Poglavlje 1

Deskriptivna statistika

Primjeri u Excelu vezani za ovu cjelinu nalaze se u dokumentu *Deskriptivnastatistika.xlsx*.

Deskriptivna statistika je dio matematičke statistike koji se koristi za opisivanje i bolje razumijevanje izmjerenog (ili zadanog) skupa podataka. Upoznat ćemo se s osnovnim pojmovima deskriptivne statistike.

- **Duljina uzorka** je broj podataka.
- **Najmanji podatak**
- **Najveći podatak**
- **Raspon** je razlika najvećeg i najmanjeg podatka.
- **Aritmetička sredina** (prosjek)
- **Medijan** je srednji podatak. Pola podataka nalazi se iznad, a pola ispod medijana.
- **Prvi kvartil** (donji kvartil) je broj od kojeg je manje ili jednako 25% podataka.
- **Drugi kvartil** je broj od kojeg je manje ili jednako 50% podataka. Drugi kvartil je isto što i medijan.
- **Treći kvartil** (gornji kvartil) je broj od kojeg je manje ili jednako 75% podataka.
- **Mod** uzorka je podatak koji se u uzorku pojavljuje najviše puta.

Napomenimo da je nulti kvartil zapravo najmanji podatak, a četvrti kvartil je najveći podatak. Kvartili dijele skup podataka na četiri dijela. Analogno bi mogli podijeliti podatke i na neki drugi broj dijelova. Podjela na 100 dijelova je podjela na percentile. Onda je primjerice sedmi percentil broj od kojeg je manje ili jednako 7% podataka.

U donjoj tablici navedene su Excel formule koje se koriste za računanje navedenih pojmova. Umjesto riječi "uzorak" u desnom stupcu upisuje se raspon polja u kojima se nalaze podatci, npr. MIN(A1:A50).

POJAM	NAREDBA
duljina uzorka	COUNT(uzorak)
najmanji podatak	MIN(uzorak)
najveći podatak	MAX(uzorak)
raspon	MAX(uzorak) - MIN(uzorak)
aritmetička sredina	AVERAGE(uzorak)
medijan	MEDIAN(uzorak)
1. kvartil	QUARTILE(uzorak; 1)
3. kvartil	QUARTILE(uzorak; 3)
mod	MODE(uzorak)
7. percentil	PERCENTILE(uzorak; 0.07)

Za bolju interpretaciju podataka bitno je znati koliko su podatci raspršeni, odnosno koliko odstupaju od prosjeka. Pretpostavimo da skup podataka ima duljinu n i označimo njegovu aritmetičku sredinu s $\bar{x}$. **Suma apsolutnih odstupanja** podataka od aritmetičke sredine (SAO) definira se kao

$$\text{SAO} = \sum_{i=1}^{n} |x_i - \bar{x}| = |x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|.$$

Prosječno apsolutno odstupanje od aritmetičke sredine (PAO) dobije se tako što se suma apsolutnih odstupanja podijeli s brojem podataka,

$$\text{PAO} = \frac{\text{SAO}}{n}.$$

Umjesto da gledamo apsolutno odstupanje od prosjeka, možemo gledati kvadratno odstupanje. Prosječno kvadratno odstupanje od aritmetičke sredine naziva se **varijanca** uzorka ($(s')^2$) i definirana je sa

$$(s')^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}.$$

Standardna devijacija (s') je korijen iz varijance. **Korigirana varijanca** uzorka (s^2) je veličina

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1},$$

a **korigirana standardna devijacija** (s) je korijen iz korigirane varijance. Za ove ćemo pojmove također navesti odgovarajuće Excel formule.

POJAM	NAREDBA
prosječno apsolutno odstupanje	AVEDEV(uzorak)
suma apsolutnih odstupanja	AVEDEV(uzorak)·COUNT(uzorak)
varijanca	VARP(uzorak)
standardna devijacija	STDEVP(uzorak)
korigirana varijanca	VAR(uzorak)
korigirana standardna devijacija	STDEV(uzorak)

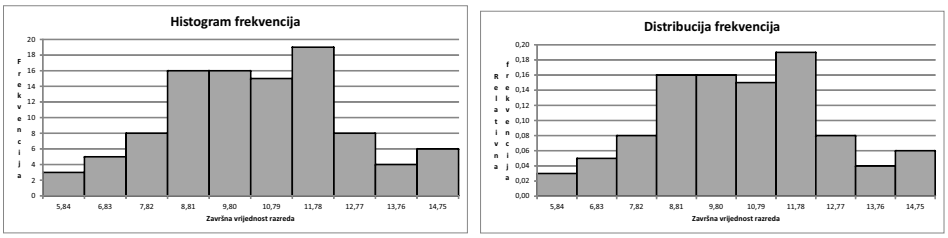
1.1 Tablica frekvencija i histogram

Zbog bolje interpretacije podataka, često je korisno podijeliti veliki skup podataka na podskupove koji se nazivaju **razredi**. Podjelu na n razreda radimo tako da interval od najmanjeg do najvećeg podatka podijelimo na n dijelova. Duljina svakog od tih n podintervala naziva se **širina razreda**, a dobije se kao kvocijent raspona i broja razreda n. (Svi su razredi jednake širine.)

Bitno je dobiti granice svakog razreda, a potom je lako rasporediti podatke po razredima. Početna vrijednost prvog razreda je najmanji podatak, a završna vrijednost jednaka je zbroju početne vrijednosti i širine razreda. Završna vrijednost prvog razreda ujedno je i početna vrijednost drugog razreda. Općenito, osim za prvi razred, početna vrijednost nekog razreda jednaka je završnoj vrijednosti prethodnog razreda. Završna vrijednost nekog razreda jednaka je zbroju njegove početne vrijednosti i širine razreda. Završna vrijednost zadnjeg razreda uvijek je jednaka najvećem podatku.

Na primjer, neka je zadan skup podataka među kojima je najmanji jednak 4, a najveći 14. Zadatak je podijeliti taj skup na 10 razreda. Tada je širina razreda jednaka $\frac{14-4}{10} = 1$. Prvi se razred proteže od 4 (što je najmanji podatak) do 5 ($4+1=5$), drugi razred od 5 do 6 ($5+1=6$), itd. Konačno, zadnji razred obuhvaća podatke od 13 do 14.

Frekvencija razreda je broj podataka u pojedinom razredu. U Excelu se ona dobije korištenjem naredbe



Slika 1.1: Graf frekvencija i graf relativnih frekvencija

FREQUENCY(uzorak; skup završnih vrijednosti razreda).

Relativna frekvencija razreda dobije se tako da se frekvencija razreda podijeli s ukupnim brojem podataka. Zbroj svih frekvencija razreda jednak je ukupnom broju podataka jer je svaki podatak ubrojen točno jednom i pripada samo jednom razredu. Zbroj svih relativnih frekvencija je 1.

RAZRED	ZAVRŠNA VRIJEDNOST	FREKVENCIJA	REL. FREKVENCIJA
1	5.84	3	0.03
2	6.83	5	0.05
3	7.82	8	0.08
4	8.81	16	0.16
5	9.80	16	0.16
6	10.79	15	0.15
7	11.78	19	0.19
8	12.77	8	0.08
9	13.76	4	0.04
10	14.75	6	0.06

Graf koji predočava frekvencije razreda naziva se **histogram frekvencija**. Crta se kao stupčasti graf gdje su na x-osi završne vrijednosti razreda, a na y-osi frekvencije. S druge strane, graf koji predočava relativne frekvencije naziva se **distribucija frekvencija**. To je također stupčasti graf, na x-osi su završne vrijednosti razreda, a na y-osi relativne frekvencije.