

Grafička prezentacija podataka

- **Osnovni oblici grafičke prezentacije:** Numeričko i grafičko sumiranje podataka.
- **Vizuelna komunikacija:** Grafički prikazi često bolji za komunikaciju.

Tipovi grafičkih prikaza

- **Stupičasti grafikoni (Bar plots):** Prikaz frekvencije na y osi i kategorije na x osi.
- **Pite (Pie charts):** Prikaz udjela svake vrijednosti kategorijske promjenljive u odnosu na cjelinu.
- **Histogrami (Histograms):** Prikaz distribucije kvantitativnih podataka.
- **Vremenski grafikoni (Timeplots):** Prikaz podataka kroz vrijeme.
- **Stemplots:** Prikaz svake pojedinačne vrijednosti, slični histogramima.

Kategorijski podaci

- **Opis:** Kategorijski podaci su podaci koji se odnose na kategorije ili grupe. Oni ne predstavljaju numeričke vrijednosti, već različite vrste ili klase.
- **Primjeri:** Boje automobila, vrste voća, spol, tipovi proizvoda.
- **Grafički prikazi:**
 - Stupičasti grafikoni (Bar plots): Koriste se za prikazivanje frekvencije ili broja pojava svake kategorije. Na y osi se prikazuje frekvencija, a na x osi kategorije. Na primjer, broj prodanih automobila po boji.
 - Pite (Pie charts): Prikazuju udio svake kategorije u odnosu na cjelinu. Svaki segment pite predstavlja procentualni udio određene kategorije. Na primjer, udio tržišta različitih brendova.

Kvantitativni podaci

- **Opis:** Kvantitativni podaci su numerički podaci koji predstavljaju količine ili mjere. Oni se mogu zbrajati, oduzimati i koristiti u matematičkim operacijama.
- **Primjeri:** Visina, težina, starost, prihodi.
- **Grafički prikazi:**
 - Histogrami (Histograms): Koriste se za prikazivanje distribucije kvantitativnih podataka. Na y osi se prikazuje frekvencija, a na x osi intervali numeričkih vrijednosti. Na primjer, raspodjela dobi u populaciji.
 - Vremenski grafikoni / vremenske serije (Timeplots): Prikazuju promjene kvantitativnih podataka kroz vrijeme. Vrijeme se prikazuje na x osi, a vrijednosti podataka na y osi. Na primjer, promjena temperature tokom godine.
 - Stemplots: Prikazuju svaku pojedinačnu vrijednost kvantitativnih podataka. Podaci se dijele na "stabljike" i "listove", gdje stabljike predstavljaju glavne

vrijednosti, a listovi pojedinačne podatke. Na primjer, ocjene studenata na ispitu.

Histogrami

- Distribucija frekvencije: Prikaz frekvencije na y osi i intervala na x osi.
- Ključni pojmovi: Centar, varijabilnost, oblik histograma, izuzeci (outliers).

Transformacija distribucije frekvencije

Relativna distribucija frekvencije

- Opis: Relativna distribucija frekvencije prikazuje udio slučajeva u ukupnom broju slučajeva, umjesto apsolutnog broja slučajeva.
- Izračunavanje: Da bismo dobili relativnu frekvenciju, dijelimo frekvenciju svake kategorije ili intervala s ukupnim brojem slučajeva.
- Primjer: Ako imamo 100 studenata i 20 od njih je dobilo ocjenu "A", relativna frekvencija za ocjenu "A" je 0.2 (ili 20%).
- Primjena: Relativna distribucija frekvencije je korisna za usporedbu različitih skupova podataka jer omogućava prikaz podataka u odnosu na cjelinu, bez obzira na veličinu uzorka.

Transformacija u %

- Opis: Histogram može biti transformisan tako da prikazuje proporciju učešća u pojedinačnom intervalu u postocima.
- Izračunavanje: Relativna frekvencija se množi sa 100 da bi se dobio postotak.
- Primjer: Ako je relativna frekvencija za određeni interval 0.2, to znači da taj interval čini 20% ukupnog broja slučajeva.
- Primjena: Transformacija histograma u postotke omogućava lakše razumijevanje i usporedbu podataka, posebno kada se radi o velikim skupovima podataka. Na primjer, možemo vidjeti koliki postotak populacije spada u određeni raspon dobi.

Oblik distribucije

Oblik distribucije podataka nam pomaže da razumijemo kako su podaci raspoređeni. Distribucija može biti simetrična ili iskošena (asimetrična).

Simetrija/iskošenost

Simetrična distribucija

- **Opis:** Simetrična distribucija je ona koja se može podijeliti na dva jednaka dijela istog oblika. To znači da su lijeva i desna strana distribucije ogledalo jedna drugoj.
- **Primjer:** Distribucija visine odraslih osoba može biti približno simetrična, gdje većina ljudi ima visinu oko prosjeka, a manje ljudi ima ekstremne visine.

Iskošenost (Skewed)

- **Opis:** Iskošenost se odnosi na asimetriju u distribuciji podataka. Distribucija može biti iskošena u lijevo ili desno.
- **Iskošenost u lijevo:** Lijeva strana histograma proteže se dalje od desne. To znači da postoji veći broj nižih vrijednosti, dok su veće vrijednosti rijetke.
 - **Primjer:** Distribucija ocjena na testu gdje većina učenika dobije visoke ocjene, a samo nekoliko njih dobije niske ocjene.
- **Iskošenost u desno:** Desna strana histograma proteže se dalje od lijeve. To znači da postoji veći broj viših vrijednosti, dok su niže vrijednosti rijetke.
 - **Primjer:** Distribucija prihoda gdje većina ljudi ima umjerene prihode, a samo nekoliko njih ima vrlo visoke prihode.

Oblik distribucije kroz Boxplot

Elementi Boxplota

- **Najveća vrijednost (whiskers):** Gornji "brk" (whisker) predstavlja najveću vrijednost podataka koja nije izuzetno udaljena od ostalih podataka (nije outlier).
- **Najmanja vrijednost (whiskers):** Donji "brk" (whisker) predstavlja najmanju vrijednost podataka koja nije izuzetno udaljena od ostalih podataka (nije outlier).
- **Medijana:** Linija unutar kutije (box) koja dijeli podatke na dva jednaka dijela. Predstavlja srednju vrijednost skupa podataka.
- **Prvi kvartil (Q1):** Donja granica kutije. Predstavlja 25. percentil podataka, odnosno vrijednost ispod koje se nalazi 25% podataka.
- **Treći kvartil (Q3):** Gornja granica kutije. Predstavlja 75. percentil podataka, odnosno vrijednost ispod koje se nalazi 75% podataka.

Informacije koje pruža Boxplot

- **Manje informacija nego histogram:** Boxplot ne prikazuje detaljnu distribuciju podataka kao histogram, ali pruža ključne statističke informacije.

- **Grafička usporedba različitih distribucija:** Boxplot omogućava jednostavnu usporedbu različitih distribucija podataka. Više boxplotova može se prikazati jedan pored drugog za usporedbu različitih skupova podataka.

Iskošenost (Skewness)

- **Veći dio u box-u određuje iskošenost:** Ako je jedan dio kutije (box) veći od drugog, to ukazuje na iskošenost distribucije.
 - **Iskošenost udesno:** Ako je desni dio kutije veći, distribucija je iskošena udesno (veće vrijednosti su češće).
 - **Iskošenost ulijevo:** Ako je lijevi dio kutije veći, distribucija je iskošena ulijevo (manje vrijednosti su češće).
- **Whiskers:** Ako su dijelovi kutije jednaki, gledamo whiskers. Duži whisker na jednoj strani ukazuje na iskošenost u tom smjeru.
 - **Simetrična distribucija:** Ako su dijelovi kutije i whiskers jednaki, distribucija je simetrična.

Važnost distribucije

Informacija o promjenljivoj

- **Opis:** Ukupan oblik distribucije pruža važne informacije o promjenljivoj. Distribucija nam govori koje vrijednosti promjenljiva uzima i koliko često uzima svaku vrijednost.
- **Primjena:** Razumijevanje distribucije pomaže u identifikaciji obrazaca, izuzetaka i općeg rasporeda podataka. To je ključno za analizu podataka i donošenje informiranih odluka.

Primjeri

- **Simetrične distribucije:** Veličine živih bića iste vrste (npr. dužina cvrčaka) obično su simetrične. To znači da su podaci ravnomjerno raspoređeni oko srednje vrijednosti.
- **Iskošene distribucije:** Podaci o prihodima (bilo da se radi o pojedincima, kompanijama ili nacijama) obično su nagnuti udesno. Postoji mnogo umjerenih prihoda, nekoliko velikih prihoda i nekoliko vrlo velikih prihoda.

Prikaz distribucije

- **Histogram:** Koristimo histogram za prikaz distribucije kvantitativne promjenljive. Histogram prikazuje frekvenciju podataka u određenim intervalima.
- **Dijagram:** Dijagrami poput boxplota također se koriste za prikaz distribucije kvantitativnih podataka, pružajući ključne statističke informacije.

Prezentacija distribucije podataka pomoću brojeva

Deskriptivna statistika

Deskriptivna statistika se koristi za opisivanje i sumiranje osnovnih karakteristika skupa podataka. Ona uključuje mjere centralne tendencije i mjere disperzije.

Mjere centralne tendencije

Mjere centralne tendencije su statističke mjere koje opisuju središnju tačku skupa podataka. Tri glavne mjere centralne tendencije su medijana, modus i prosječna vrijednost.

Medijana

- **Opis:** Medijana je srednja tačka distribucije. To je vrijednost koja dijeli skup podataka na dva jednaka dijela, gdje je polovina posmatranja manja od medijane, a druga polovina veća.
- **Izračunavanje:**
 - **Neparan broj posmatranja:** Ako je broj posmatranja neparan, medijana je srednja vrijednost u uređenoj listi.
 - **Paran broj posmatranja:** Ako je broj posmatranja paran, medijana je prosjek dvije srednje vrijednosti u uređenoj listi.
- **Primjer:** Ako imamo skup podataka [3, 5, 7, 9, 11], medijana je 7. Ako imamo skup podataka [3, 5, 7, 9], medijana je $(5+7)/2 = 6$.

Modus

- **Opis:** Modus je najčešće pojavljujuća vrijednost u skupu podataka. Skup podataka može imati jedan modus (unimodalna distribucija), više modusa (multimodalna distribucija) ili nijedan modus (ako se nijedna vrijednost ne ponavlja).
- **Primjer:** U skupu podataka [2, 4, 4, 6, 8], modus je 4 jer se pojavljuje najviše puta.

Prosječna vrijednost (Aritmetička sredina)

- **Opis:** Prosječna vrijednost je tačka ravnoteže podataka, izračunata kao aritmetička sredina svih vrijednosti u skupu podataka.
- **Izračunavanje:** Prosječna vrijednost se izračunava tako što se zbroje sve vrijednosti u skupu podataka i podijele s brojem posmatranja.

$$\text{Prosječna vrijednost} = \frac{\sum \text{vrijednosti}}{\text{broj posmatranja}}$$

- **Primjer:** Ako imamo skup podataka [3, 5, 7, 9, 11], prosječna vrijednost je $(3+5+7+9+11)/5 = 7$.

Kvartili

Prvi kvartil (Q1)

- **Opis:** Prvi kvartil (Q1) je medijana zapažanja koja se nalaze lijevo od ukupne medijane u uređenoj listi podataka.
- **Izračunavanje:**
 1. Rasporedite sva zapažanja po veličini, od najmanjeg do najvećeg.
 2. Pronađite ukupnu medijanu (M) skupa podataka.

3. Prvi kvartil (Q1) je medijana zapažanja koja se nalaze lijevo od ukupne medijane. Ako je broj zapažanja neparan, ukupna medijana se ne uključuje u izračunavanje Q1.
- **Primjer:** U skupu podataka [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ukupna medijana je 5.5. Prvi kvartil (Q1) je medijana zapažanja [1, 2, 3, 4, 5], što je 3.

Treći kvartil (Q3)

- **Opis:** Treći kvartil (Q3) je medijana zapažanja koja se nalaze desno od ukupne medijane u uređenoj listi podataka.
- **Izračunavanje:**
 1. Rasporedite sva zapažanja po veličini, od najmanjeg do najvećeg.
 2. Pronađite ukupnu medijanu (M) skupa podataka.
 3. Treći kvartil (Q3) je medijana zapažanja koja se nalaze desno od ukupne medijane. Ako je broj zapažanja neparan, ukupna medijana se ne uključuje u izračunavanje Q3.
- **Primjer:** U skupu podataka [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ukupna medijana je 5.5. Treći kvartil (Q3) je medijana zapažanja [6, 7, 8, 9, 10], što je 8.

Rezime distribucije od pet brojeva

Rezime distribucije od pet brojeva sastoji se od pet ključnih statističkih mjera koje pružaju osnovne informacije o raspodjeli podataka. Ove mjere su:

1. **Minimum:** Najmanja vrijednost u skupu podataka.
2. **Prvi kvartil (Q1):** Medijana zapažanja lijevo od ukupne medijane.
3. **Medijana (M):** Srednja tačka distribucije, gdje je polovina posmatranja manja, a druga polovina veća.
4. **Treći kvartil (Q3):** Medijana zapažanja desno od ukupne medijane.
5. **Maksimum:** Najveća vrijednost u skupu podataka.

Primjer rezimea distribucije od pet brojeva

U skupu podataka [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], rezime distribucije od pet brojeva bi bio:

- **Minimum:** 1
- **Prvi kvartil (Q1):** 3
- **Medijana (M):** 5.5
- **Treći kvartil (Q3):** 8
- **Maksimum:** 10

Mjere disperzije

Mjere disperzije su statističke mjere koje opisuju raspršenost ili varijabilnost skupa podataka. One nam pomažu da razumijemo koliko su podaci raspršeni oko srednje vrijednosti. Tri glavne mjere disperzije su raspon, standardna devijacija i varijansa.

Raspon

- **Opis:** Raspon je najjednostavnija mjera disperzije. On predstavlja razliku između najveće i najmanje vrijednosti u skupu podataka.
- **Izračunavanje:**

$$\text{Raspon} = \text{Maksimum} - \text{Minimum}$$

- **Primjer:** Ako imamo skup podataka [3, 7, 8, 15, 20], raspon je $20 - 3 = 17$.
- **Primjena:** Raspon daje osnovnu informaciju o širini distribucije podataka, ali ne uzima u obzir sve vrijednosti u skupu podataka.

Standardna devijacija

- **Opis:** Standardna devijacija mjeri prosječnu udaljenost posmatranja od njihove srednje vrijednosti. Ona pokazuje koliko su podaci raspršeni oko prosječne vrijednosti.
- **Izračunavanje:**
 1. Pronađite srednju vrijednost (aritmetičku sredinu) skupa podataka.
 2. Izračunajte razliku između svake vrijednosti i srednje vrijednosti, a zatim kvadrirajte te razlike.
 3. Izračunajte prosjek kvadrata tih razlika (varijansa).
 4. Standardna devijacija je kvadratni korijen varijanse.

$$\text{Formula: } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- Gdje je x_i svaka pojedinačna vrijednost, $\bar{x}$ srednja vrijednost, a n broj posmatranja.
- **Primjer:** Ako imamo skup podataka [2, 4, 4, 4, 5, 5, 7, 9], srednja vrijednost je 5. Standardna devijacija se izračunava kao:
 - Korak 1: Srednja vrijednost = 5
 - Korak 2: Razlike od srednje vrijednosti: [-3, -1, -1, -1, 0, 0, 2, 4]
 - Korak 3: Kvadrati razlika: [9, 1, 1, 1, 0, 0, 4, 16]
 - Korak 4: Varijansa = $(9+1+1+1+0+0+4+16) / 7 = 4.57$
 - Korak 5: Standardna devijacija = $\sqrt{4.57} \approx 2.14$
- **Primjena:** Standardna devijacija je korisna za razumijevanje varijabilnosti podataka i usporedbu raspršenosti različitih skupova podataka.

Varijansa

- **Opis:** Varijansa je mjera disperzije koja predstavlja prosjek kvadrata razlika između svake vrijednosti i srednje vrijednosti. Varijansa je kvadrat standardne devijacije.
- **Izračunavanje:**

$$\text{Formula: } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Gdje je x_i svaka pojedinačna vrijednost, $\bar{x}$ srednja vrijednost, a n broj posmatranja.
- **Primjer:** Koristeći isti skup podataka [2, 4, 4, 4, 5, 5, 7, 9], varijansa se izračunava kao:
 - Korak 1: Srednja vrijednost = 5
 - Korak 2: Razlike od srednje vrijednosti: [-3, -1, -1, -1, 0, 0, 2, 4]
 - Korak 3: Kvadrati razlika: [9, 1, 1, 1, 0, 0, 4, 16]
 - Korak 4: Varijansa = $(9+1+1+1+0+0+4+16) / 7 = 4.57$
- **Primjena:** Varijansa je korisna za kvantifikaciju varijabilnosti podataka, ali zbog kvadriranja razlika, jedinice varijanse nisu iste kao jedinice originalnih podataka.

Uticaj transformacija na prosječnu vrijednost i standardnu devijaciju

Dodavanje konstantne vrijednosti

- **Opis:** Kada se svim vrijednostima u skupu podataka doda ista konstanta, cijela distribucija se pomjera za tu konstantu.
- **Uticaj na prosječnu vrijednost:** Prosječna vrijednost se povećava za tu konstantu.
 - **Primjer:** Ako imamo skup podataka [2, 4, 6] i dodamo konstantu 3, novi skup podataka je [5, 7, 9]. Prosječna vrijednost se povećava za 3.
- **Uticaj na standardnu devijaciju:** Standardna devijacija ostaje nepromijenjena jer se udaljenosti između vrijednosti ne mijenjaju.
 - **Primjer:** Standardna devijacija skupa [2, 4, 6] je ista kao i standardna devijacija skupa [5, 7, 9].

Multiplikacija konstantom

- **Opis:** Kada se sve vrijednosti u skupu podataka pomnože istom konstantom, mijenjaju se i prosječna vrijednost i standardna devijacija.
- **Uticaj na prosječnu vrijednost:** Prosječna vrijednost se množi tom konstantom.
 - **Primjer:** Ako imamo skup podataka [2, 4, 6] i pomnožimo ga konstantom 2, novi skup podataka je [4, 8, 12]. Prosječna vrijednost se množi sa 2.
- **Uticaj na standardnu devijaciju:** Standardna devijacija se također množi tom konstantom.

- **Primjer:** Standardna devijacija skupa [2, 4, 6] se množi sa 2 kada se skup podataka promijeni u [4, 8, 12].

Ekstremne vrijednosti (outliers)

Opis

- **Opis:** Ekstremne vrijednosti (outliers) su vrijednosti koje se numerički značajno razlikuju od ostatka skupa podataka. One mogu biti mnogo veće ili mnogo manje od ostalih vrijednosti.
- **Primjer:** U skupu podataka [2, 3, 4, 5, 100], vrijednost 100 je outlier.

Uticaaj

- **Uticaaj na prosječnu vrijednost:** Srednja vrijednost je pod uticajem outliera jer ekstremne vrijednosti mogu značajno povećati ili smanjiti prosječnu vrijednost.
 - **Primjer:** U skupu podataka [2, 3, 4, 5, 100], prosječna vrijednost je $(2+3+4+5+100)/5 = 22.8$, što nije reprezentativno za većinu podataka.
- **Uticaaj na standardnu devijaciju:** Outlieri također povećavaju standardnu devijaciju jer povećavaju varijabilnost podataka.

Izbor statističkog rezimea

Medijana i kvartili

- **Opis:** Medijana i kvartili su manje pogođeni ekstremnim vrijednostima i iskošenom distribucijom jer se temelje na poziciji podataka, a ne na njihovim vrijednostima.
- **Primjena:** Koriste se za opisivanje distribucija koje sadrže outliere ili su iskošene.
 - **Primjer:** U skupu podataka [2, 3, 4, 5, 100], medijana je 4, što je reprezentativnije za većinu podataka.

Aritmetička sredina

- **Opis:** Aritmetička sredina je osjetljiva na ekstremne vrijednosti i iskošene distribucije jer se temelji na svim vrijednostima u skupu podataka.
- **Primjena:** Koristi se samo za simetrične distribucije bez odstupanja.
 - **Primjer:** U skupu podataka [2, 3, 4, 5, 6], prosječna vrijednost je 4, što je reprezentativno za većinu podataka.