaju dominantnu vrijednost. Skup P sadrži 4 različita rezultata od kojih se 3 najveća rezultata pojavljuju jednak broj puta (3 puta), dok su svi rezultati koji su sadržani u skupu R frekvencije jedan (pojavljuju se samo po jednom), što znači da ni jedan od ova dva skupa ne sadrži vrijednost koja dominira tim skupom.

* 1. Geometrijska sredina

Geometrijska sredina, poznata i pod nazivom logaritamska sredina, jedna je od mjera centralne tendencije. Geometrijska sredina se kao mjera centralne tendencije najčešće koristi kao mjera prosječne brzine nekih promjena, odnosno u slučajevima kada podaci slijede geometrijsku progresiju, tj. pri analizi vremenskih nizova podataka. Primjerice, u ekonomiji pri analizi podataka. Dozvoljeno je računati je u slučajevima kada broj nije negativan ili jednak nuli. Najčešća oznaka koja se koristi za označavanje geometrijske sredine je .

Ako ukupan broj svih elemenata nekog skupa podataka označimo malim latiničnim slovon n, te umnožak svih elemenata skupa podataka označimo sa , tada geometrijsku sredinu možemo definirati kao n-ti korijen umnoška svih elemenata promatranog skupa. Za računanje geometrijske sredine možemo koristiti matematičku formulu (4).

(4)

Na primjer, ukoliko je zadan skup podataka Q={5, 1, 8, 13, 12, 15} i potrebno je izračunati njegovu geometrijsku sredinu, to ćemo učiniti na sljedeći način: =1.43005377999, odnosno ako rezultat zaokružimo na dvije decimale, geometrijska sredina zadanog skupa je 1.43.

* 1. Harmonijska sredina

Harmonijska sredina je još jedna mjera centralne tendencije. Najmanja je od poznatih brojevnih sredina. Obično se koristi pri računanju prosjeka nekih odnosa, odnosno pri računanju prosječnog vremena koje je potrebno da bi se obavila neka aktivnost. U usporedbi s aritmetičkom i geometrijskom sredinom koristi se vrlo rijetko. Kao geometrijsku sredinu, dozvoljeno je računati je u slučajevima kada broj nije negativan ili jednak nuli. Mjerna jedinica harmonijske sredine određena je varijablom za koju se računa harmonijska sredina, odnosno jednaka je mjernoj jedinici te varijable. Za označavanje harmonijske sredine obično se koristi veliko latinično slovo H.

Harmonijska sredina je kao recipročna vrijednost aritmetičke sredine od recipročnih vrijednosti podataka određenog skupa. Ukoliko ukupan broj podataka određenog skupa označimo malim latiničnim slovom n, a zbroj recipročnih vrijednosti svih podataka iz tog skupa označimo sa za računanje harmonijske možemo koristiti matematičku formulu (5).

(5)

Na primjer, neka je zadan skup podataka R={3, 17, 13, 14, 5, 32}. Ukoliko želimo izračunati njegovu harmonijsku sredinu, prvi korak bio bi da izračunamo recipročne vrijednosti svih podataka iz zadanog skupa, nakon čega je potrebno izračunati aritmetičku sredinu od svih tih recipročnih vrijednosti, te je zatim potrebno izračunati recipročnu vrijednost od te aritmetičke sredine kako bi dobili vrijednost harmonijske sredine zadanog skupa, što možemo učiniti na sljedeći način: .

1. Organiziranje podataka i prikazivanje rezultata

Prvi korak pri svakom procesu istraživanja je prikupljanje podataka iz kojih se mogu dobiti informacije koje su potrebne. Podatke je moguće prikupiti na više načina, primjerice iz nekih javnih izvora kao što su časopisi, knjige, novine, internet i slično ili iz nekih dizajniranih eksperimenata kao što su ankete koje istraživač vrši nad eksperimentalnim jedinkama koje su raspoređene u različite skupine s obzirom na njihove osobine, te nakon izvršavanja eksperimenta bilježi potrebne informacije.  
 Na primjer, ukoliko neka farmaceutska kompanija vrši testiranja kako bi se provjerio utjecaj određenog proizvoda. To se obično radi na način da se najprije jedinke nad kojima se vrše testiranja podijele u dvije skupine. Jednoj skupini se daje farmaceutski proizvod koji se testira, dok se drugoj skupini daje neki proizvod koji izgleda isto kao i proizvod koji se testira, ali nema nikakvog značajnog utjecaja na jedinku (placebo efekat). Zatim se potrebne informacije, odnosno traženi podaci dobivaju na osnovu odgovora na pitanja iz ankete koju istraživač provodi nad eksperimentalnim jedinkama, te na osnovu promatranja koje istraživač vrši nad eksperimentalnim jedinkama u njihovom prirodnom okruženju.

Nakon što su potrebni podaci prikupljeni nad njima je potrebno izvršiti aktivnosti organizacije, prikazivanja i statističkog obrađivanja, što nam omogućavaju metode deskriptivne statistike. Proces organiziranja podataka zasniva se na kvantificiranju svake varijable, odnosno kodiranju, te njihovim unošenjem u podatkovnu tablicu.

Kod varijabli koje su izmjerene na metričkoj skali proces kodiranja je vrlo jednostavan jer su podaci već u brojevnoj formi pa je potrebno samo unijeti podatke u podatkovnu tablicu.

Na primjer, ukoliko se na kraju školske godine vrši istraživanje o uspješnosti učenika s obzirom na njihove zaključne ocjene, u tom slučaju rezultate učenika, odnosno njihove zaključne ocjene već imamo u brojčanoj formi (od 1 do 5) pa ih je potrebno samo unijeti u podatkovnu tablicu.

Za razliku od procesa kodiranja varijabli koje su izmjerene na nominalnoj skali, proces organiziranja varijabli koje nisu već kvatificirane je složeniji jer podaci nisu u brojevnoj formi pa je potrebno opisne vrijednosti koje su sadržane u varijablama pretvoriti u brojeve.

Na primjer, ukoliko se vrši istraživanje prilikom kojeg ispitanici dolaze iz nekoliko različitih država, što je bitno za istraživanje. Neka je primjerice jedan dio ispitanika iz Hrvatske, jedan dio iz Bosne i Hercegovine, a jedan dio iz Srbije. U ovom slučaju varijablu države iz koje ispitanik dolazi ne možemo koristiti pri statističkoj obradi ukoliko se opisna vrijednost koja je sadržana unutar varijable nekako ne prebaci u brojčanu formu. To se može učiniti na više načina, primjerice da kodu „Hrvatska“ dodjelimo broj 1, kodu „Bosna i Hercegovina“ dodijelimo broj 2, a kodu „Srbija“ dodijelimo broj 3. Tek sada se ovi podaci mogu unijeti u podatkovnu tablicu.

Prilikom procesa kodiranja često se događa da neki od ispitanika ne daju odgovore ili da dati odgovor nije čitljiv. U takvim situacijama potrebno je odlučiti kako će se ti podaci kodirati i koristiti pri statističkoj analizi. To se može učiniti na više načina, no obično se dodaje novi kod, odnosno brojčana vrijednost koja će označavati takve vrijednosti i one se neće miješati s ostalim vrijednostima koje su sadržane unutar te varijable.

Podaci se u podatkovnu tablicu unose tako što se u tablične redove smjeste pojedini ispitanici, a u tablične stupce unose se varijable koje su izmjerene prilikom istraživanja.

* 1. Grupiranje podataka

Često se događa da podaci unutar određenih varijabli nakon unosa u podatkovnu tablicu budu veoma raspršeni, zbog čega ih je veoma teško prikazati na smislen način. Kako bi se njihov prikaz olakšao, obično se koristi metoda grupiranja podataka po razredima.

Nakon što je proces prikupljanja podataka završen, podatke je potrebno organiziriat na najbolji mogući način kako bi se mogli prikazati uz pomoć grafičkog ili tabličnog prikaza, te kako bismo mogli provjeriti način na koji su ti podaci raspodjeljeni i da li su pogodni za primjenu u određenim statističkim postupcima. Prvi korak pri procesu statističke obrade obično je grafičko prikazivanje podataka, što će nam olakšati grupiranje podataka. Ukoliko među prikupljenim podacima broj mogućih vrijednosti nije velik, tada je takvo grupiranje besmisleno raditi, no u suprotnom slučaju grupiranje podataka može biti veoma korisno.

Prvi korak pri procesu grupiranja rezultata obično je da se odredi ukupan broj razreda u koje će se rezultati grupirati, što se određuje proizvoljno s obzirom na karakteristike prikupljenih podataka.

Nakon što se odredi ukupan broj razreda, određuje se interval razreda. Da bi odredili interval razreda potrebno je znati koliki je totalni raspon, koji se računa tako što se od vrijednosti najvećeg rezultata oduzme vrijednost najmanjeg rezultata i dobivenu vrijednost, odnosno razliku tih dvaju rezultata uvećamo za jedan. Za računanje totalnog raspona možemo koristiti matematičku formulu (6), gdje je TR oznaka za totalni raspon, oznaka za vrijednost najvećeg rezultata, a oznaka za vrijednost najmanjeg rezultata.

)+1 (6)

Nakon što je poznato koliko je totalni raspon možemo izračunati interval razreda, koji se računa na način da se totalni raspon podijeli sa ukupnim brojem razreda. Ukoliko interval razreda označimo oznakom IR, te ukupan broj razreda oznakom BR, za računanje intervala razreda možemo koristiti matematičku formulu (7). Kada izračunamo interval razreda, dobivena vrijednost zaokružuje se na veću vrijednost kako bi se osiguralo da svaki rezultat bude pridružen razredu koji je za njega predviđen.

(7)

Sljedeći korak je određivanje gornje i donje granice za svaki od razreda. Granice razreda obično se određuju na način da svojom preciznošću odgovaraju izmjerenim podacima. Na primjer, ukoliko su rezultati koji su grupirani u određeni razred zapisani kao cijeli brojevi, tada će i granice biti cijeli brojevi.

Posljednji korak pri procesu grupiranja podataka je prikazivanje raspodjele rezultata, odnosno određivanje frekvencije za sve rezultate u svakom od razreda. Frekvencija nekog rezultata predstavlja broj pojavljivanja tog rezultata, odnosno njegovu učestalost. Pored određivanja frekvencije, postoji mogućnost računanja proporcije, odnosno relativne frekvencije, te postotka za svaki od rezultata. Proporcija rezultata se računa tako što se frekvencija podijeli s ukupnim brojem podataka, a postotak rezultata se računa tako što proporciju pomnožimo sa 100.

Postoji nekolko smjernica kojih se potrebno pridržavati pri procesu grupiranja podataka. Te smjernice su:

* Intervali, odnosno rasponi svih razreda moraju biti jednake veličine.
* Poželjno je da ukupan broj razreda bude neparan.
* Razredi se ne smiju preklapati, odnosno svaki podatak mora biti sadržan u točno jednom razredu (samo u onom razredu čijem intervalu odrešeni podatak pripada).
* Raspodjele s praznim razredima potrebno je izbjegavati.
* Pri odabiru ukupnog broja razreda poželjno je da se pokuša raspodjela na što više mogućih načina kako bi se mogao odabrati broj razreda koji najbolje odgovara raspodjeli.

Na primjer, neka je skupina od 20 studenata pristupila ispitu iz kolegija „Uvod i vjerojatnost i statistiku“, te neka skup S={23, 54, 11, 94, 97, 4, 33, 11, 76, 55, 55, 94, 12, 32, 33, 75, 88, 90, 2, 15} predstavlja skup svih rezultata (bodova) koje su studenti ostvarili na ispitu. Ukoliko želimo grupirati rezultate u razrede, prvi korak bio bio da odredimo broj razreda. U ovom slučaju za ukupan broj razreda uzet ćemo broj BR=5 jer imamo veoma mali broj rezultati, a uz to želimo i da ukupan broj razreda bude neparan. Sljedeći korak je računanje intervala razreda, što je u biti razlika između totalnog raspona i ukupnog broja razreda. U ovom slučaju totalni raspon jednak je TR=(97-2)+1=96 pa će stoga interval razreda biti jednak =19.2. Tu vrijednost potrebno je zaokružiti na IR=20. Zatim je potrebno odrediti donju i gornju granicu za svaki od razreda. Za donju granicu prvog razreda odabrat ćemo broj 1 zbog zaokruživanja koje je prethodilo, iako je broj jedan manji od vrijednosti najmanjeg rezultata. Budući da nam je interval razreda IR=20, za gornju granicu prvog razreda uzet ćemo broj 21 kako bi omogućili da se u taj razred može staviti 20 različitih rezultata, odnosno sve rezultate s vrijednostima između 1 i 20 (uključujuči i rezultate s vrijednostima 1 i 20). Na isti način određuju se donje i gornje granice za ostale razrede.Na kraju, nakon što su određene donje i gornje granice svih razreda, potrebno je rezultate smjestiti u razrede kojima pripadaju. U tu svrhu koristit ćemo se tablicom u koju unosimo jedan po jedan rezultat, te ih tako i označavamo. Kao primjer pomoćne tablice s podacima grupiranim u razrede pogledajte tablicu ispod (*Tablica br. 1*) u kojoj su prikazani rezultati iz ovog primjera grupirani u razrede.

Tablica br. Pomoćna tablica rezultata iz prethodnog primjera grupiranih u razrede

|  |  |  |
| --- | --- | --- |
| **Razred** | **Granice razreda** | **Broj rezultata** |
| 1. | 1-20 | 6 |
| 2. | 21-40 | 4 |
| 3. | 41-60 | 3 |
| 4. | 61-80 | 2 |
| 5. | 81-100 | 5 |

* 1. Tabličnini prikaz podataka

Tablični prikaz podataka jedna je od metoda za prikazivanje podataka koji su prikupljeni pri procesu istraživanja. Koristi se u slučajevima kada su izvorne vrijednosti mjerenja bitnije od organiziranosti skupa podataka, u suprotnim slučajevima koristi se grafički prikaz podataka.

Postoje smjernice kojih se potrebno pridržavati prilikom tabličnog prikazivanja podataka. Te smjernice su:

* Tablica mora biti jasnog sadržaja, mora biti opisana brojem i kratkim naslovom koji nam govori što ta tablica predstavlja, te ukoliko je potrebno uz tablicu je moguće dodati pojašnjenja.
* Naslov tablice potrebno je pisati malo manjim fontom od fonta u kojem je pisan ostatak teksta, mora se nalaziti iznad tablice, te ga je potrebno centrirati.
* Redovi i stupci tablice moraju biti kratko i jasno označeni, vrijednosti koje se nalaze u tabličnim ćelijama moraju biti smisleno poredane, te ukoliko je potrebno istaknuti neki podatak to se može učiniti korištenjem zvjezdice, te je potrebno značenje zvjezdice objasniti ispod tablice.
* U dijelu teksta za čije razumjevanje je potrebna tablica, potrebno je navesti redni broj tablice koju je potrebno pogledati.
* Vodoravne linije koriste se za odvajanje većih tabličnih cjelina kao što su zaglavlje i podnožje tablice, dok se uspravne linije trebaju izbjegavati.

Kao primjer tabličnog prikaza podataka pogledajte tablicu stanja nogometnih klubova u skupini C grupne faze europske Lige Prvaka nakon 3. takmičarskog kola (*Tablica br. 2*).

Tablica br. Tablica stanja nogometnih klubova u skupini C grupne faze europske Lige Prvaka

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Naziv Kluba | Broj odigranih utakmica | Broj postignutih pogodaka | Broj primljenih pogodaka | Gol razlika | Bodovi |
| Manchester City | 3 | 10 | 1 | 9 | 9 |
| Dinamo Zagreb | 3 | 6 | 4 | 2 | 4 |
| Shakhtar | 3 | 4 | 6 | -2 | 4 |
| Atalanta | 3 | 2 | 11 | -9 | 0 |

* 1. Grafički prikaz podataka

Pored tehnike tabličnog prikazivanja podataka, podatke možemo prikazati i uz pomoć različitih vrsta grafičkog prikazivanja podataka. Grafičko prikazivanje podataka se koristi kako bi se podaci dobiveni provođenjem nekog istraživanja slikovito prikazali u svrhu što jasnijeg i preciznijeg razumjevanja rezultata koji su dobiveni procesom istraživanjea.

Uz pomoć tehnika ekstrapolacije i interpolacije, grafički prikaz podataka može biti koristan i pri procjenjivanju određenih vrijednosti koje nisu prethodno određene nekom vrstom mjerenja. Ovaj način prikazivanja podataka je vrlo bitan za otkrivanje nekih posebnih osobina podataka koje nisu očekivane, a može biti koristan i pri uspoređivanju različitih rezultata.

Pri procesu grafičkog prikazivanja podataka potrebno je posebnu pažnju posvetiti čitljivosti i jasnoći prikazanih podataka. Kao i kod tabličnog prikazivanja podataka, postoje smjernice kojih se potrebno pridržavati pri grafičkom prikazivanju podataka. Te smjernice su:

* Svaki grafički prikaz mora biti opisana rednim brojem i kratkim naslovom

koji nam govori što taj grafički prikaz predstavlja.

* Ukoliko je potrebno, uz grafički prikaz moguće je dodati neka pojašnjenja u svrhu lakšeg razumjevanja.
* Za razliku od tabličnog prikazivanja, kod grafičkog prikazivanja podataka naslov i redni broj pišu se ispod grafičkog prikaza, te ih je potrebno pisati malo manjim fontom od fonta u kojem je pisan ostatak teksta.
* U većini slučajeva uz grafički prikaz je potrebno dodati „Legendu“ u kojoj su sadržana objašnjenja potrebna za jasno razumjevanje tog grafičkog prikaza.
* Grafički prikaz mora biti čitljiv i u potpunosti razumljiv.
* U dijelu teksta za čije razumjevanje je potreban grafički prikaz, potrebno je navesti redni broj grafičkog prikaza koji je potrebno pogledati.

Postoji više različitih tehnika za grafičko prikazivanje podataka između kojih se odlučujemo ovisno o vrsti podataka koji su prikupljeni provođenjem aktivnosti istraživanja. Načini koji se najčešće koriste su:

* kružni dijagram (torta-dijagram),
* stupčani dijagram,
* histogram i
* poligon frekvencija

Pri korištenju stupčanih dijagrama i poligona frekvencija potrebno je pridržavati se sljedećih smjernica:

* Dužina ordinate treba biti veća od dužine apscise.
* Jedinice na apscisi i ordinati moraju biti jasno označene.
* U većini slučajeva potrebno je dodati legendu koja pomaže pri razumjevanju dijagrama.
* Na ordinati je obično predstavljena frekvencija, ali u određenim slučajevima moguće je koristiti postotke.
  + 1. Kružni dijagram

Kružni diagram, poznat i pod nazivom torta-dijagram, jedan je od načina grafičkog prikazivanja podataka. Koristi se obično pri procesu prikazivanja distribucije podataka u nekoj određenoj varijabli. Kod ovog dijagrama je bitno da su prikazane sve potrebne kategorije za određenu varijablu i postotak pripadajućih podataka za svaku od tih kategorija. Pri korištenju kružnog dijagrama cilj je predstaviti neke od glavnih informacije do kojih smo došli istraživanjem. Ovaj način grafičkog prikazivanja podataka preporučljivo je koristiti u slučajevima kada se radi o distribuciji varijabli koje nemaju veliki broj kategorija. U većini slučajeva potrebno je uz dijagram dodati i legendu koja pojašnjava značenje određenih stavki u tom dijagramu kako bi on bio što jasniji i lakši za razumijevanje.

Na primjer, neka je na kraju školske godine u nekoj školi 15% učenika imalo odličan uspjeh, 25% učenika vrlodobar uspjeh, 34% učenika dobar uspjeh, 14% učenika dovoljan uspjeh, te je 12% učenika palo, odnosno imalo nedovoljan uspjeh. Kao primjer grafičkog prikazivanja podataka uz pomoć kružnog dijagrama pogledajte sliku ispod (*Slika 1.*) koja prikazuje uspješnost učenika iz ovog primjera na kraju školske godine.

****

Slika 1. Uspješnost učenika iz prethodnog primjera na kraju školske godine

* + 1. Stupčani dijagram

Stupčani dijagram, kao što samo ime kaže, građen je u obliku stupaca. Kao i kružni dijagram, stupčani dijagram koristi se pri grafičkom prikazivanju podataka. Obično se koriste za prikazivanje odnosa između određenih varijabli i njihovih frekvencija. Sastoje se od više stupaca, a visina svakog stupca nam govori o frekvenciji određene kategorije. Na apscisi su predstavljene vrijednosti kategorija, dok su na ordinati obično predstavljene frekvencije.

Poseban oblik stupčanog dijagrama predstavlja histogram. To je stupčani dijagram čije su varijable kontinuirane. Histogrami se sastoje od niza stupaca čije visine predstavljaju frekvencije intervala, a dodavanjem točke na sredinu svakog od tih stupaca lako ga je pretvoriti u poligon frekvencija.

Na primjer, ukoliko želimo grafički prikazati broj novorođenih dječaka i djevojčica u prethodnih 5 godina u nekom gradu. Recimo da je u nekom gradu u 2014. godini rođeno 30 dječaka i 44 djevojčice, u 2015. godini 39 dječaka i 38 djevojčica, u 2016. godini 35 dječaka i 41 djevojčica, u 2016. godini 32 dječaka i 31 djevojčica, u 2017. godini 29 dječaka i 34 djevojčice, te u 2018. godini 26 dječaka i 26 djevojčica. Kao primjer grafičkog prikazivanja uz pomoć stupčanog dijagrama, u ovom slučaju histograma, pogledajte histogram (*Slika 2.*) koji grafički prikazuje broj novorođenih dječaka i djevojčica iz ovog primjera za svaku od prethodnih 5 godina.



Slika 2. Broj novorođenih dječaka i djevojčica iz prethodnog primjera za svaku od prethodnih 5 godina

* + 1. Poligon frekvencija

Poligon frekvencija je još jedan od često korištenih načina za grafičko prikazivanje podataka. Koristi se pri prikazivanju odnosa između određenih varijabli i njihovih frekvencija. Sastoji se od linija koje su opisane točkama. Visine tih točaka nam daju informacije o frekvencijama intervala.

Na primjer, recimo da uz pomoć poligona frekvencija želimo grafički prikazati usporedbu neke dvojice nogometaša ovisno o broju postignutih pogodaka u prethodnih 5 sezona. Recimo da je određeni nogometaš (Nogometaš broj 1) u sezoni 2014./2015. postigao 17 pogodaka, u 2015./2016. sezoni 13 pogodaka, u 2016./2017. sezoni 25 pogodaka, u 2017./2018. sezoni 20 pogodaka, te u 2018./2019. sezoni 29 pogodaka. A drugi nogometaš (Nogometaš broj 2) u sezoni 2014./2015. postigao 19 pogodaka, u 2015./2016. sezoni 15 pogodaka, u 2016./2017. sezoni 19 pogodaka, u 2017./2018. sezoni 20 pogodaka, te u 2018./2019. sezoni 27 pogodaka. Kao primjer grafičkog prikaza podataka uz pomoć poligona frekvencija pogledajte sliku ispod (*Slika 3.*) na kojoj je prikazano koliko su nogometaši iz primjera postigli pogodaka u svakoj od prethodnih 5 sezona.



Slika 3. Broj postignutih pogodaka koje su ostvarili nogometaši iz prethodnog primjera u prethodnih 5 sezona

1. Mjere disperzije

Mjere disperzije, odnosno mjere raspršenja, računaju se s ciljem dobivanja informacija o međusobnom razlikovanju rezultata unutar nekog određenog skupa. Te informacije moraju biti dostupne prilikom procesa istraživanja.

Postoji više mjera disperzije, među kojima se najčešće koriste:

* raspon rezultata,
* varijanca,
* standardna devijacija i
* koeficijent varijabilnosti.

Pored već nabrojanih, najčešće korištenih, postoje još neke mjere disperzije koje se mnogo rijeđe koriste. Neke od takvih mjera disperzije su primjerice indeks srednjeg odstupanja i poluinterkvartilno raspršenje (interkvartilni raspon) koje se može računati uz centralnu vrijednost (medijan) nad nekim uređenim skupovima podataka, a opisano je kao razlika među rezultatima koji se nalaze na granicama trećeg (gornjeg) i prvog (donjeg) kvartila.

* 1. Raspon rezultata

Raspon rezultata je jedna od mjera disperzije, a definiran je kao razlika između najvećeg i najmanjeg rezultata u nekom uređenom skupu (u skupu u kojem su rezultati poredani po veličini). Ova mjera disperzije je jednostavna i lako razumljiva, no budući da se vrijednost raspona rezultata temelji na samo dvije vrijednosti, najvećoj i najmanjoj vrijednosti u skupu, raspon rezultata je vrlo osjetljiv na skupove s ekstremnim vrijednostima. Pored toga što na raspon rezultata mogu negativno utjecati ekstremne vrijednosti, ova mjera disperzije obično raste zajedno s porastom broja rezultata unutar skupa pa se zbog toga smatra vrlo nesigurnom.

Ukoliko raspon rezultata označimo velikim latiničnim slovom R, najmanju vrijednost unutar određenog skupa označimo sa , te najveću vrijednost unutar tog istog skupa označimo sa , za računanje vrijednosti raspona rezultata možemo koristiti matematičku formulu (8).

R= (8)

Na primjer, neka je zadan uređeni skup rezultata S=(4, 4, 7, 7, 7, 8, 9, 11, 11, 13, 13, 13), raspon rezultata za zadani skup možemo izračunati na sljedeći način: R=13-4=9. Dakle, raspon rezultata zadanog skupa je 9.

* 1. Varijanca

Varijanca je mjera disperzije koja je opisana kao središnja vrijednost zbroja kvadriranih vrijednosti odstupanja svakog pojedinog rezultata od aritmetičke sredine, a dozvoljeno je računati je samo uz aritmetičku sredinu. Za označavanje varijance obično se koristi oznaka ili veliko latinično slovo V.

Računanje varijance je vrlo korisno pri procesu provođenja kompliciranijih statističkih analiza, ali kao samostalna vrijednost se koristi vrlo rijetko. Prvi korak pri procesu računanja varijance je računanje razlike među svim pojedinim rezultatima i aritmetičke sredine, nakon toga te razlike se kvadriraju i zbrajaju, te se naposljetku taj zbroj podijeli s ukupnim brojem rezultata. Ukoliko sa označimo sve pojedine rezultate mjerenja, sa označimo aritmetičku sredinu, te malim latiničnim slovom n označimo ukupan broj rezultata, varijancu možemo računati korištenjem matematičke formule (9) koja se obično koristi kada se radi s podacima dobivenim iz ukupne populacije.

(9)

Prilikom računanja varijance u slučajevima kada se radi s podacima koji su dobiveni procesom mjerenja na uzorku, točnije je koristiti modificiranu matematičku formulu (10).

(10)

Na primjer, ukoliko smo mjerenjem nad nekom vrlo malom populacijom dobili rezultate koji su sadržani u uređenom skupu S=(7, 7, 9, 9, 9, 11, 11), najprije moramo izračunati aritmetičku sredinu rezultata, a to ćemo učiniti na sljedeći način: . Nakon što smo izračunali aritmetičku sredinu slijedi računanje varijance, što možemo učiniti na sljedeći način: : .

* 1. standardna devijacija

Standardna devijacija je mjera disperzije koja je definirana kao drugi korijen iz središnje vrijednosti zbroja kvadriranih vrijednosti odstupanja svakog pojedinog rezultata od aritmetičke sredine, odnosno drugi korijen iz vrijednosti varijance. Kao i varijancu, standardnu devijaciju dozvoljeno je računati samo uz aritmetičku sredinu. Točnost kojom aritmetička sredina predstavlja neki skup retultata ovisi o vrijednosti standardne devijacije. Što je vrijednost standardne devijacije veća, aritmetička sredina nepreciznije predstavlja skup rezultata, odnosno što je vrijednost standardne devijacije manja, to aritmetička sredina preciznije prestavlja skup rezultata. Za označavanje standardne devijacije obično se koristi oznaka SD.

Vrijednost standardne devijacije dobije se na sličan način kao i vrijednost varijance. Postoje dvije matematičke formule koje se koriste pri računanju standarde devijacije ovisno o načinu na koji su dobiveni podaci s kojima se radi. Razlika je u tome što se jedna od ovih formula koristi kada se radi s podacima koji su dobiveni iz ukupne populacije, dok se druga koristi kada su podaci dobiveni procesom mjerenja uzoraka. Prvi korak pri procesu računanja standardne devijacije je računanje razlika među svim pojedinim rezultatima i aritmetičke sredine, nakon čega se te razlike kvadriraju i zbrajaju, a zatim se taj zbroj podijeli s ukupnim brojem rezultata. Posljednji korak je računanje drugog korijena iz vrijednosti koju smo dobili tim dijeljenjem. Ukoliko sa označimo sve pojedine rezultate mjerenja, sa označimo aritmetičku sredinu, te malim latiničnim slovom n označimo ukupan broj rezultata, standardnu devijaciju možemo računati korištenjem matematičke formule (11) koja se obično koristi kada se radi s podacima dobivenim iz ukupne populacije.

(11)

Pri procesu računanja standardne devijacije u slučajevima kada se radi s podacima koji su dobiveni procesom mjerenja na uzorku, točnije je koristiti modificiranu matematičku formulu (12) koja se mnogo češće koristi u praksi jer su slučajevi u kojima je moguće izvršiti mjerenje nad svim jedinkama neke određene populacije vrlo rijetki pa se podaci obično dobivaju provođenjem mjerenja na uzorku.

S (12)

Na primjer, ukoliko smo mjerenjem nad nekom vrlo malom populacijom dobili rezultate koji su sadržani u uređenom skupu S=(8, 8, 8, 10, 10, 10, 13, 13), najprije moramo izračunati aritmetičku sredinu rezultata, a to ćemo učiniti na sljedeći način: . Nakon što smo izračunali aritmetičku sredinu slijedi računanje varijance, što možemo učiniti na sljedeći način: : . Posljednji korak pri računanju standarde devijacije je računanje drugog korijena iz varijance, to ćemo učiniti na sljedeći način: =2.07.

* 1. Koeficijent varijabilnosti

Koeficijent varijabilnosti je mjera disperzije koja se računa u slučajevima kada su nam potrebne informacije o razlikama u variranju u nekim određenim svojstvima između dvije ili više skupina podataka. Za označavanje koeficijenta varijabilnosti obično se koristi oznaka CV.

Koeficijent varijabilnosti se računa na način da najprije izračunamo aritmetičku sredinu i standardnu devijaciju, a nakon toga te vrijednosti uvrstimo u matematičku formulu (13) za računanje koeficijenta varijabilnosti, gdje je SD oznaka za standardnu devijaciju, a oznaka za aritmetičku sredinu.

(13)

Na primjer, ukoliko su zadana dva skupa rezultata P=(2, 2, 5, 5, 5, 9, 10, 10) i R=(12, 12, 15, 16, 20, 20, 20, 20), najprije računamo aritmetičku sredinu i standardnu devijaciju za skup rezultata P. To ćemo učiniti na sljedeći način: i =3.3. Nakon što smo izračunali aritmetičku sredinu i standardnu devijaciju, na sljedeći način ćemo izračunati koeficijent varijabiliteta za skup rezultata P: Isti postupak ponovimo i za skup rezultata R. Najprije računamo aritmetičku sredinu na sljedeći način: zatim računamostandardnu devijaciju na sljedeći način:

=13, te naposljetku računamo koeficijent varijabiliteta na sljedeći način: Sada možemo usporediti koeficijente varijabiliteta skupova P i R, iz čega možemo zaključiti da je varijabilitet veći u skupu R.

1. Zaključak

Deskriptivna statistika je dio statistike koji podrazumjeva tehnike i metode potrebne za rad s prikupljenim podacima u svakodnevnom životu. Ovaj dio statistike pruža informacije potrebne za lakše razumjevanje i rješavanje većine problema s kojima se svakodnevno susrećemo. Obuhvaća moguće načine za organizaciju podataka i njihovo jasno i jednostavno prikazivanje. Proučavanjem i razumjevanjem metoda i tehnika deskriptivne statistike rad s prikupljenim podacima postaje mnogo brži i jednostavniji.

1. Literatura
2. Bubuić, A., *Osnove statistike u društvenim i obrazovnim znanostima*, Filozofski fakultet, Split, 2015.
3. Gogala, Z., *Osnove statistike*, Sinergija d.o.o., Zagreb, 2001.
4. Papić, M., *Primijenjena statistika u MS Excelu*, ZORO d.o.o., Zagreb, 2005.
5. Šošić, I., *Primijenjena statistika*, Školska knjiga, Zagreb, 2006.
6. <https://ldap.zvu.hr/~oliverap/VjezbeIzStatistike/3_Mjere%20centralne%20tendencije%20vje%C5%BEbe.pdf>, 9.9.2019.
7. <http://matematika.fkit.hr/novo/statistika_i_vjerojatnost/predavanja/1%20-%20Deskriptivna%20statistika.pdf>, 9.9.2019.
8. <http://neuron.mefst.hr/docs/katedre/neuroznanost/Osnove_Prikaz_podataka_2010.pdf>, 9.9.2019.
9. <https://www.pmf.unizg.hr/_download/repository/PREDAVANJE7.pdf>, 9.9.2019.
10. <https://www.ucg.ac.me/skladiste/blog_22018/objava_32492/fajlovi/Deskriptivna%20statistika.pdf>, 23.10.2019.