Agencija za
strukovno obrazovanje
i obrazovanje odraslih

OSNOVE STATISTIKE

Vladimira Majdandžić

MODERNIZACIJA SUSTAVA STRUKOVNOG OBRAZOVANJA I OSPOSOBLJAVANJA

Projekt je sufinancirala Europska unija iz Europskog socijalnog fonda.

Autor:
Vladimira Majdandžić

Sektor:
Elektrotehnika i računarstvo

Naziv standarda kvalifikacije:
Tehničar za programiranje / Tehničarka za programiranje

Razina HKO: 4.2

Modul:
Osnove umjetne inteligencije

Skup ishoda učenja:
Osnove statistike

Razina SIU: 4

Ishodi učenja:

1. Opisati populacije i uzorke, metode uzorkovanja i vrste uzoraka te numerički i grafički prikazati univarijantne i bivarijantne podatke
2. Prepoznati aberantne podatke, njihov utjecaj na statističko zaključivanje te opisati način postupanja s njima za zadani scenarij
3. Provesti normizaciju i standardizaciju podataka te interpretirati postupke za usklađivanje uzorka s populacijskim obilježjima
4. Interpretirati različite korelacijske pokazatelje i odabrati pokazatelje ovisno o podacima te opisati odnose analize i testiranja hipoteze za zadani problem

Ključni pojmovi:
populacija, uzorak, deskriptivna, interferencijalna statistika, statistički skup, populacija, modalitet, mod, relativna frekvencija, aritmetička sredina; donji kvartil; medijan; percentil; varijacija; devijacija

CSVET: 1

Datum izrade DOM-a:
19. srpnja 2023.

Napomena:

Riječi i pojmovni sklopovi koji imaju rodno značenje korišteni u ovom dokumentu (uključujući nazive strukovnih kvalifikacija, zvanja i zanimanja) odnose se jednako na oba roda (muški i ženski) i na oba broja (jedninu i množinu), bez obzira na to jesu li korišteni u muškom ili ženskom rodu, odnosno u jednini ili množini.

Agencija za
strukovno obrazovanje
i obrazovanje odraslih

OSNOVE STATISTIKE

Vladimira Majdandžić

MODERNIZACIJA SUSTAVA STRUKOVNOG OBRAZOVANJA I OSPOSOBLJAVANJA

Europska unija "Zajedno do fondova EU"

EUROPSKI STRUKTURNI I INVESTICIJSKI FONDOVI

Projekt je sufinancirala Europska unija iz Europskog socijalnog fonda.

SADRŽAJ

UVOD

Razni autori o statistici

„Postoje tri vrste laži: laž, besramna laž i statistika.“ (Mark Twain)

„Smrt jednog čovjeka je tragedija, smrt miliona je statistika.” (Josif Staljin)

„Statistika je bajka razuma.” (Martin Kessel)

”Statistika je skup točnih podataka koji daje pogrešan rezultat.” (Anonimus)

”Statistika naša dika: štogod hoćeš, ona slika.” (Vladimir Bulatović Vib)

”Ja jedem kupus, ti jedeš meso – u prosjeku jedemo sarmu.” (Anonimus)

Pojam statistike

Statistika je grana matematike koja obuhvaća sakupljanje, analizu, interpretaciju i prezentaciju podataka te izradu predviđanja koja se temelje na tim podacima.

Najstariji zapisi o korištenju statistike potječu iz 9. stoljeća u zapisima arapskog znanstvenika Al-Kindi-ja koji koristi statistiku u svrhu izučavanja kodiranih poruka, a u 14. stoljeću nastaju zapisi Nuova Cronica (povijest Firenze) koje sadrže niz statističkih podataka o populaciji, edukaciji i sl.

Statistiku kao znanstvenu disciplinu dijelimo na **opisnu** ili deskriptivnu i **interferencijalnu**.

Opisna (deskriptivna) statistika temelji se na konkretnim rezultatima dobivenima nekim istraživanjima ili mjerenjima. U tim se istraživanjima skup svih promatranih objekata obuhvaća u cijelosti. Zadaća opisne statistike je “opisati” dobivene rezultate, tj. srediti ih i sažeti tako da budu što pregledniji, razumljiviji i pogodniji za interpretaciju, daljnju analizu i primjenu.

Inferencijalna statistika temelji se na djelomičnu obuhvatu skupa svih promatranih objekata. Rezultati se dobivaju ispitivanjem određenog dijela (uzorka) promatranih objekata. Koristeći induktivnu metodu izvode se zaključci o karakteristikama (obilježjima) svih promatranih objekata.

Slika 1. Statistička analiza podataka

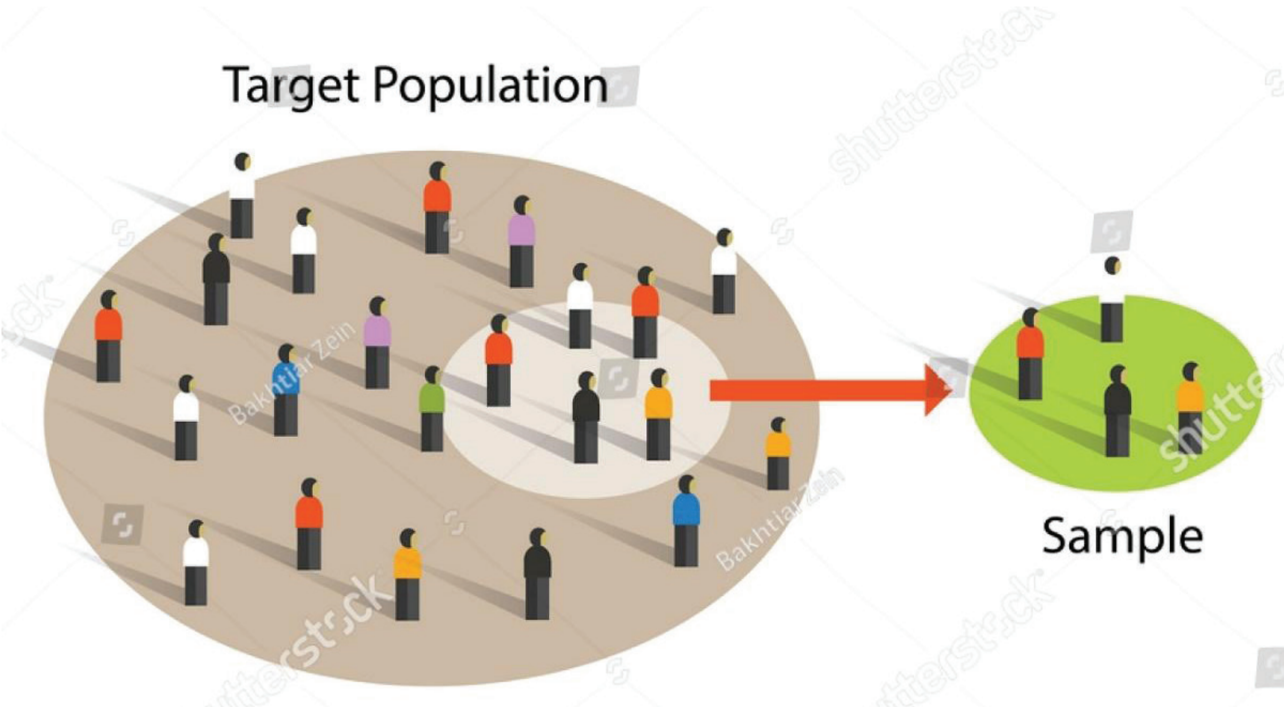
1. OSNOVE STATISTIKE

U ovom materijalu bavit ćete se pojmovima, prikazima i postupcima deskriptive statistike. Opisat ćete populacije i uzorke, metode uzorkovanja i vrste uzoraka. Prikazat ćete numerički i grafički univarijantne i bivarijantnte podatke. Bit ćete sposobni prepoznati aberantne podatke, njihov utjecaj na statističko istraživanje i opisati način postupanja s njima za zadani scenarij. Provest ćete normizaciju i standardizaciju podataka i interpretirati na primjerima postupke za usklađivanje uzorka s populacijskim obilježjima. Interpretirat ćete različite korelacijske pokazatelje i opisati odnose analize i testiranja hipoteze za zadani problem. Svi ovi pojmovi i postupci bit će u ovom materijalu potkrijepljeni primjerima za lakše razumijevanje i produbljivanje stečenog znanja. Preporučamo da svaki primjer riješite, najprije uz upute u materijalu, a zatim samostalno.

Da bismo mogli napraviti dublju analizu podatka, potrebno je krenuti od osnovnih pojmova u statistici.

Osnovni pojmovi

Statistički skup je osnovni pojam u statistici. Tvore ga **statističke jedinice** ili **elementi** (osobe, poslovni subjekti, regije, države, predmeti itd.) koji imaju barem jedno zajedničko svojstvo (**obilježje** ili **varijabla**) koje od elementa do elementa očituje statističku promjenjivost. Ukupan broj elemenata statističkoga skupa naziva se **opseg statističkoga skupa**. Prema **opsegu**, statističke skupove (kao i sve druge skupove) dijelimo na **konačne** i **beskonačne**, a prema **vrsti** elemenata na **stvarne** (realne) i **zamišljene** (hipotetične). **Stvarni** statistički skupovi sastoje se od elemenata koji postoje u tekućem vremenu i prostoru, a **zamišljeni** statistički skupovi sastoje se od elemenata generiranima različitim modelima procesa ili pokusa.



Slika 2. Uzorkovanje

Već je istaknuto da statistički skup tvore elementi koji imaju barem jedno zajedničko svojstvo. Stoga za svaki element statističkoga skupa znamo odgovarajući podatak o tome svojstvu. Skup svih takvih podataka naziva se **osnovni** skup ili **populacija**.

Iz definirane populacije uzima se u statističko istraživanje **uzorak**, a individualni član populacije čija svojstva mjerimo nazivamo **element**.

Populacija ili statistički skup mora biti definiran pojmovno, prostorno i vremenski.

Pojmovna definicija utvrđuje pripadnost bilo kojega uzorka navedenom skupu, **prostorna** definicija utvrđuje prostor kojemu pripadaju svi elementi populacije, a **vremenska** utvrđuje vremensko razdoblje ili vremenski trenutak za koji su vezani svi uzorci.

Primjer 1.

Promatra se statistički skup koji tvore svi učenici prvih razreda Srednje strukovne škole u Samoboru na dan 1. rujna 2023. Definirajmo navedeni skup pojmovno, prostorno i vremenski:

Rješenje:

pojmovna definicija: učenici prvih razreda Srednje strukovne škole

prostorna definicija: Srednja strukovna škola u Samoboru

vremenska definicija: 1. rujna 2023.

Skup je realan i konačan, a njegov opseg iznosi 145 učenika. Primjer osnovnog skupa vezanog za promatrani statistički skup jest skup svih zaključnih ocjena iz predmeta matematika. Jedan od uzoraka toga skupa je npr. Top 20, tj. skup 20 najboljih učenika iz predmeta matematika na kraju školske godine.

Osobitosti pojedinih populacija predočavaju se različitim analitičkim veličinama, pri čemu, ukoliko se radi o brojčanoj veličini, govorimo o **parametru**. Procjenjujemo li parametar na temelju uzorka, funkcija vrijednosti uzorka naziva se **procjeniteljem**, a vrijednost te funkcije je **procjena parametra**.

Primjer 1:

U Hrvatskoj je 31. prosinca 2022. godine bilo 323 945 registriranih poslovnih subjekata. Za svaki subjekt evidentiran je, pri Državnom zavodu za statistiku (DZS), broj zaposlenih na taj dan, veličina dugotrajne imovine i oblik vlasništva.

a) O kojem je statističkom skupu riječ? Identificirajte populacije.

b) Navedite neke od mogućih karakteristika populacije koje bi se mogle dobiti primjenom statističkih metoda.

c) Odaberite jednu populaciju i objasnite što bi u danom primjeru predočavao uzorak. Navedite moguću primjenu uzorka i značajke dobivenog suda.

Rješenje:

a) Statistički skup čine svi registrirani poslovni subjekti na dan 31.12. 2022. Jedinica tog skupa (jedinica promatranja ili element) je jedna pravna osoba. Skup je realan i definiran pojmovno (zakonska definicija poslovnog subjekta) i vremenski (vremenska točka 31.12. 2022.) Skup broji 323.945 elemenata. Populacije su: (1) 323 945 podataka o broju zaposlenih, (2) 323 945 podataka o veličini dugotrajne imovine, (3) 323 945 podataka o vrsti vlasništva.

b) Primjenom statističkih metoda može se, npr., odrediti najčešći broj zaposlenih, prosječna vrijednost dugotrajne imovine, struktura prema obliku vlasništva. Ovi pokazatelji nazivaju se parametrima jer se određuju pomoću podataka za svaki element statističkog skupa.

c) Ako donosimo sud o prosječnoj veličini kapitala na osnovu statističkog uzorka, odaberemo uzorak registriranih poslovnih subjekata, npr. njih 100 i utvrdimo odgovarajuće podatke u svakom uzorku. Na osnovu uzorka ne može se odrediti vrijednost parametra, već njegova procjena brojem. Iskoristimo li slučajni statistički uzorak, osim procjene brojem možemo odrediti granice intervala procjene u kojemu s određenom vjerojatnošću očekujemo da će se nalaziti parametar. Informacije dobivene pomoću uzorka **sadrže pogrešku** jer su dobivene pomoću dijela, a ne svih podataka.

Vrste i izvori statističkih podataka

Statistički podaci su općenito rezultati mjerenja svojstava elemenata statističkog skupa, njegova podskupa ili eksperimentalnih jedinica. Obilježja (varijable) elemenata statističkog skupa su odgovarajuća svojstva prema kojima uspoređujemo elemente. Svako obilježje se javlja u više pojavnih oblika (**modaliteta**). Skup svih modaliteta nekoga obilježja naziva se skala. Tako je npr. skala modaliteta obilježja ''dan u tjednu'' ponedjeljak, utorak, srijeda, četvrtak, petak, subota i nedjelja, skala modaliteta ''mjesec u godini'' siječanj, veljača, ožujak, travanj, svibanj, lipanj, srpanj, kolovoz, rujan, listopad, studeni i prosinac i slično.

Skale obilježja mogu se podijeliti sukladno određenim metričkim svojstvima. Razlikujemo nominalnu, redoslijednu (ordinalnu ili rang), intervalnu i omjernu skalu.

Nominalna skala zadaje se u obliku skupa čiji su elementi nenumeričke vrijednosti: slovne oznake, pridjevi, kategorije itd. Poredak elemenata skale prilikom definiranja skale je proizvoljan. Mogu se odnositi na opisno izražen varijabilitet danog svojstva (atributivna skala) iil na prostorne jedinice(geografska skala).

U nekim se slučajevima nenumeričke oznake zamjenjuju brojčanima, pri čemu navedene brojčane oznake služe isključivo kao identifikatori, pa s njima nema smisla provoditi nikakve računske operacije. Npr. zamijenimo li opisni modalitet ponedjeljak brojčanim modalitetom 1, a opisni modalitet utorak modalitetom 2, onda nema smisla zapisati niti računati npr. 1+2, 1 · 2 i slično. **Nominalna skala** ponajprije **služi za razvrstavanje (kategorizaciju)** elemenata osnovnoga skupa u kategorije, a ukoliko je njihov broj vrlo velik, skala se utvđuje dogovorno i u takvim slučajevima govorimo o nomenklaturi. Tipični primjeri su različite nomenklature djelatnosti, industrijskih proizvoda, usluga itd.

Redoslijedna (ordinalna, rang) skala elementima statističkoga skupa pridružuje slovne oznake, simbole ili brojeve prema intenzitetu mjerenoga svojstva, pa se elementi potom klasificiraju i uređuju sukladno stupnju promatranoga svojstva. Poredak modaliteta ovdje nije proizvoljan: modalitete obično navodimo od onoga koji označava najmanji stupanj svojstva do onoga koji označava najveći stupanj svojstva ili obrnuto. Tipični primjeri redosljedne skale su skala ocjena (1, 2, 3, 4, 5), skala stupnjeva stručne spreme, skala stupnjeva zadovoljstva kvalitetom nastave predmetom Tjelesna i zdravstvena kultura i slično. Iako ova skala ima bolja metrička svojstva od nominalne, **osnovni nedostatak** joj je **nemogućnost preciznoga utvrđivanja razlika mjerenoga svojstva**: npr. znamo da je ocjena 5 bolja od ocjene 2, ali ne možemo npr. reći da je ocjena 5 za 3 bolja od ocjene 2.

Ukoliko elementima statističkoga skupa pridružujemo realne brojeve sukladno intenzitetu promatranoga svojstva, dobit ćemo **intervalnu** skalu. Ona **uvijek ima definiranu mjernu jedinicu i dogovorno utvrđenu nulu**, pa ima bolja metrička svojstva od prethodnih dviju skala jer **omogućuje i razvrstavanje i rangiranje i brojčano utvrđivanje razlika mjerenoga svojstva**.

S podacima dobivenima mjerenjem na intervalnoj skali jasno su definirane operacije zbrajanja i oduzimanja, pa tako jednake razlike brojeva na skali govore o jednakoj razlici mjerenih svojstava.

Primjer 2:

Ako je temperatura zraka na Zavižanu -10°C, u Zagrebu 0°C, a u Rijeci 12°C, onda se može reći da je u Zagrebu za 10°C toplije nego na Zavižanu, a za 12°C hladnije nego u Rijeci. Također, dogovorno utvđena nula znači da element statističkoga skupa posjeduje određeno svojstvo i da je modalitet toga svojstva jednak nuli. Konkretno, činjenica da je temperatura zraka u Zagrebu 0°C znači da zrak u Zagrebu ima svoju temperaturu i da je ta temperatura jednaka 0°C.

Skala s najboljim metričkim svojstvima je tzv. **omjerna** skala. Ona se dobije tako da se elementima statističkoga skupa pridruže realni brojevi sukladno intenzitetu mjerenoga svojstva, a osim definirane mjerne jedinice, ima i nulu koja označava nepostojanje svojstva. Zbog toga se na takvoj skali mogu definirati sve četiri osnovne računske operacije jer razlike i omjeri vrijednosti imaju suvislo tumačenje.

Primjer 3:

Ako Petar na tekućem računu ima 1 220 €, a Marko 122 €, onda Petar na svom tekućem računu ima 10 puta više novca, odnosno za 1 098 € više od Marka.

Varijable mjerene na numeričkoj skali (intervalnoj i omjernoj) nazivaju se numeričkim (**kvantitativnim**) varijablama. Numerička skala sadrži konačan broja vrijednosti ili teoretski, beskonačno. Među numeričkim varijablama razlikujemo diskretne i kontinuirane varijable.

Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti.

Primjer 4:

Diskretne numeričke varijable su

- broj bodova na državnoj maturi iz matematike,
- broj ulovljenih komaraca u klopku,
- broj dana u godini s temperaturom zraka višom od 35°C.

Skup je mogućih vrijednosti **kontinuiranih** numeričkih varijabli cijeli skup realnih brojeva ili neki interval.

Primjer 5:

Sljedeće su numeričke varijable kontinuirane:

- postotak prolaznosti na pojedinim ispitima tijekom jedne nastavne godine,
- temperatura mora,
- vodostaj neke rijeke

Podaci dobiveni mjerenjima na nominalnoj i redoslijednoj skali ubrajaju se među **kvalitativne** podatke.

U svrhu prikaza podataka i nekih statističkih analiza, vrijednosti se numeričke varijable također mogu svrstati u **kategorije**. Za razliku od kategorija kvalitativnih varijabli, među kategorijama numeričke varijable uvijek se može prepoznati prirodan poredak.

Primjer 6:

Istraživanjem se ispituje preferencija pojedinih vrsta kave pomoću uzoraka osoba starijih od 18 godina na području jedne regije. U upitniku se između ostalih, nalaze i sljedeća pitanja (varijable):

(1) Spol _______ (muški – 0, ženski - 1)

(2) Dob _______ (navršene godine života)

(3) Prosječan broj konzumiranih napitaka kave tijekom jednog mjeseca

(4) Rang marke kave prema preferenciji ______ (marka A, marka B, marka C, marka D). U rangiranju se uz odabranu marku navodi rang izražen brojem pri čemu se broj 1 pridružuje marki koju se najviše preferira, broj 2 marki s manjom preferencijom i tako redom.

(5) Važnost cijene pri izboru marke: _______ (1) cijena je izrazito važna, (2) cijena je važna, (3) cijena nije odlučujuća, (4) cijena nije važna, (5) cijena je najmanje važna.

Podaci u Tablici 1. prikupljeni su anketiranjem 1000 ispitanika i upisani u datoteku. Dio datoteke (matrice podataka) jest:

Tablica 1. Rezultati ankete o preferencijama vrsta kave osoba starijih od 18 godina:

Red. br.	SPOL	DOB	POTROŠNJA	RANG A	RANG B	RANG C	RANG D	CIJENA
1	0	21	10	2	1	3	4	4
2	1	34	25	1	4	3	2	1
3	1	19	28	2	3	1	4	5
4	0	42	5	4	3	2	1	3
5	1	27	15	1	2	3	4	2

Izvor: autor prema Vrste i izvori statističkih podataka, Primijenjena statistika, Školska knjiga 2006., stranica 10.

a) Klasificirajte varijable u upitniku prema metričkim svojstvima. U koju skupinu, obzirom na izvor, pripadaju prikupljeni podaci?

b) Protumačite značenje elemenata navedenog dijela matrice podataka.

Rješenje:

a) (1) nominalna (atributivna) varijabla; (2) numerička kontinuirana varijabla pretvorena u diskretnu varijablu, (3) numerička diskretna varijabla, (4) i (5) su rang – varijable. Istražuje se uzorkom.

b) U matrici podataka prvi je ispitanik muškog spola (kod 0), navršio je 21 godinu, mjesečno u prosjeku 10 puta konzumira kavu, marki A daje rang 2, marku B stavlja na prvo mjesto, marku C na treće, a marku D na četvrto mjesto, pri izboru cijena mu nije važna.

Slika 3. Grupa prijatelja

Pitanje:

Znate li da Državni Zavod za statistiku prikuplja i obrađuje mnoštvo podataka vezanih uz stanovništvo Republike Hrvatske, primjerice, prosječnu plaću, stope inflacije, podatke energetske učinkovitosti, nezaposlenosti, noćenju turista i sličnih? Osim Državnog Zavoda za statistiku provjerite koje podatke objavljuje Hrvatska narodna banka. Pronađite podatke o gospodarskim kretanjima i međunarodnoj robnoj razmjeni naše države. Za više podataka posjetite www.dzs.hr, www.hnb.hr, www.zse.hr, www.hgk.hr, www.worldbank.org.

Analizirajte i metode istraživanja javnog mnjijenja Gallupova instituta na temelju informacija danih na stranici www.gallup.com.

2. UREĐIVANJE I PRIKAZ PODATAKA

Kako su statistički podaci u pravilu mnogobrojni, često nije moguće izravno zaključivanje o osobitostima pojava koje predočavaju. Uređivanjem statističkih podataka nastaju statistički nizovi. Ukoliko se radi o uređivanju modaliteta nominalne varijable, doći ćemo do **nominalnog** niza, **redoslijedni** niz nastaje uređenjem podataka o rang-varijabli. Kronološko nizanje podataka dovodi do **vremenskog** niza.

Jedan se statistički niz pregledno prikazuje tablično u **jednostavnoj tablici**. Više nizova nastalih sređivanjem podataka prema modalitetima iste varijable predočavamo **skupnom tablicom**. Podaci grupirani istodobno prema modalitetima dviju ili više varijabli prikazuju se u kombiniranoj tablici koju još nazivamo **tablicom kontingencije**.

Uređeni podaci prikazani statističkim tablicama omogućuju na jednostavan način uvid u osobitosti analiziranih pojava, a često jasniji uvid u osobitost analizirane pojave pruža grafički prikaz analiziranog statističkog niza.

Predmet analize mogu biti podaci o jednoj varijabli (jednodimenzionalni nizovi) ili o više njih (višedimenzionalna analiza).

Ako u kvalitativnoj analizi proučavamo vrlo mali broj varijabli, navodimo ih nekim redom odabranim po volji ili prema intenzitetu prema mjernoj skali, u protivnom ih grupiramo.

Pri grupiranju se vodimo osnovnom idejom: podijeliti statistički skup na podskupove prema svim modalitetima koji tvore skalu obilježja. Pritom treba poštivati dva načela:

- Isključivost; svaki element statističkog skupa ne može istodobno biti u barem dva različita podskupa
- Iscrpnost; svaki element statističkog skupa mora biti obuhvaćen grupiranjem podataka.

Elementarna analiza podataka u sklopu deskriptivne statistike provodi se pomoću relativnih brojeva. Najčešće su u primjeni: postoci, proporcije, relativne frekvencije, indeksi i relativni brojevi koordinacije (koeficijenti, indikatori).

Utvrdimo osnovne pojmove:

- **Apsolutna frekvencija modaliteta** – ukupan broj pojavljivanja nekoga modaliteta u osnovnom skupu (Pravilo: Zbroj apsolutnih frekvencija cijele skale modaliteta jednak je opsegu statističkog skupa).
- **Relativna frekvencija modaliteta p_i** – omjer je apsolutne frekvencije i opsega statističkog skupa, a postotna relativna frekvencija je relativna frekvencija iskazana u postocima. (Pravilo: Zbroj relativnih frekvencija cijele skale modaliteta jednak je 1 (odnosno, 100%))

Primjer 1:

Za osnovni skup: {1, 1, 2, 2, 2, 2, 4}

Apsolutna frekvencija modaliteta 2: 4;

Postotna relativna frekvencija modaliteta 2: 57.14%

Kod uređenih numeričkih nizova koristimo i kumulativne apsolutne frekvencije „manje od" i „veće od":

- **Kumulativna apsolutna frekvencija "manje od"** nekoga modaliteta – ukupan broj svih elemenata osnovnog skupa koji su **manji ili jednaki tom modalitetu**
- **Kumulativna relativna frekvencija "manje od" nekoga modaliteta** – omjer kumulativne apsolutne frekvencije "manje od" i opsega statističkog skupa
- **Kumulativna apsolutna frekvencija "veće od"** nekoga modaliteta – ukupan broj svih elemenata osnovnoga skupa koji su **veći ili jednaki tom modalitetu**
- **Kumulativna relativna frekvencija "veće od" nekoga modaliteta** – omjer kumulativne apsolutne frekvencije "veće od" i opsega statističkog skupa

Primjer 2:

Za osnovni skup: {1, 2, 3, 4, 5}

Kumulativna apsolutna frekvencija "manje od" modaliteta 3: 3;

Kumulativna relativna frekvencija "manje od" modaliteta 3: 60%;

Kumulativna apsolutna frekvencija "ve će od" modaliteta 2: 4;

Kumulativna relativna frekvencija "manje od" modaliteta 2: 80%;

Primjer 3:

U Tablici 2. prikazana je razdioba osoba koje imaju prebivalište u Gradu Zagrebu i traže zaposlenje, grupirane prema starosti. Podaci su prikupljeni 31. 12. 2021.

Slika 4. Uspješno poslovno okruženje

Tablica 2. Razdioba osoba koje imaju prebivalište u Gradu Zagrebu i traže zaposlenje prema starosti:

Čekanje na zaposlenje [prema starosti]	Broj osoba
15 - 19	254
20 - 24	1 032
25 - 29	1 741
30 - 34	1 453
35 - 39	1 430
40 - 44	1 548
45 - 49	1 640
50 - 54	1 634
55 - 59	2 073
60 i više	1 725
Ukupno:	14 530

Izvor: Statistički ljetopis Grada Zagreba za 2022. godinu, stranica 151.

a) Odredite statistički skup i njegov opseg.

b) Odredite osnovni skup i njegov opseg.

c) Odredite obilježje prema kojemu su podijeljeni elementi statističkoga skupa.

d) Odredite apsolutnu frekvenciju razreda 25 - 29 i objasnite njezino značenje.

e) Izračunajte relativnu frekvenciju razreda 35 - 39 i objasnite njezino značenje.

f) Izračunajte kumulativnu apsolutnu frekvenciju „više od“ razreda 1 – 3 i objasnite njezino značenje.

g) Izračunajte kumulativnu relativnu frekvenciju „manje od“ razreda 50 – 54 i objasnite njezino značenje.

h) Zadanu razdiobu prikažite grafički histogramom. (Za razmak dviju uzastopnih frekvencija na osi ordinata uzmite 500.) Uz grafikon navedite sve potrebne oznake.

Rješenje:

a) Statistički skup tvore osobe koje imaju prebivalište u Gradu Zagrebu i traže zaposlenje 31. 12. 2021. Opseg skupa je 14 530 (osoba).

b) Osnovni skup jednak je statističkom skupu. Opseg je 14 530 (osoba).

c) Obilježje prema kojemu su podijeljeni elementi statističkoga skupa jest čekanje na zaposlenje.

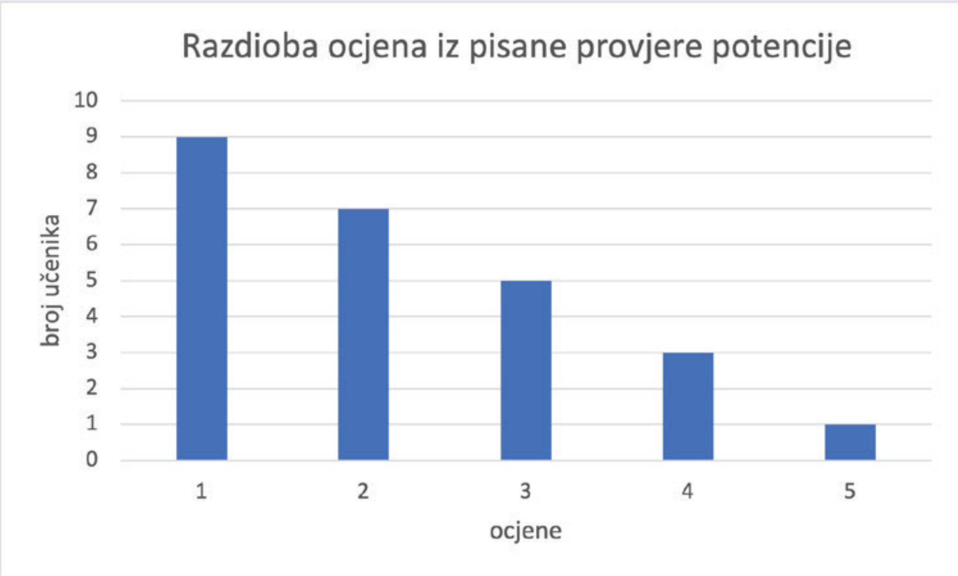
d) Apsolutna frekvencija razreda 25 – 29 jednaka je 1741. Njezino značenje je: 1741 promatrana osoba je čekala na zaposlenje, a ima između 25 i 29 godina.

e) Relativna frekvencija razreda 35 – 39 iznosi (približno) 9.84 %. Njezino značenje je: (Približno) 9.84% promatranih osoba je čekalo na zaposlenje, a u dobnoj je granici 35-39.

f) Kumulativna apsolutna frekvencija „više od“ razreda 20 – 24 iznosi 14 276. Njezino značenje je: 14 276 promatrana osoba je čekala na zaposlenje, a starija je od 19 godina.

g) Kumulativna relativna frekvencija „manje od“ razreda 50 – 54 iznosi (približno) 73.86%. Njezino značenje je: (Približno) 73.86% promatranih osoba je čekala na zaposlenje, a ima najviše 54 godina.

h)



Razdioba svih osoba prema starosti koje imaju prebivalište u Gradu Zagrebu i traže zaposlenje 31. 12. 2021. (izradio: autor)

Prikaz podataka tablicom (tabeliranje)

Tabeliranje je postupak svrstavanja podataka u redove i stupce tablice prema određenom pravilu. Tabličnim načinom prikazivanja olakšava se praćenje statističkih podataka, a time i zaključci o pojavama koje oni predočuju.

Razlikujemo izvještajne i analitičke tablice.

Izvještajne su tablice one u kojima se navode svi prikupljeni podaci.

Analitička tablica sadrži dio uređenih podataka izdvojenih za određenu analizu. Analitičke su tablice manjih dimenzija.

Primjer 1:

Prosječne mjesečne isplaćene bruto plaće po zaposlenome u pravnim osobama svih oblika vlasništva (analitička tablica):

		2020.		2021.	
		kune	euri	kune	euri
ukupno		**7811**	**1037**	**8306**	**1102**
A	Poljoprivreda, šumarstvo i ribarstvo	6585	874	6986	927
B	Rudarstvo i vađenje	11816	1568	12074	1602
C	Prerađivačka industrija	7796	1035	8336	1106
D	Opskrba eleketričnom energijom, plinom, parom i klimatizacija	9884	1312	10365	1376
E	Opskrba vodom; uklanjanje otpadnih voda, gospodarenje otpadom, te djelatnosti sanacije okoliša	6782	900	7167	951
F	Građevinarstvo	6061	804	6298	836
G	Trgovina na veliko i malo; popravak motocikala i vozila	7221	958	7684	1020
H	Prijevoz i skladištenje	7227	959	7484	993
I	Djelatnosti pružaja smještaja te pripreme i usluživanja hrane	4836	642	5121	680
J	Informacije i komunikacije	9831	1305	11541	1532
K	Financijske djelatnosti i djelatnosti osiguranja	10962	1455	11541	1532
L	Poslovanje nekretninama	6830	906	7418	985
M	Stručne, znanstvenei tehničke djelatnosti	8263	1097	8841	1173
N	Administrativne i pomoćne uslužne djelatnosti	5459	725	5605	744
O	Javna uprava i obrana; obavezno socijalno osiguranje	8497	1128	9066	1203
P	Obrazovanje	7804	1036	8352	1109
Q	Djelatnosti zdravstvene zaštite i socijalne skrbi	8528	1132	9254	1228
R	Umjetnost, zabava i rekreacija	7074	939	7536	1000
S	Ostale uslužne djelatnosti	6187	821	6434	854

Izvor: Statistički ljetopis Grada Zagreba za 2022. godinu, stranica 156.

Primjer 2:

Prosječne mjesečne isplaćene neto plaće po zaposlenom u pravnim osobama prema spolu i stupnju stručne spreme (izvještajna tablica):

		spol		stupanj stručne spreme			
	ukupno	muškarci	žene	visoka stručna sprema	viša stručna sprema	srednja stručna sprema	niža stručna sprema
2015.	6626	7040	6198	9518	6929	5211	4047
2016.	6740	7238	6243	9598	7020	5327	3989
2017.	6990	7489	6481	9843	7249	5492	4130
2018.	7187	7740	6635	10067	7393	5654	4249
2019.	7468	7988	6944	10374	7726	5885	4459
2020.							
kune	7679	8123	7212	10682	7995	6014	4580
euri	1019	1078	957	1418	1061	798	608

Izvor: Statistički ljetopis Grada Zagreba za 2022. godinu, stranica 158.

Osim na izvještajne i analitičke, tablice dijelimo na: jednostavne, skupne i kombinirane.

Jednostavna statistička tablica sadrži jedan statistički niz.

U **skupnoj statističkoj tablici** nalazimo dva ili više statističkih nizova. Nizovi prikazani u skupnoj tablici odnose se na podatke različitih skupova, uređenih prema oblicima iste varijable (obilježja).

U **kombiniranoj tablici** prikazani su podaci grupirani istodobno prema dva ili više obilježja.

Primjer 3:

Jednostavna statistička tablica:

Učenici u osnovnim školama grada Zagreba, u školskoj godini 2020./2021.	
Spol	**Broj učenika**
muški	60663
ženski	29529
ukupno	**90192**

Primjer 4:

Skupna statistička tablica:

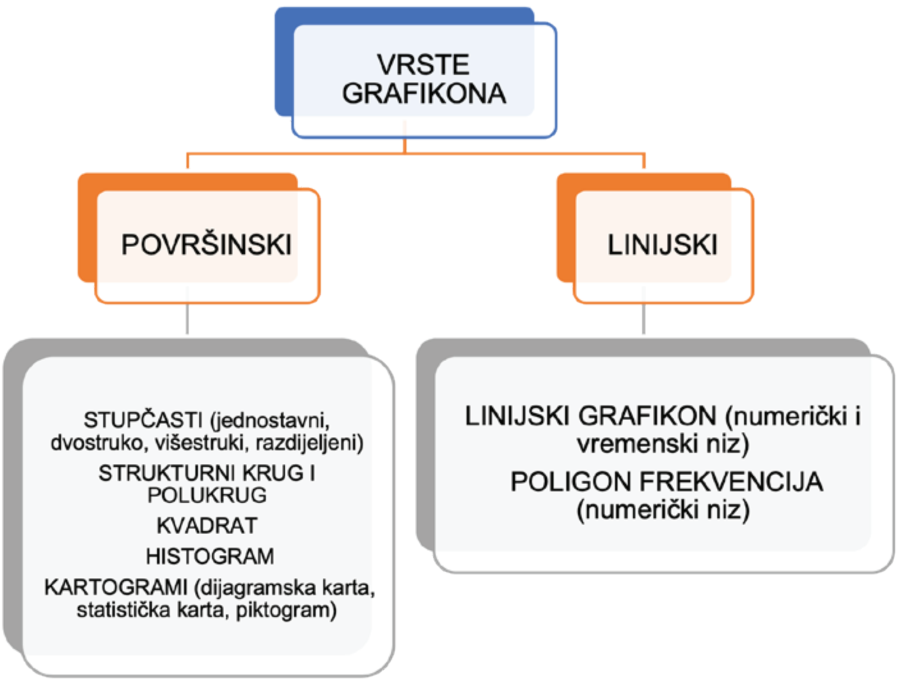
Stanovništvo, stanovi i kućanstva najvećih gradova u Republici Hrvatskoj prema popisu:			
Grad	**Broj stanovnika**	**Broj stanova**	**Broj kućanstava**
Zagreb	769944	393433	300650
Split	161312	83360	60015
Rijeka	108622	63333	48028
Osijek	96848	51556	39359
ukupno	**1136726**	**591682**	**448052**

Izvor: Popis stanovništva, kućanstava i stanova 2021.

Primjer 5:

Kombinirana statistička tablica – prikazuje podatke grupirane prema dva ili više obilježja je tablica u Primjeru 2: Prosječne mjesečne isplaćene neto plaće po zaposlenom u pravnim osobama prema spolu i stupnju stručne spreme.

Grafičko prikazivanje podataka



Primjer 6:

Jednostavni stupčasti:

Tablica: Anketirani inozemni putnici u Hrvatskoj od lipnja do prosinca 1998.		
Vrsta prijelaza	**broj anketiranih putnika**	**Struktura u %**
	f_i	P_i
cestovni	17253	87,5
zračni	999	5,1
željeznički	773	3,9
pomorski	693	3,5
ukupno	**19718**	**100,0**

Prikaz nominalnog niza jednostavnim stupcima

Primjer 7:

Jednostavni stupčasti s položenim stupcima:

Anketirani inozemni putnici u Hrvatskoj od lipnja do prosinca 1998.			
Država	**Broj anketiranih putnika**	**Struktura u %**	**Kumulativni postotak**
Slovenija	4559	23,12	23,12
Njemačka	3411	17,30	40,42
Bosna i Hercegovina	3212	16,29	56,71
Italija	2962	15,02	71,73
Austrija	1937	9,83	81,56
Mađarska	901	4,57	86,13
Republika Češka	545	2,76	88,89
Ostale države	2191	11,11	100
ukupno	**19718**	**100,0**	

Prikaz niza (nominalni, geografski) položenim stupcima

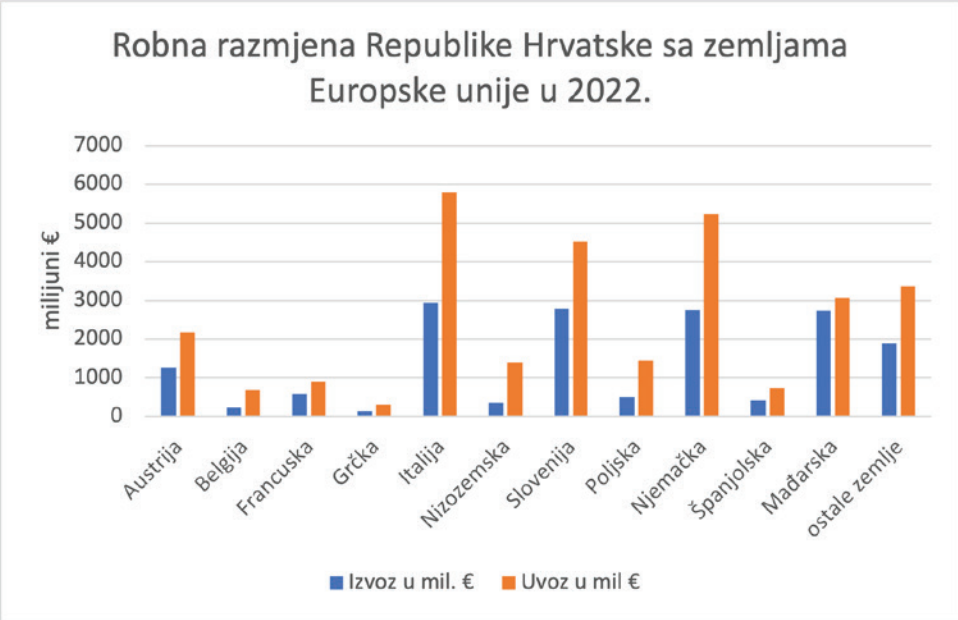
Primjer 8:

Dvostruki stupci:

Tablica: Uvoz i izvoz Republike Hrvatske u zemlje Europske unije 2022. godine

Zemlja	Izvoz u mil. €	Uvoz u mil €	Struktura izvoza u %	Struktura uvoza u %	Koeficijent pokrivenosti uvoza
Austrija	1275	2169	7,68%	7,33%	58,80
Belgija	233	688	1,40%	2,33%	33,79
Francuska	584	887	3,52%	3,00%	65,87
Grčka	147	298	0,89%	1,01%	49,40
Italija	2947	5799	17,76%	19,60%	50,82
Nizozemska	352	1397	2,12%	4,72%	25,18
Slovenija	2789	4519	16,81%	15,28%	61,73
Poljska	508	1449	3,06%	4,90%	35,08
Njemačka	2740	5226	16,51%	17,67%	52,42
Španjolska	408	724	2,46%	2,45%	56,41
Mađarska	2730	3060	16,45%	10,34%	89,23
ostale zemlje	1883	3365	11,34%	11,37%	55,95
ukupno	**16597**	**29582**	**100,00%**	**100,00%**	**56,11**

Izvor: Robna razmjena Republike Hrvatske s inozemstvom u 2022., DZS

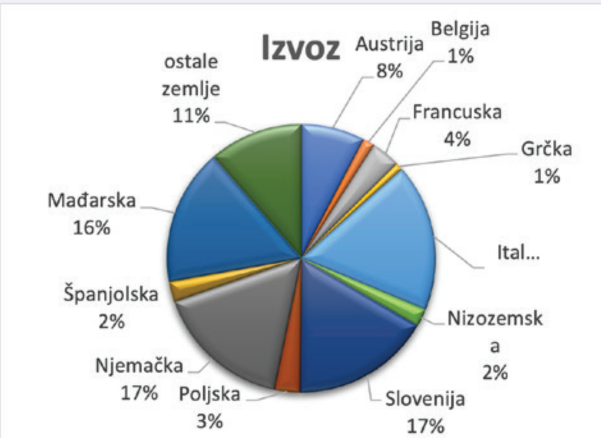


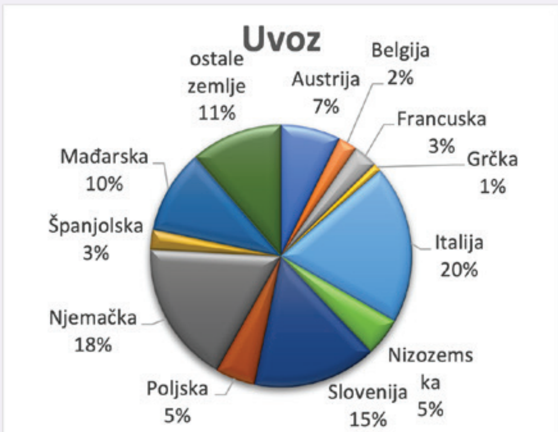
Usporedba uvoza i izvoza dvostrukim stupcima

Uspoređuje li se struktura dvaju ili više nizova strukturnim krugovima, potrebno je nacrtati krugove istih radijusa. Predodžba o razlikama strukture dobije se tako da se usporede veličine sektora kruga koje predočavaju veličine vezane za iste modalitete nizova. Broj stupnjeva svakog kruga određuje se pomoću izraza $s_i = \frac{f_i}{N}360$, odnosno $s_i = P_i \cdot 3.6$, $N = \sum_{i=1}^{k} f_i$

Primjer 9:

Strukturni krugovi jednakih polumjera:



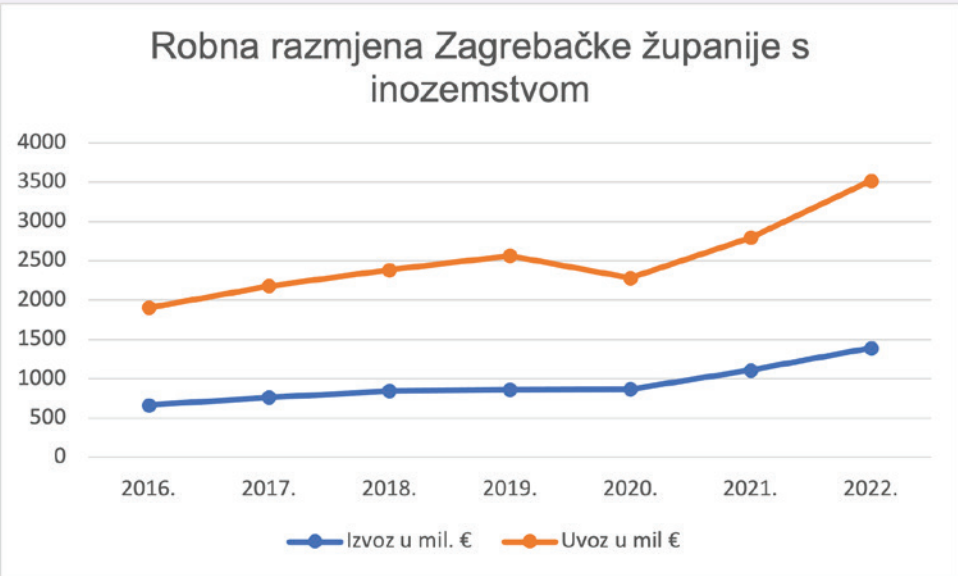


Linijski grafikon služi za grafičko prikazivanje vremenskih nizova, nizova kumulativnih frekvencija, te za usporedbu dvaju ili više nizova kvantitativnih podataka.

Tablica: Uvoz i izvoz Zagrebačke županije

	Izvoz u mil. €	**Uvoz u mil €**
2016.	664,84	1906,505
2017.	764,61	2179,609
2018.	846,751	2383,084
2019.	859,452	2566,804
2020.	867,458	2278,874
2021.	1107,117	2793,229
2022.	1391,703	3518,757
ukupno	**6501,931**	17626,862

Izvor: Robna razmjena s inozemstvom – pregled po županijama, DZS



Usporedba dvaju nizova kvantitativnih podataka

Jeste li znali?

Osim gore navedenih najčešće korištenih grafikona u uporabi su kartogrami koji služe za prikaz geografskih statističkih nizova. Vrste kartograma su dijagramska karta, piktogram i statistička karta.

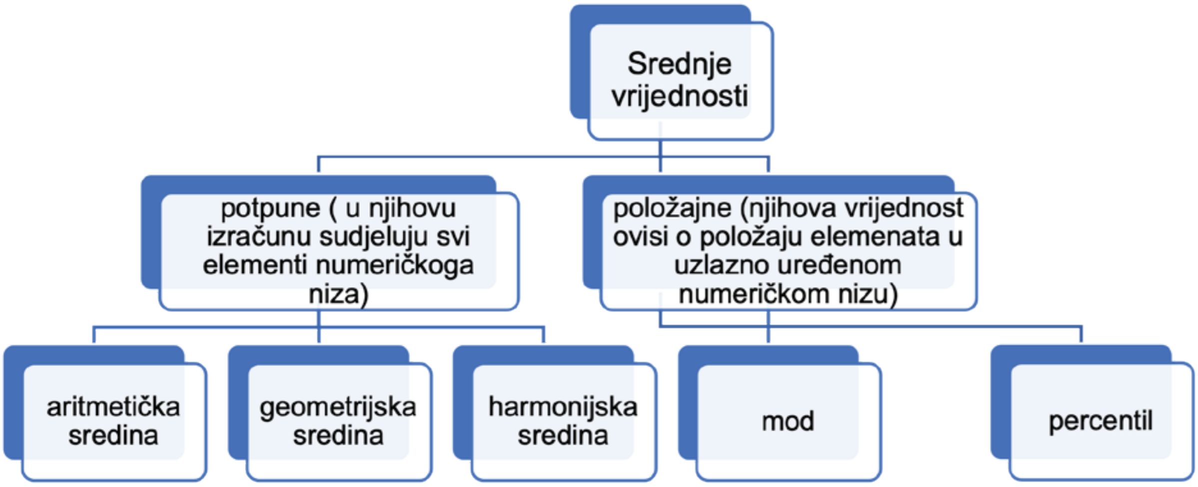
3. OSNOVE OPISNE STATISTIKE

Statistički skup, a samim tim i osnovni skup vrlo često znaju imati vrlo veliki opseg. Primjer tako velikog skupa je oko 3,8 milijuna podataka o jednom obilježju prikupljenih u posljednjem popisu stanovništva Republike Hrvatske.

Ako je riječ o **kvantitativnom** obilježju, takvi podaci se obavezno **grupiraju u razrede**.

Ipak, unatoč takvom grupiranju, radi kratkoće i jednostavnosti opisivanja prikupljenih podataka potrebno je imati pokazatelje koji će dovoljno dobro reprezentirati osnovni skup (npr. 3.8 milijuna podataka zamijeniti jednim podatkom).

U tu se svrhu uvode **srednje vrijednosti**. One se najvećim dijelom odnose isključivo na modalitete kvantitativnoga obilježja koje tvore konačan numerički niz podataka.



Potpune srednje vrijednosti

Aritmetička sredina – predstavlja **prosjek svih vrijednosti elemenata numeričkoga niza**. Ona je dobar reprezentant numeričkoga niza ako taj niz ne sadrži ekstremno velike i/ili ekstremno male vrijednosti.

Kod opisivanja **prosjeka relativnih promjena neke pojave** (npr. godišnja proizvodnja čokolade) koristi se i **geometrijska** sredina. Njezina primjena je najvećim dijelom u računanju tzv. verižnih i skupnih indeksa.

Harmonijska sredina primjenjuje se kad je veličina kojoj se određuje prosječna vrijednost obrnuto razmjerna veličini kojom se određuje prosječna vrijednost (npr. kad se prosječna produktivnost određuje utrošenim vremenom za izradu nekog proizvoda). Mjerna jedinica harmonijske sredine jednaka je mjernoj jedinici varijable za koju se određuje.

Sve tri navedene vrijednosti mogu se računati za negrupirane podatke, odnosno procijeniti za grupirane podatke (podatke grupirane u razrede).

Srednja vrijednost	Izračun iz negrupiranih podataka	Izračun iz grupiranih podataka
Aritmetička sredina	$\underline{x}=\frac{\sum_{i=1}^{n} x_i}{n}=\frac{x_1+x_2+\ldots+x_n}{n}$	$\underline{x}=\frac{\sum_{i=1}^{k} f_i \cdot x_i}{f_1+f_2+\ldots+f_k}$
Geometrijska sredina	$G=\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$	$G=\sqrt[\sum_{i=1}^{k} f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_k^{f_k}}$
Harmonijska sredina	$H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$	$H=\frac{f_1+f_2+\ldots+f_k}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots+\frac{f_k}{x_k}}=\frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k}\frac{f_i}{x_i}}$

Napomena: U slučaju grupiranih podataka su x_i i f_i redom razredna sredina, odnosno apsolutna frekvencija i-toga razreda, za svaki $i=1,\ldots,k$.

Ukoliko se u nizu podataka pojavljuje ekstremna vrijednost, podatak koji znatno odstupa od ostalih, taj podatak znatno utječe na aritmetičku sredinu. Tada aritmetička sredina nije dobra mjera srednje vrijednosti, već će takve podatke puno bolje opisati mjera srednje vrijednosti medijan.

Položajne srednje vrijednosti

Mod

- Oznaka Mo – to je modalitet s najvećom apsolutnom/relativnom frekvencijom (srednja vrijednost koja se može određivati i za kvalitativna i za kvantitativna obilježja).
- Za razliku od ostalih srednjih vrijednosti, mod ne mora biti jedinstven. Tako razdiobe mogu biti unimodalne (imaju točno jedan mod), bimodalne (imaju točno dva različita moda) ili multimodalne (imaju barem tri različita moda).
- Ako je razdioba unimodalna, onda je mod dobar reprezentant numeričkoga niza ako pripadna relativna frekvencija moda nije manja od 50%.
- Analogni kriteriji vrijede i za ostala dva tipa razdioba (zavisno o broju modova).
- Ako su **podaci grupirani u prave razrede**, onda vrijednost moda procjenjujemo prema sljedećoj formuli:

$$Mo=L_i+\frac{b-a}{2\cdot b-(a+c)}\cdot s$$

gdje su:

L_i – donja granica modalnoga razreda;

b – korigirana apsolutna/relativna frekvencija modalnoga razreda (dobivena dijeljenjem originalne frekvencije i originalne razredne širine);

a – korigirana apsolutna frekvencija koja neposredno prethodi frekvenciji b (dobivena analogno kao frekvencija b)

c – korigirana apsolutna/relativna frekvencija koja neposredno slijedi iza frekvencije b (dobivena analogno kao frekvencija b)

s – razredna širina modalnog razreda.

Napomena: Ako svi razredi imaju jednaku širinu, nije potrebno računati korigirane frekvencije (mogu se koristiti originalne frekvencije).

Pretpostavimo da je numerički niz uzlazno uređen, tj. da su svi numerički podaci poredani od najmanjega prema najvećemu.

Percentili su položajne srednje vrijednosti koje dijele numerički niz na ukupno 100 jednakobrojnih dijelova. Ima ih ukupno 99. Označavaju se s P_k, za k = 1, 2, ..., 99.

Za svaki k = 1, ..., 99 vrijednost P_k iz negrupiranih podataka računamo prema formuli:

$$P_k = \begin{cases} x_{\left[\frac{k \cdot n}{100}\right]}, & ako\ n\ nije\ djeljivo\ sa\ 100 \\ \frac{1}{2} \cdot \left(x_{\frac{k \cdot n}{100}} + x_{\frac{k \cdot n}{100}+1} \right), & ako\ je\ n\ djeljiv\ sa\ 100. \end{cases}$$

Ako su podaci grupirani u prave razrede, ove vrijednosti procjenjujemo po formuli:

$$P_k = L_1 + \frac{\frac{k}{100} \cdot n - \sum_{i=1}^{m-1} f_i}{f_m} \cdot h$$

gdje su:

L_i – donja granica razreda kojemu pripada P_k.

m – redni broj razreda (u uzlaznom poretku svih razreda)

h – širina razreda kojemu pripada P_k

f_i – apsolutna/relativna frekvencija i-toga razreda.

Značenje vrijednosti P_k je:

- k% elemenata niza nije veće od P_k.
- Ekvivalentno, $(100-k)$% elemenata niza nije manje od P_k.

U praksi se najčešće koriste 25., 50. i 75. percentil. Oni imaju svoja posebna imena i oznake, tako 25. percentil označava se s Q_1 i naziva prvi ili donji kvartil, 50. percentil označava se s Q_2 ili Me, te naziva drugi kvartil ili medijan, 75. percentil označava se s Q_3 i naziva treći ili gornji kvartil.

Kvartil	**Izračun iz negrupiranih podataka**	**Izračun (procjena) iz grupiranih podataka**
Q_1	$Q_1 = \begin{cases} x_{\left[\frac{n}{4}\right]}, & ako\ n\ nije\ djeljiv\ s\ 4 \\ \frac{1}{2} \cdot \left(x_{\frac{n}{4}} + x_{\frac{n}{4}+1} \right), & ako\ je\ n\ djeljiv\ s\ 4. \end{cases}$	$Q_1 = L_1 + \frac{\frac{1}{4} \cdot n - \sum_{i=1}^{m-1} f_i}{f_m} \cdot h$
$M_e = Q_2$	$Q_2 = \begin{cases} x_{\left[\frac{n}{2}\right]}, & ako\ je\ n\ neparan \\ \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & ako\ je\ n\ paran. \end{cases}$	$M_e = Q_2 = L_1 + \frac{\frac{1}{2} \cdot n - \sum_{i=1}^{m-1} f_i}{f_m} \cdot h$
Q_3	$Q_3 = \begin{cases} x_{\left[\frac{3n}{4}\right]}, & ako\ n\ nije\ djeljiv\ s\ 4 \\ \frac{1}{2} \cdot \left(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1} \right), & ako\ je\ n\ djeljiv\ s\ 4. \end{cases}$	$Q_3 = L_1 + \frac{\frac{3}{4} \cdot n - \sum_{i=1}^{m-1} f_i}{f_m} \cdot h$

Napomena:

n je opseg osnovnoga skupa (duljina numeričkoga niza);

L_i je donja granica pravoga razreda kojemu pripada dotični kvartil;

m je redni broj;

h je širina toga razreda;

f_i je apsolutna/ relativna frekvencija i-tog razreda.

Reprezentativnost aritmetičke sredine, medijana i kvartila utvrđuje se pomoću posebnih statističkih pokazatelja. Riječ je o tzv. mjerama raspršenja (disperzije). Također, prilikom procjenjivanja srednjih vrijednosti iz podataka grupiranih u neprave (nominalne) razrede potrebno je te razrede pretvoriti u prave, pa potom provesti procjene.

Postupak pretvorbe nominalnih razreda u prave razrede je sljedeći:

- Korak 1. Donja granica prvoga razreda i gornja granica posljednjeg a razreda ostaju nepromijenjene.
- Korak 2. Gornja granica i-toga pravoga razreda jednaka je aritmetičkoj sredini gornje granice i-toga nominalnoga razreda i donje granice (i+1)-vog nominalnoga razreda, za i = 1, 2, ..., n.
- Korak 3. Donja granica (i+1)-vog razreda jednaka je gornjoj granici i-toga razreda, za i = 1, 2, ..., n–1.

Primjer 1.

Za grafički zadanu razdiobu odredite i objasnite značenje sljedećih statističkih pokazatelja:

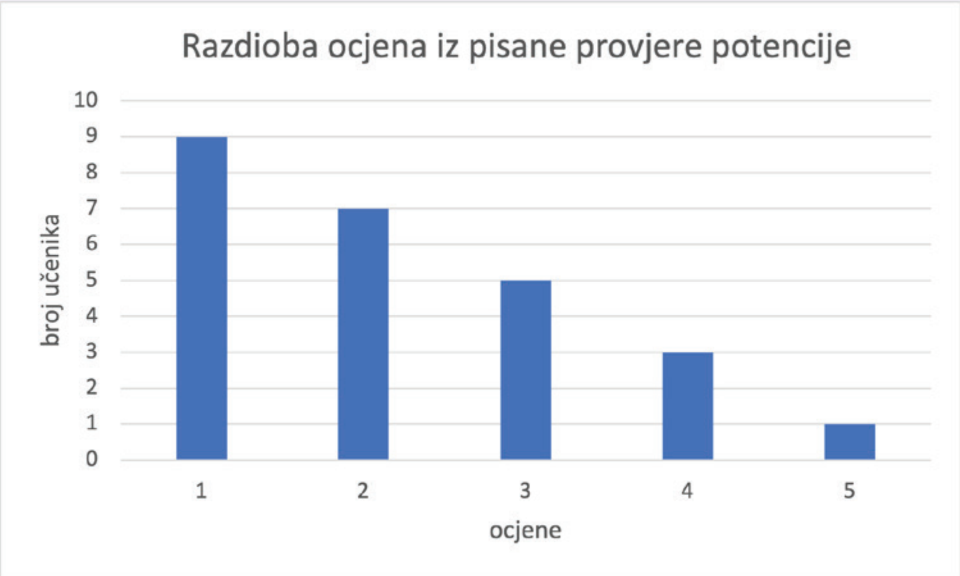
a) aritmetička sredina;

b) mod;

c) donji kvartil;

d) medijan;

e) gornji kvartil.



Rješenje:

Računamo redom:

$$n = 9 + 7 + 5 + 3 + 1 = 25$$

$$\underline{x} = \frac{9 \cdot 1 + 7 \cdot 2 + 5 \cdot 3 + 3 \cdot 4 + 1 \cdot 5}{25} = 2.2$$

$$M_o = 1$$

$$Q_1 = x_{\left\lceil \frac{25}{4} \right\rceil} = x_7 = 1 \text{ (25 nije djeljivo s 4)}$$

$$Me = x_{\left\lceil \frac{25}{2} \right\rceil} = x_{13} = 2$$

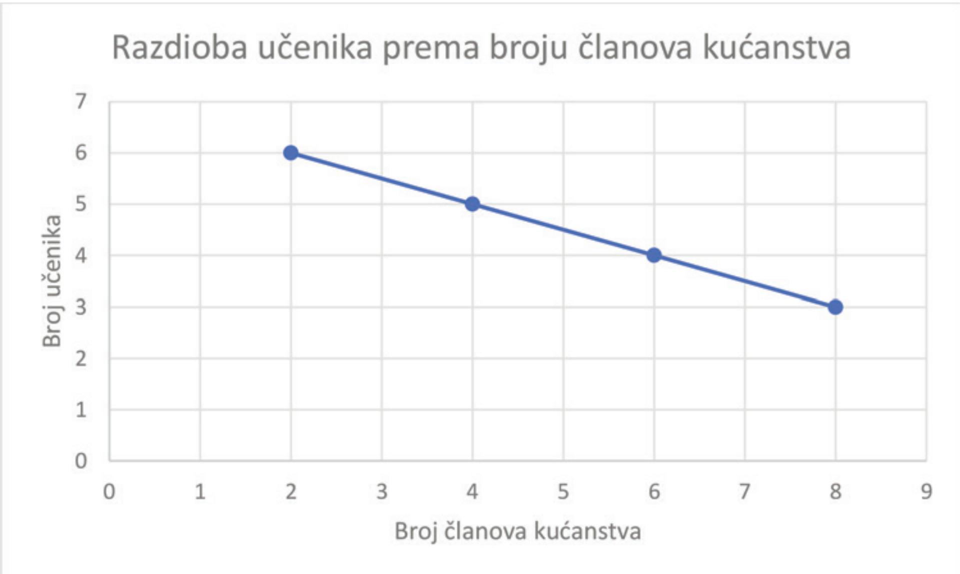
$$Q_3 = x_{\left\lceil \frac{3 \cdot 25}{4} \right\rceil} = x_{19} = 3$$

Interpretacije su redom:

- Prosjek svih ocjena po jednom učeniku je 2.2. (Prosječne ocjena ispita je 2.2) (AS)
- Najčešća ocjena na ispitu je 1(mod)
- 25% svih učenika je ocijenjeno ocjenom nedovoljan (ili manjom od 1). (donji kvartil)
- Polovica svih učenika ocijenjena je ocjenom nedovoljan ili dovoljan (Me)
- 75% svih učenika ocijenjeno je ocjenom dobar ili manjom od dobar. (gornji kvartil)

Provjerite svoje znanje

Provedite interpretaciju i izračune za sljedeći grafikon:



Odredite:

a) aritmetičku sredinu;

b) mod;

c) donji kvartil;

d) medijan;

e) gornji kvartil.

4. MJERE RASPRŠENJA (DISPERZIJE)

Mjerama raspršenja brojčano se opisuje stupanj varijabilnosti statističkih podataka. U tu se svrhu uvode mjere raspršenja. Njihova je svrha opisati raspršenost elemenata osnovnoga skupa u odnosu na aritmetičku sredinu, odnosno medijan.

Razlikujemo apsolutne i relativne mjere raspršenja.

U **apsolutne** mjere raspršenja ubrajamo:

- raspon varijacije;
- interpercentil;
- srednje apsolutno odstupanje;
- varijancu (disperziju);
- standardnu devijaciju

Raspon varijacije (oznaka: R) je razlika najveće i najmanje vrijednosti elemenata osnovnoga skupa, $R_x = x_{max} - x_{min}$.

On je relativno gruba i neprecizna mjera raspršenja. Ne procjenjuje se za podatke grupirane u razrede.

Interpercentil je razlika bilo kojih dvaju percentila P_k i P_l uz uvjet k ≥ l. On predstavlja raspon varijacije središnjih (k – l)% elemenata niza.

Ta mjera je bolja od raspona varijacije, a dodatno zahtijeva uzlazno uređivanje polaznoga niza. Za k = 75 i l = 25 dobiva se interkvartil (oznaka: I_q). On predstavlja raspon varijacije središnje polovice elemenata niza.

Najčešće korištena mjera je **standardna devijacija**. Međutim, razvojem računalnih programa sve više se koristi i srednje apsolutno odstupanje. Formule za izračunavanje tih veličina iz negrupiranih podataka, te procjenu njihove vrijednosti iz podataka grupiranih u razrede navedene su u sljedećoj tablici:

Mjera raspršenja	Izračun iz negrupiranih podataka	Izračun (procjena) iz grupiranih podataka
Srednje apsolutno odstupanje	$MAD = \frac{\lvert x_1 - \underline{x}\rvert + \lvert x_2 - \underline{x}\rvert + \ldots + \lvert x_n - \underline{x}\rvert}{n}$	$MAD = \frac{f_1 \cdot \lvert x_1 - \underline{x}\rvert + f_2 \cdot \lvert x_2 - \underline{x}\rvert + \ldots + f_n \cdot \lvert x_n - \underline{x}\rvert}{f_1 + f_2 + \ldots + f_n}$
varijanca	$\sigma^2 = \frac{\lvert x_1 - \underline{x}\rvert^2 + \lvert x_2 - \underline{x}\rvert^2 + \ldots + \lvert x_n - \underline{x}\rvert^2}{n}$	$\sigma^2 = \frac{f_1\lvert x_1 - \underline{x}\rvert^2 + f_2\lvert x_2 - \underline{x}\rvert^2 + \ldots + f_n\lvert x_n - \underline{x}\rvert^2}{f_1 + f_2 + \ldots + f_n}$
standardna devijacija	$\sigma = \sqrt{\sigma^2}$ (drugi korijen iz varijance)	$\sigma = \sqrt{\sigma^2}$

U **relativne** mjere raspršenja ubrajamo:

- **koeficijent varijacije** (mjera raspršenja u odnosu na aritmetičku sredinu);
- **koeficijent kvartilne devijacije** (mjera raspršenja u odnosu na medijan).

Ove mjere nemaju mjernu jedinicu, pa se njihove vrijednosti najčešće iskazuju u postotcima, one služe i kao najbolji pokazatelji varijabiliteta elemenata osnovnoga skupa.

Koeficijent varijacije definira se kao relativno srednje odstupanje od aritmetičke sredine. Označava se s V i računa prema formuli: $V = \frac{\sigma}{\underline{x}}$.

Koristeći taj koeficijent, varijabilitet se obično klasificira ovako:

V	0 - 10	10 - 30	30 - 50	50 - 70	≥70
interpretacija	vrlo slab	relativno slab	umjeren	relativno jak	jak

Koeficijent kvartilne devijacije predstavlja relativnu mjeru raspršenja elemenata osnovnoga skupa u odnosu na medijan. Označava se s V_q i računa prema formuli:

$$V_q = \frac{I_q}{Q_1 + Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Koristeći taj koeficijent, varijabilitet se obično klasificira ovako:

V_q	0.0 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 - 0.5	0.5 - 1
interpretacija	vrlo slab	relativno slab	umjeren	relativno jak	jak

Primjer:

U tablici su zadani podaci o broju stanovnika i liječnika u 15 država.

a) Izračunajte i interpretirajte vrijednost koja dijeli niz podataka o broju stanovnika od najmanje do najveće vrijednosti u omjeru 1:3.

b) Izračunajte i interpretirajte vrijednost koja dijeli niz podataka o broju stanovnika od najmanje prema najvećoj vrijednosti u omjeru 3:1.

c) Izračunajte i interpretirajte relativnu mjeru disperzije središnjih 50% broja liječnika.

d) Izračunajte i interpretirajte relativnu mjeru disperzije središnjih 50% broja liječnika.

e) Izračunajte i interpretirajte raspon podataka o broju stanovnika.

f) Izračunajte i interpretirajte vrijednost koja odvaja 30% najvećeg broja liječnika.

g) Razlikuju li se države više po broju stanovnika ili broju liječnika?

Broj stanovnika (milijuni)	broj liječnika (tisuće)
25	30
11	28
100	16
4	5
0,6	1,6
12	4
3	1,2
11	7
12	8

Broj stanovnika (milijuni)	broj liječnika (tisuće)
54	20
8	6
10	3
13	11
25	16
150	30

Rješenje (unesite u tablicu Excela podatke počevši od ćelije A1), a zatim provedite račune i interpretacije:

a) Q_1 = QUARTILE(A2:A16;1) (=9.00) (Interpretacija: jedna četvrtina država ima manje ili jednako 9 milijuna stanovnika, a tri četvrtine više.)

b) Q_3 = QUARTILE(A2:A16;3) (=25.00) (Interpretacija: tri četvrtine država imaju manje ili jednako 25 milijuna stanovnika.)

c) Izračunajte prvi i treći kvartil za broj liječnika: Q_1 = 4.50; Q_3 = 18.00; Apsolutna mjera disperzije za srednjih 50% bit će interkvartil: $I_q = Q_3 - Q_1 = 13.50$ (Interpretacija: Raspon središnjih 50% država po broju liječnika je 13,50 tisuća)

Izračunajte koeficijent kvartilne devijacije (jer ne računamo za sve podatke već samo za središnjih 50%):

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.6$$

(Interpretacija: obzirom da je koeficijent kvartilne devijacije veći od 0,5 varijabilnost srednjih 50% broja liječnika je vrlo velika.)

d) $R_x = x_{max} - x_{min}$ = 149,40 (Raspon država prema broju stanovnika iznosi 149.40 milijuna stanovnika.)

e) Izračunajte centile: =PERCENTILE(A2:A16;10%) (Interpretacija: 10% država imaju manje ili jednako 3.4 milijuna stanovnika.)

f) Izračunajte centile: =PERCENTILE(B2:B16;70%); (Interpretacija: 30% država imaju više od 16 milijuna stanovnika; odnosno 70% država imaju manje ili 16 tisuća liječnika.)

g) Pogledajte kolika je disperzija; izračunajte varijabilnost u postotcima; najprije koeficijent varijacije za broj stanovnika (prvo izračunajte standardnu devijaciju koju zatim podijelite s prosjekom: V=STDEVP(A2:A16)/AVERAGE(A2:A16) =139%

Koeficijent varijacije za broj liječnika: V=STDEVP(B2:B16)/AVERAGE(B2:B16) =80%

Zaključak: više variraju po broju stanovnika jer je koeficijent varijacije za broj stanovnika veći.

Regresija i korelacija

Mnoštvo je slučajeva koji se odnose na istraživanje međusobnog odnosa dvaju ili više pojava – promjena jedne pojave uvjetovana je promjenama druge ili drugih pojava.

Povezanost pojava može biti:

- funkcionalna – veze se mogu predočiti izrazima na temelju kojih se točno utvrđuje vrijednost jedne za danu vrijednost druge (drugih) vrijednosti
- statistička – jednoj vrijednosti jedne pojave odgovara više vrijednosti druge (drugih) pojava

Pri istraživanju masovnih pojava analizom utvrđujemo vezu među pojavama po obliku (linearna ili zakrivljena), smjeru (pozitivna ili negativna) i jakosti (funkcionalna ili statistička).

Model koji sadrži jednu zavisnu i jednu nezavisnu varijablu naziva se modelom jednostavne regresije, a model sa dvije ili više nezavisnih varijabli model višestruke regresije.

Za određivanje oblika regresije kao jednostavno sredstvo služi dijagram rasipanja koji se konstruira tako da se u koordinatni sustav unose parovi vrijednosti varijable X i Y, tj. on se sastoji od točaka (xi, yi), a zatim iz rasporeda točaka zaključujemo o obliku, smjeru i jakosti veze.

Koeficijent korelacije je mjera koja nam pokazuje koliko su određene veličine povezane).

Najpoznatiji koeficijenti korelacije:

- Pearsonov: varijable mjerene na intervalnoj ili omjernoj ljestvici mjerenja, zahtjeva **linearan** i normalan raspored podataka te dovoljno velik uzorak
- Spearmanov: jedna ili obje varijable mjerene ordinalnom mjernom ljestvicom, ne postavlja uvjet linearnosti, simetričnosti niti veličine uzorka.

U programu MS Excel oba koeficijenta možemo lako izračunati koristeći se funkcijama i formulama:

- Za Paersonov koeficijent =PAERSON(;)
- Za Spearmanov koeficijent najprije formiramo dva nova stupca (Rx i Ry) u kojima rangiramo varijable naredbom =RANK(; raspon podataka), zatim kreiramo novi stupac di razliku rankova, zatim kreiramo novi stupac u koji računamo di2. Podatke iz ovog stupca zbrojimo i dobijemo vrijednost $\sum d_i^2$.

  Računamo $r = 1 - \frac{6 \cdot \sum d_i^2}{n^3 - n}$

Jačina povezanosti dana je tablicom:

Apsolutna vrijednost koeficijenta korelacije	Jačina povezanosti između varijabli
$\lvert r\rvert = 1$	Potpuna korelacija
$0{,}8 \le \lvert r\rvert < 1$	Jaka korelacija
$0{,}5 \le \lvert r\rvert < 0{,}8$	Srednje jaka korelacija
$0{,}2 \le \lvert r\rvert < 0{,}5$	Relativno slaba korelacija (*)
$0 < \lvert r\rvert < 0{,}2$	Neznatna korelacija
$\lvert r\rvert = 0$	Potpuna odsutnost korelacije

Primjer 1:

Zadani su podaci o vremenu u učenju za ispit i ocjeni na ispitu 10 učenika. Izračunajte i interpretirajte Pearsonov i Spearmanov koeficijent korelacije između vremena provedenog učenju i postignute ocjene na ispitu.

Vrijeme učenja (sati)	Ocjena na ispitu
5	4
4	2
6	3
9	5
10	4
12	3
11	4
15	5
20	3
5	2

Rješenje: Pearsonov koeficijent = PEARSON(A2:A11;B2:B11) (=0,33). Korelacija između vremena učenja i dobivene ocjene je relativno slaba (*), ali pozitivna. To znači da vrijeme provedeno u učenju nije jedini uvjet za postizanje boljeg uspjeha.

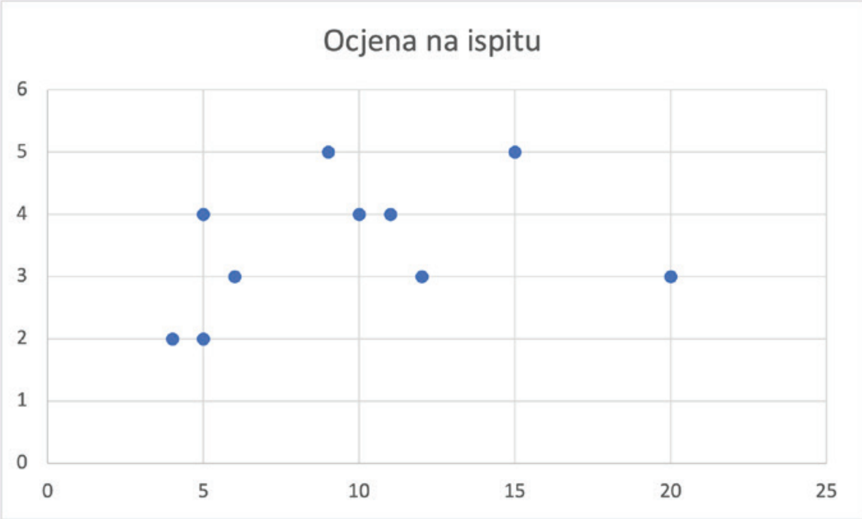
Spearmanov koeficijent: Prvi stupac (rangiranje; ćelija c2) =RANK(A2;$A$2:$A$11));Drugi stupac (rangiranje; ćelija d2 =RANK(B2;$B$2:$B$11)), formiramo stupac razlika rangova: (=c2-d2), formiramo stupac d22, a zatim zbrojimo elemente u posljednjem stupcu.)

Računamo $r = 1 - \frac{6 \cdot \sum d_i^2}{n^3 - n}$

K7 f_x

	A	B	C	D	E	F	G	H	I
1	Vrijeme učen	Ocjena na ispitu	Rx	Ry	di	di^2			
2	5	4	8	3	5	25			
3	4	2	10	9	1	1			
4	6	3	7	6	1	1			
5	9	5	6	1	5	25			
6	10	4	5	3	2	4			
7	12	3	3	6	-3	9			
8	11	4	4	3	1	1			
9	15	5	2	1	1	1			
10	20	3	1	6	-5	25			
11	5	2	8	9	-1	1			
12									
13					suma di	93			
14									
15					r	0,44			
16									
17									
18									
19									
20									

Pogledajte pripadni dijagram raspršenosti i donesite zaključke.



Primjer 2:

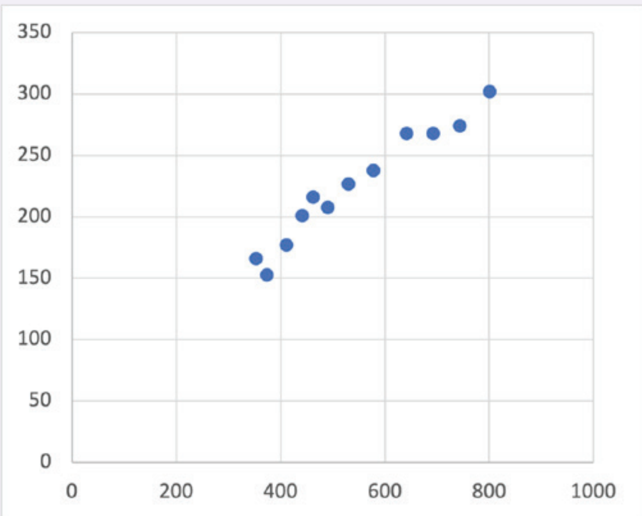
U sljedećoj tablici naći ćete sljedeće podatke: varijabla X predočava ukupno vrijeme za reklame na nacionalnoj televiziji u minutama po mjesecima tijekom jedne godine, a varijabla Y prodaju proizvoda, u tisućama komada, po mjesecima te godine. Izračunajte vrijednost Pearsonova koeficijenta korelacije. Kako glasi linearna regresijska jednadžba s procijenjenim parametrima?

x	352	373	411	441	462	490	529	577	641	692	743	801
y	166	153	177	201	216	208	227	238	268	268	274	302

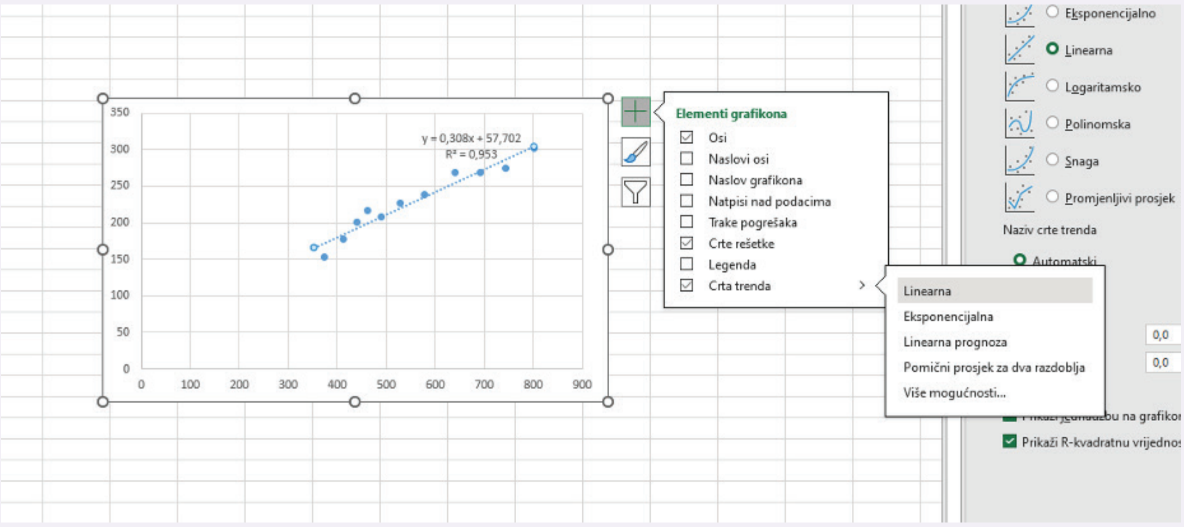
Rješenje:

Izračunajte u programu MS Excel najprije Pearsonov koeficijent. On iznosi 0.98, što nam govori o jakosti korelacije.

Dijagram raspršenosti:



U programu MS Excel lako dobivamo pravac regresije:



Provjerite svoje znanje

Riješite sljedeći zadatak:

Zadani su podaci o ulaganju u marketing i ostvarenom godišnjem profitu 15 turističkih agencija. Odredite jednadžbu modela linearne regresije koja opisuje zavisnost profita o ulaganju u marketing i objasnite značenje dobivenih parametara. Podatke prikažite grafički dijagramom raspršenosti.

Redni broj agencije	Ulaganje u marketing (000 €)	Godišnji profit (000€)
1	15	80
2	22	95
3	25	100
4	30	120
5	40	110
6	45	1445
7	50	130
8	60	180
9	70	210
10	80	200
11	95	280
12	100	320
13	120	350
14	130	375
15	150	480

Upute: Nezavisna varijabla je ulaganje, zavisna je prihod. Prikažite dijagram raspršenosti, a zatim dodajte pravac regresije. Svakako izračunajte i Pearsonov koeficijent i donesite zaključke.

5. STATISTIČKA OBRADA VELIKE KOLIČINE PODATAKA

Metoda uzoraka

Često je nemoguće ili iznimno financijski i organizacijski zahtjevno prikupiti podatke o obilježjima svih jedinica statističkog skupa. Stoga ispitujemo samo dio populacije, tj. njezin uzorak. Uzorak će biti reprezentativan ako po svojim karakteristikama nalikuje na populaciju (uzorak je umanjena slika osnovnog skupa).

Uzorkom ćemo procijeniti karakteristike osnovnog skupa, metoda uzoraka nam omogućuje donošenje odluke o prihvaćanju ili odbacivanju određene hipoteze koja se odnosi na karakteristiku neke populacije.

Obzirom na način izbora jedinica razlikujemo:

- **Namjerni** uzorak (potrebno je dobro poznavati populaciju da bi a uzorak odabrali one elemente koji će osigurati strukturu obzirom na sva relevantna obilježja populacije; primjenjuju se metode deskriptivne statistike)
- **Slučajni** uzorak (svaki član ima vjerojatnost izbora u uzorak veću od 0; podaci se analiziraju po načelima inferencijalne statistike; procjenjuju se parametri i testiraju hipoteze o njima i oblicima rasporeda osnovnih skupova; izbor se vrši pomoću tablice slučajnih brojeva ili sistematski).

Procjena aritmetičke sredine osnovnog skupa μ je parametar koji se procjenjuje:

- Brojem (npr. želimo li izračunati prosjek primanja stanovnika nekog područja, odabiremo uzorak od n stanovnika tog područa), računamo aritmetičku sredinu uzorka i zaključujemo da je taj broj istovjetan prosjeku primanja cijelog područja)
- Intervalom (formiramo interval određene širine ovisno o pouzdanosti pri čemu pazimo na z_I – koeficijent pouzdanosti provjere)

Najčešće se formiraju intervali procjene s 95% - tnom pouzdanosti, u tom slučaju koeficijent pouzdanosti iznosi 1.96 (očitavamo iz tablice površine ispod normalne krivulje, pri čemu uzimamo 0.95:2=0.4750)

Interval procjene glasi: $P\left\{\underline{x} - z_{\frac{\gamma}{2}}\sigma_{\underline{x}} \leq \mu \leq \underline{x} + z_{\frac{\gamma}{2}}\sigma_{\underline{x}}\right\}$

Središnja točka intervala je aritmetička sredina uzorka $\underline{x}$ oko koje gradimo interval u kojem se nalazi aritmetička sredina intervala μ.

P – pouzdanost

γ – vjerojatnost pogreške u procjeni aritmetičke sredine populacije

$(1-\gamma)$ – pouzdanost intervalne procjene

Osim koeficijenta pouzdanosti, treba izračunati i standardnu pogrešku aritmetičke sredine. Simboli koje pritom koristimo:

$\sigma_{\underline{x}}$ – standardna pogreška procjene aritmetičke sredine populacije

σ – standardna devijacija populacije

s - standardna devijacija uzorka

$\hat{\sigma}$ – standardna devijacija populacije procijenjena pomoću uzorka

N – opseg populacije

n – opseg uzorka

f – frakcija uzorka – odnos veličine uzorka i populacije $c = n/N$.

U ovisnosti o poznavanju standardne devijacije populacije i frakcije uzorka određuje se izraz za standardnu pogrešku aritmetičke sredine populacije na sljedeći način:

Izraz za standardnu pogrešku aritmetičke sredine populacije	Uvjeti za primjenu izraza
$\sigma_{\underline{x}} = \frac{\sigma}{\sqrt{n}}$	σ poznata i $f < 0.05$
$\sigma_{\underline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	σ poznata i $f \geq 0.05$
$\sigma_{\underline{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$	σ nije poznata i $f < 0.05$
$\sigma_{\underline{x}} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	σ nije poznata i $f \geq 0.05$
$\sigma_{\underline{x}} = \frac{s}{\sqrt{n-1}}$	σ nije poznata i $f < 0.05$
$\sigma_{\underline{x}} = \frac{s}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$	σ nije poznata i $f \geq 0.05$

Za uzorke **veće od 30 jedinica** koeficijent pouzdanosti z uzimamo iz tablice površina ispod normalne razdiobe:

- z = 1.96 (uz 95% pouzdanosti procjene)
- z = 2.58 (uz 99% pouzdanosti procjene)

Za uzorke **manje od 30 jedinica** koeficijent pouzdanosti z ne možemo uzeti iz tablice površina ispod normalne razdiobe već ga uzimamo iz Studentove tablice za $k = n-1$ stupnjeva slobode, uz željenu vjerojatnost procjene (najčešće 0.95 ili 0.99).

Primjer 1:

a) Procijenimo prosječno vrijeme obrade korisnika banke u poslovnici, ako je slučajni uzorak veličine n = 64 uz razinu pouzdanosti 95°, ako je zabilježeno vrijeme u minutama preko terminala banke dano u tablici uz frakciju izbora manju od 5%:

Trajanje obrade							
16,10	12,79	11,85	9,37	7,26	10,12	12,53	9,93
9,23	8,07	10,96	12,69	7,91	2,00	9,87	11,88
11,81	9,88	10,86	12,93	12,54	12,70	7,18	9,57
16,20	7,80	4,90	7,44	8,05	8,10	9,93	5,60
2,00	9,64	10,36	12,25	11,89	11,09	2,50	11,72
4,00	13,15	9,72	15,50	9,88	13,55	10,67	10,84
8,44	10,13	8,65	9,70	10,42	10,98	11,09	8,03
5,30	4,20	12,81	8,90	9,08	11,04	6,26	9,54

b) U kojim se granicama može očekivati prosječno vrijeme obrade naloga?

Slika 5. Razgovor u poslovnici

Rješenje:

Uzorak je velik (n>30) i izabran uz frakciju f < 0.05. Nije poznat oblik distribucije u osnovnom skupu, aritmetička sredina niti standardna devijacija. Tako da:

a) Procjenitelj aritmetičke sredine osnovnog skupa je aritmetička sredina uzorka:

$$\underline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{64}\left[16.10 + 9{,}23 + \ldots + 4{,}20\right] = 9{,}70906$$

Procjena prosječnog trajanja svih obrada iznosi 9.71 minutu.

b) Intervalni procjenitelj s razinom pouzdanosti 100(1 $-\gamma$)% u općem je obliku:

$$P\left\{\underline{x} - z_{\frac{\gamma}{2}}\sigma_{\underline{x}} \le \mu \le \underline{x} + z_{\frac{\gamma}{2}}\sigma_{\underline{x}}\right\}.$$

Razina pouzdanosti je 95%, odnosno 0.95, pa je vrijednost koeficijenta prema tablicama površina ispod normalne krivulje $z_{0.025}$ = 1.96

Računamo standardnu pogrešku procjene aritmetičke sredine, ali prvo treba izračunati vrijednost procjene standardne devijacije osnovnog skupa pomoću uzorka:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}} = \sqrt{\frac{16.10^2 + \ldots + 9.54^2 - \frac{621.38^2}{64}}{63}} = 3.06977$$

Kako je frakcija izbora $f < 0.05$, standardna devijacija nepoznata, to je:

$$\sigma_{\underline{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{3.06977}{\sqrt{64}} = 0.38372$$

Vrijednosti granice procjene su:

$$\underline{x} - z_{\frac{\gamma}{2}}\sigma_{\underline{x}} = 9.70906 - 1.96 \cdot 0.38372 = 8.95697$$

$$\underline{x}+z_{\frac{\gamma}{2}}\sigma_{\underline{x}}=9.70906+1.96\cdot 0.38372=10.46116$$

Dakle, 95% tni interval pouzdanosti procjene aritmetičke sredine osnovnog skupa glasi:

$$P\left\{\underline{x}-z_{\frac{\gamma}{2}}\sigma_{\underline{x}}\leq\mu\leq\underline{x}+z_{\frac{\gamma}{2}}\sigma_{\underline{x}}\right\};\ P\{8.95697\leq\mu\leq 10.46116\}.$$

Tumačimo ovako: s vjerojatnošću od 0.95 očekuje se da je prosječno vrijeme obrade naloga svih korisnika poslovnice banke između 8.96 i 10.46 minuta.

Testiranje hipoteza o parametru

U prethodnom poglavlju uvidjeli smo na koji način možemo iz slučajno odabranog dijela populacije procijeniti određenu karakteristiku populacije s mogućnošću kontrole vjerojatnosti greške u procjeni.

Drugi pristup omogućuje pretpostavku da neka karakteristika populacije, koja nam je nepoznata, ima određenu numeričku vrijednost. Na osnovi teorije ili iskustva postavljaju se pretpostavke ili hipoteze.

Svaki postupak polazi od nulte hipoteze i alternativne hipoteze. Sadržaj hipoteza određuje istraživač, a sadržaj alternativne hipoteze uvijek proturječi sadržaju nulte hipoteze.

U postupku odlučivanja vrši se provjera istinitosti hipoteze na sljedeći način:

- izabire se uzorak iz populacije i izračuna se karakteristika uzorka
- izračunata karakteristika uspoređuje se s pretpostavljenom karakteristikom populacije
- ako razlika nije velika može se smatrati slučajnom, a hipoteza mogućom
- ako je razlika prevelika, te se ne bi mogla smatrati slučajnom, hipoteza se odbacuje kao neprihvatljiva

Cilj testiranja je odbacivanje što više lažnih i što manje istinitih hipoteza.

U postupku odlučivanja pojavljuju se dvije vrste pogrešaka:

- Pogreška tipa I – učini se kad se odbaci istinita nulta hipoteza
- Pogreška tipa II – učini se kad se prihvati lažna nulta hipoteza

Postupak testiranja i njegovi ishodi nalaze se u sljedećoj tablici:

Odluka	**Nulta hipoteza je**	
	istinita	**lažna**
Prihvatiti nultu hipotezu	odluka ispravna	pogreška tipa II
Odbaciti nultu hipotezu	pogreška tipa I.	odluka ispravna

Postupak odlučivanja počiva na svojstvima sampling-distribucije test-veličine, odnosno na sadržaju i specifikaciji nulte hipoteze.

Kako je svaka sampling-distribucija distribucija vjerojatnosti, pogreška tipa I prikazuje se vjerojatnošću odbacivanja istinite nulte hipoteze α. Još se naziva se i razinom značajnosti (razinom signifikantnosti). β je vjerojatnost da se prihvati lažna nulta hipoteza, odnosno da se napravi pogreška tipa II. Vjerojatnost odbacivanja lažne nulte hipoteze naziva se snagom statističkog testa. Ta je vjerojatnost jednaka (1-β).

Testiranje hipoteze o aritmetičkoj sredini populacije

Testiranje se provodi pomoću slučajnog uzorka veličine n:

- Za n > 30: veliki uzorak (z-test)
- Za n < 30: mali uzorak (t-test)

Koraci testiranja:

1) određivanje sadržaja nulte i alternativne hipoteze,
2) određivanje izraza za testnu veličinu i izračunavanje njezine vrijednosti,
3) odabir razine značajnosti i određivanje kritičnih granica koje dijele područje prihvaćanja nulte hipoteze od područja njezina odbacivanja,
4) donošenje zaključka o ishodu testa

Veliki uzorak:

- ako je nulta hipoteza istinita sampling-distribucija sredina uzoraka oblika je normalne distribucije ili približno tog oblika;
- oznake: μ = nepoznata aritmetička sredina populacije,
  μ_0 = pretpostavljena aritmetička sredina populacije

VRSTA TESTA	Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hipoteze	Područje odbacivanja nulte hipoteze
dvosmjeran	$H_0 \ldots \mu = \mu_0$	$H_1 \ldots \mu \neq \mu_0$	$\lvert z \rvert < z_{\frac{\alpha}{2}}$	$\lvert z \rvert > z_{\frac{\alpha}{2}}$
jednosmjeran, na gornju granicu	$H_0 \ldots \mu \leq \mu_0$	$H_1 \ldots \mu > \mu_0$	$z < z_\alpha$	$z > z_\alpha$
jednosmjeran, na donju granicu	$H_0 \ldots \mu \geq \mu_0$	$H_1 \ldots \mu < \mu_0$	$z > -z_\alpha$	$z < -z_\alpha$

Izraz nulta hipoteza skraćeno pišemo H_0, a izraz alternativna hipoteza H_1. Empirijski z – omjer, odnosno test – veličina dana je formulom: $z = \frac{\underline{x} - \mu_0}{\sigma_{\underline{x}}}$.

Oznake:

$\sigma_{\underline{x}}$ – standardna pogreška

$\underline{x}$ – aritmetička sredina uzorka

μ_0 – pretpostavljena veličina aritmetičke sredine osnovnog skupa

Odluka se donosi alternativno pomoću kritičnih granica, a za dvosmjeran test kritične granice prihvaćanja nulte hipoteze su: $c_1 = \mu_0 - z_{\frac{\alpha}{2}} \cdot \sigma_{\underline{x}}, c_2 = \mu_0 + z_{\frac{\alpha}{2}} \cdot \sigma_{\underline{x}}$

Kad je riječ o jednosmjernom testu na gornju granicu, kritična je granica $c_2 = \mu_0 + z_\alpha \cdot \sigma_{\underline{x}}$. Odluka u jednosmjernom testu na donju granicu donosi se pomoću $c_1 = \mu_0 - z_\alpha \cdot \sigma_{\underline{x}}$.

Aritmetička sredina uzorka veća od donje granice upućuje na prihvaćanje nulte hipoteze, a vrijednost sredine manja od donje kritične granice na njeno odbacivanje.

Testiranje hipoteze o pretpostavljenoj aritmetičkoj sredini populacije koja je normalno raspoređena s **nepoznatom standardnom devijacijom** temelji se na Studentovoj distribuciji kao sampling-distribuciji sredina, polazi se od nulte hipoteze kao istinite i koristi se t-test, pri čemu je: $t = \frac{\underline{x} - \mu_0}{\sigma_{\underline{x}}}$.

Ako je nulta hipoteza istinita, test – veličina pripada t-distribuciji (Studentovoj distribuciji) s (n-1) stupnjem slobode. Oblici hipoteza tada su:

VRSTA TESTA	Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hipoteze	Područje odbacivanja nulte hipoteze
dvosmjeran	$H_0 \ldots \mu = \mu_0$	$H_1 \ldots \mu \neq \mu_0$	$\lvert t \rvert < t_{\frac{\alpha}{2}}$	$\lvert t \rvert > t_{\frac{\alpha}{2}}$
jednosmjeran, na gornju granicu	$H_0 \ldots \mu \leq \mu_0$	$H_1 \ldots \mu > \mu_0$	$t < t_\alpha$	$t > t_\alpha$
jednosmjeran, na donju granicu	$H_0 \ldots \mu \geq \mu_0$	$H_1 \ldots \mu < \mu_0$	$t > -t_\alpha$	$t < -t_\alpha$

Primjer 1.

Radi povećanja produktivnosti rada strojeva tipa ABC, predložena je njihova preinaka. Prema proračunima, preinaka je poslovno opravdana ako se postigne povećan broj operacija po satu i ako u prosjeku iznosi više od 120. Na jednom je stroju provedena preinaka i evidentiran je broj operacija po satu 144. Prosječan broj operacija po satu u provedenom je ispitivanju iznosio 125. Zbroj kvadrata vrijednosti mjerenja iznosi 2307600. Do kojeg se zaključka dolazi na temelju provedenog ispitivanja? Vjerojatnost odbacivanja istinite nulte hipoteze iznosi 5%.

Slika 6. Modificiran stroj

Rješenje:

Test o povećanom učinku modificiranog stroja provodi se na temelju velikog uzorka izabranog iz beskonačnog skupa. Promatramo odstupanje naviše u odnosu prema pretpostavljenom prosječnom broju operacija po satu koji je 120. Zaključujemo da je riječ o jednosmjernom testu na gornju granicu. Provodimo međurezultate:

$$n=144, \underline{x}=125, \hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\underline{x}^2}{n-1}}=\sqrt{\frac{2307600-144\cdot 125^2}{143}}=20.06981$$

$$\sigma_{\underline{x}}=\frac{\hat{\sigma}}{\sqrt{n}}=1.67248$$

Hipoteze su:

$$H_0 \ldots \mu \leq 120;\ H_1 \ldots \mu > 120$$

Test veličina je $z=\frac{\underline{x}-\mu_0}{\sigma_{\underline{x}}}=\frac{125-120}{1.67248}=2.99.$

Razina značajnosti je α = 0.05, kako je test jednosmjeran i uzorak velik, područje prihvaćanja nulte hipoteze je:

$$z<z_{\alpha}, z<1.65, \text{ jer je } \alpha=0.05, z_{0.05}=1.65$$

Odluka jest: empirijski z – omjer veći je od kritične vrijednosti, 2.99>1.65. Odbacuje se nulta hipoteza na danoj razini značajnosti.

Alternativno, do zaključka se dolazi pomoću kritične granice:

$$c_2=\mu_0+z_{\frac{\alpha}{2}}\cdot\sigma_{\underline{x}}=120+1.65\cdot 1.67248=122.76$$

Aritmetička sredina uzorka je 125 i veća je od gornje granice pa se ne prihvaća nulta hipoteza. Odstupanje aritmetičke sredine uzorka naviše značajno je pa se prihvaća pretpostavka da je preinaka strojeva gospodarski opravdana.

ZAKLJUČAK

Statističke metode znanstvene su metode koje se bave prikupljanjem, analizom i obradom podataka različite vrste, stoga je široko primjenjiva u različitim znanostima, poglavito pri obradi rezultata prikupljenim istraživanjem ili eksperimentom, a potom i radi zaključivanja iz konkretnih slučajeva na općenite zaključke.

Danas se općenito u statističkim istraživanjima i zaključivanju uvelike oslanjamo na niz računalnih programa, od poznatog nam MS Excela koji nudi pregršt funkcija, do sofisticiranijih, kao što su SAS Studio, JASP, Statistica i slični.

Dodatno istražite teorijsku distribuciju vjerojatnosti procjenitelja parametra, tzv. Sampling distribuciju aritmetičkih sredina, proporcija i varijanci.

IMPRESUM

Autor:
Vladimira Majdandžić

Nakladnik:
Agencija za strukovno obrazovanje i obrazovanje odraslih

Izvoditelj:
Profil Klett d.o.o.

Recenzent:
Zlatko Damijanić

Urednik:
Vedrana Gregurić

Lektor:
Zdravka Krpina

Grafička priprema:
Aria dizajn, Zagreb

Stručnjak za didaktičko-metodičko oblikovanje:
Matea Thompson

Voditelj projekta:
Zvonimir Stanić

Izvori fotografija:
Arhiva autora, Shutterstock

Literatura:

1. Boris Petz: Osnovne statističke metode za nematematičare, Naklada slap, Jastrebarsko, 2002.
2. Ivan Šošić: Primijenjena statistika, Školska knjiga, Zagreb,2006.
3. Milan Papić: Primijenjena statistika u MS Excelu, Zoro d.o.o. Zagreb, 2008.
4. Mr.sc. Bojan Kovačić: Poslovna statistika URL: https://bkovacic.weebly.com/uploads/7/4/0/7/7407552/skripta.pdf (pristupljeno 18.7.2023.)

Sadržaj ove publikacije isključiva je odgovornost Agencije za strukovno obrazovanje i obrazovanje odraslih.

Projekt je sufinancirala Europska unija iz Europskog socijalnog fonda.

Za više informacija o EU fondovima posjetite web stranicu Ministarstva regionalnoga razvoja i fondova Europske unije: www.strukturnifondovi.hr.

Agencija za
strukovno obrazovanje
i obrazovanje odraslih

MODERNIZACIJA SUSTAVA STRUKOVNOG OBRAZOVANJA I OSPOSOBLJAVANJA

Europska unija
"Zajedno do fondova EU"

EUROPSKI STRUKTURNI
I INVESTICIJSKI FONDOVI

ESF UČINKOVITI LJUDSKI POTENCIJALI

Projekt je sufinancirala Europska unija iz Europskog socijalnog fonda.