Vjerojatnost i statistika

Mjere centralne tendencije i raspršenosti podataka

---

19. prosinca 2023.

Mjere centralne tendencije podataka

Aritmetička sredina

Definicija 1. Aritmetička sredina (eng. mean) niza izmjerenih vrijednosti (podataka) $x_1, x_2, \ldots, x_n$ varijable X definirana je kao

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- npr. neka su $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$ izmjerene vrijednosti jedne varijable
- s obzirom da ih ima ukupno devet, aritmetička sredina ovog skupa podataka je

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42$$

Medijan

Definicija 2. Medijan niza izmjerenih vrijednosti varijable X je broj sa svojstvom da je barem pola vrijednosti veće ili jednako od njega.

- u sortiranom nizu podataka, medijan je srednji podatak po veličini
- način njegovog određivanja ovisi o tome imamo li neparan ili paran broj podataka

Medijan

- ukoliko imamo **neparan broj** izmjerenih vrijednosti, onda postoji podatak koji je na srednjoj poziciji u uređenom skupu izmjerenih vrijednosti, pa njega definiramo kao medijan
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3$ izmjerene vrijednosti jedne varijable
- najprije ih poredamo po veličini: $1, 1, 2, 2, 2, 2, 3, 5, 5, 6, 7$
- s obzirom da ih ima ukupno jedanaest, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2

Medijan

- ukoliko imamo **paran broj** izmjerenih vrijednosti, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka
- medijan se tada definira kao njihova aritmetička sredina
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable
- najprije ih poredamo po veličini: $1, 1, 2, 2, 2, 2, 3, 3, 5, 5, 6, 7$
- s obzirom da ih ima dvanaest, "sredinu" čine šesti i sedmi podatak (brojevi 2 i 3) pa je medijan 2.5

Mod podataka

Definicija 3. Mod je vrijednost iz niza izmjerenih vrijednosti varijable X kojoj pripada najveća frekvencija.

- mod ne mora biti jedinstven
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable $\Rightarrow$ vrijednost 2 je izmjerena najviše puta pa je 2 mod ovog skupa podataka
- npr. neka su $1, 2, 3, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable $\Rightarrow$ najviše puta su izmjerene dvije vrijednosti 2 i 3 pa mod ovog skupa nije jedinstven

Mjere raspršenosti podataka

Postotna vrijednost, donji i gornji kvartil

Definicija 4. Neka je $p \in (0, 100)$. **p-postotna vrijednost** x'_p niza izmjerenih vrijednosti varijable X je broj sa svojstvom da je barem $p\%$ vrijednosti manje ili jednako x'_p i barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p

- **25%-tna vrijednost** zove se **donji kvartil**
- **75%-tna vrijednost** zove se **gornji kvartil**
- **50%-tna vrijednost** zove se **medijan**

Postotna vrijednost, donji i gornji kvartil

- najprije je potrebno danih n podataka poredati u rastućem poretku
- zatim odredimo poziciju $j = np/100$
- ako j nije prirodan broj, onda je p-postotna vrijednost podatak na sljedećoj cjelobrojnoj poziciji
- ako je j prirodan broj, onda se p-postotna vrijednost računa kao aritmetička sredina podataka na pozicijama j i $j + 1$

Postotna vrijednost, donji i gornji kvartil

- npr. neka su $1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3$ izmjerene vrijednosti neke varijable
- poredamo ih po veličini: $1, 1, 2, 3, 3, 3, 3, 3, 5, 6, 6, 7$
- želimo li odrediti donji kvartil, računamo $j = 12 \cdot 25/100 = 3$
- treći podatak u gornjem skupu je broj 2, a četvrti 3 $\Rightarrow$ donji kvartil je 2.5
- analogno, deveti broj u gornjem skupu podataka je broj 5, a deseti 6 $\Rightarrow$ gornji kvartil je 5.5

Minimum i maksimum, raspon podataka

Definicija 5. Ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable X, najmanju od njih zovemo **minimum** i označavamo $x_{\min}$, a najveću od **maksimum** i označavamo $x_{\max}$. **Raspon** (eng. range) vrijednosti je razlika razlika maksimuma i minimuma vrijednosti.

Maksimalno odstupanje od prosjeka

Definicija 6. Maksimalno odstupanje od prosjeka izmjerenih vrijednosti varijable X je veći od brojeva $(\bar{x}_n - x_{\min})$ i $(x_{\max} - \bar{x}_n)$.

- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti neke varijable
- $x_{\min} = 1$, $x_{\max} = 7$, $\bar{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25$
- maksimalno odstupanje izmjerenih vrijednosti ove varijable od prosjeka je

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75$$

Varijanca i standardna devijacija podataka

Definicija 7. Varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x_n})^2,$$

a **standardna devijacija** s_n je kvadratni korijen iz varijance.

Varijanca i standardna devijacija podataka

- npr. neka su izmjerene vrijednosti jedne varijable
  $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$
- aritmetička sredina ovog skupa podataka je približno 5.42, pa varijanca i standardna devijacija iznose

$$s_n^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 = 7.87,$$

$$s_n = \sqrt{s_n^2} = 2.81$$

Kutijasti dijagram

- pet numeričkih karakteristika skupa podataka su
  - minimum
  - donji kvartil
  - medijan
  - gornji kvartil
  - maksimum
- uz njih se često navode aritmetička sredina, standardna devijacija i veličina uzorka
- za prikazivanje skupa podataka zajedno s numeričkim karakteristikama koristi se **kutijasti dijagram** (eng. boxplot, box-and-whisker plot)
- na njemu se označavaju i **stršeće vrijednosti** skupa podataka, ako postoje

Stršeće vrijednosti

Stršeća vrijednost je podatak koji je značajno veći ili manji u odnosu na druge čije je pojavljivanje najčešće vezano uz neke od sljedećih razloga:

- podatak je netočno izmjeren ili krivo unesen u bazu podataka,
- podatak dolazi iz druge populacije,
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji.

Primjer - trgovacki-centri.sta

Mjere centralne tendencije i raspršenosti dobivamo pomoću:

Statistics → Basic Statistics → Descriptive Statistics → Variables → Advanced → označiti Mean, Median, Mode, Standard deviation, Variance, Range, Minimum & maximum i Lower & upper quartiles, Range → Summary

Kutijasti dijagram dobivamo na sljedeći način:

Graphs → Box → Advanced → Graph Type: Box-Whiskers, Regular → Variables → OK

Zadatak - djelatnici.sta

a) Kojeg su tipa varijable dane baze?

b) Izradite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli *Odjel* te nacrtajte pripadne histograme.

c) Procijenite vjerojatnost da je visina djelatnika veća od 180.

d) Izradite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli *Obrazovanje* kategorizirane prema varijabli *Spol* te nacrtajte pripadne strukturirane krugove.

e) Koliki je udio ispitanika ženskoga spola kojima je *Placa_prije* veća od 20 000?

Zadatak - djelatnici.sta

f) Odredite numeričke karakteristike varijable *Placa_prije*.

g) Kolika je minimalna, a kolika maksimalna izmjerena dob ispitanika? Kolika je prosječna dob i standardna devijacija? Nacrtajte kutijasti dijagram.

h) Razvrstajte varijablu *Placa_poslije* na disjunktne intervale duljine 5000 počevši od 0, a zatim na 8 disjunktnih intervala počevši od najmanje vrijednosti.

i) Nacrtajte kutijasti dijagram varijable *Placa*_konkurencija. Interpretirajte vrijednost gornjeg kvartila.

Statistički kalkulator

- kada je poznata distribucija iz koje podaci dolaze, možemo egzaktno odrediti vjerojatnosti koje nas zanimaju
- koristimo

> Statistics $\rightarrow$ Calculators $\rightarrow$ Distributions $\rightarrow$ odabrati distribuciju $\rightarrow$ odznačiti "Fixed Scaling"

Zadatak

Šećer se strojno pakira u papirnata pakovanja nominalne mase 1kg. Pakovanje je normalno distribuirano s očekivanjem jednakim nominalnoj masi i standardnom devijacijom od 30g. Sva pakovanja čija masa odstupa od nominalne za više od 3 standardne devijacije smatraju se škartom. Proizvođač je odlučio zamijeniti postojeći stroj novim ukoliko se pokaže da je udio škarta u proizvodnji veći od 5%.

a) Kolika je vjerojatnost da će slučajno odabrano pakovanje biti proglašeno škartom? Na osnovu toga zaključite hoće li proizvođač zamijeniti postojeći stroj novim.

b) Kolika je vjerojatnost da masa pakovanja bude veća od 1050g?

Zadatak

Standardizirani IQ testovi su dizajnirani tako da rezultati ispitanika imaju normalnu distribuciju s očekivanjem 100 i standardnom devijacijom 15.

a) Genijalnošću se smatra rezultat IQ testa veći od 140. Kolika je vjerojatnost da slučajno odabrana osoba bude genij?

b) Kolika je vjerojatnost da slučajno odabrana osoba bude prosječno inteligentna, tj. ima IQ između 90 i 110?