Numeričke metode opisa podataka

Vrste numeriških opisnih mjera

- Mjere – opisuju svojstva nekog skupa podataka
  - mjere srednje vrijednosti – opisuju položaj oko kojeg se gomilaju podaci
  - mjere disperzije – mjere varijabilnost podataka
  - mjere položaja – opisuju relativni položaj nekog podatka u odnosu na ostale podatke
  - mjere asimetrije

Mjere srednje vrijednosti

- Najčešćca je aritmetička sredina, x

- Aritetička sredina skupa od n podataka x1,x2,...,xn definiše se kao

  x = Suma svih podataka / Broj podataka

- Sljedeća mjera srednje vrijednosti je medijan, M. Medijan je vrijednost sa svojstvom da je pola podataka manje ili jednako njoj, a pola podataka veće ili jednako njoj.

Primjeri

- Primjer 1. Nađite medijan za sljedeći skup podataka: 7, 4, 3, 5, 3.
  - Rješenje M = 4

- Medijan M se za skup od n podataka x1,x2,...,xn definiše na sljedeći način.
  - Za neparan n – medijan je podatak u sredini, odnosno podatak na rednom mjestu (n +1)/2
  - Za paran n – medijan je jednak aritmetičkoj sredini podataka na rednom mjestu n/2 i n/2+1 Sljede´ca mjera srednje vrijednosti je mod
  - Podaci su poređani po veličini

Mjere srednje vrijednosti 2

- Mod je vrijednost s najvećom frekvencijom. Ako su podaci grupisani po intervalima, mod se definiše kao središte intervala sa najvećom frekvencijom, a taj interval nazivamo modalnim intervalom

- Mod (za razliku od prethodno spomenutih mjera) ima smisla i za kategorijske podatke. Npr. rezultati prodaje ljetnih majica prikazani su u sljedećoj tablici. Odrediti mod.

veličina	frekvencija
S	9
M	30
L	16
XL	40
XXL	13
$\sum$	108

Kako opisati prosječnog stanovnika Crne Gore?

- Primjer 2. Prema MONSTAT-u prosječan Crnogorac je star 37 godina, zaposlen je sa platom 510 eura, banci je dužan 1900 eura, ima srednju stručnu spremu i zove se Nikola. Koje mjere srednje vrijednosti su upotrijebljene?

Primjer 3

- Dat je niz podataka. Oderditi mjere srednje vrijednosti. Ako se najve'a vrijednost zamijeni sa 255000, kako se mijenjaju mjere srednje vrijednosti?
- Za podatke koji su izrazito asimetrični, bolja procjena srednje vrijednosti može biti medijan, jer manje zavisi od ekstremnih vrijednosti
- Za simetrične podatke, medijan i aritmetička sredina imaju približno jednaku vrijednost

99000	45000	61500	78400	48500
123000	60000	155000	77000	56400
65700	50000	140000	49600	59500
115000	45500	112000	58500	25000
63000	70000	62000	46000	110000
76000	77100	61900	36500	25000
58000	45500	55000	38000	89500
87000	63500	31700	44900	90000
68000	51600	75300	40000	32000
50500	79000	47000	48000	103000

Mjere disperzije

- Mjere disperzije ili varijabilnosti posmatranog skupa podataka - koliko se podaci međusobno razlikuju
- Najjednostavnija je raspon, R
- Raspon skupa podataka se definiše kao razlika najveće i najmanje vrijednosti: $R = x_{max} - x_{min}$
- Što je raspon manji to je manje prostora unutar kojeg podaci mogu varirati
- Određen je vrijednostima samo dva podatka, te stoga ne zavisi od varijabilnosti ostalih podataka

Varijansa

- Korisnija mjera je varijansa, s^2 – mjeri odstupanje svakog podatka od aritmetičke sredine: $x_i - x_{avg}$.
- Varijansa uzorka od n podataka se računa kao prosjek kvadrata pojedinačnih odstupanja

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

- Što je varijansa veća, to je više odstupanja među podacima. Uz gornju definiciju, često ćemo koristiti i sljedeću, ekvivalentnu formulu za varijansu:

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

Dokaz

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_i \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_i x_i^2 - 2\sum_i x_i\bar{x} + \sum_i \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n-1}\sum x_i^2 - n\bar{x}^2.
\end{aligned}
$$

Primjer 4

- Odredite varijansu za sljedeći skup podataka: 3, 7, 2, 1, 8.

	x	x^2
	3	9
	7	49
	2	4
	1	1
	8	64
$\sum$	21	127

$$s^2 = \frac{\sum x_i^2 - nx^2}{n-1} = \frac{127 - 5 \cdot 4.2^2}{4} = \frac{38.8}{4} = 9.7.$$

Sa ovom formulom imamo manji broj računskih operacija (reda 2n) za razliku od originalne formule iz definicije (reda 3n)

Varijansa 2

- Ako među podacima ima više jednakih, pri čemu se vrijednost x_i pojavljuje sa frekvencijom f_i, onda iz formule (2) slijedi

$$s^2 = \frac{1}{\sum_i f_i - 1}\left(\sum_i f_i x_i^2 - n\bar{x}^2\right)$$

pri čemu se aritmetička sredina računa kao

$$\bar{x} = \frac{1}{\sum_i f_i}\sum_i f_i x_i$$

Primjer 5

- Pet novčića smo bacali 1000 puta i zabilježili broj glava. Broj bacanja u kojima je palo 0, 1, 2, 3, 4, ili 5 glava zabilježen je u sljedećoj tablici. Odredite aritmetičku sredinu i varijansu.

broj glava	broj bacanja
0	38
1	144
2	342
3	287
4	164
5	25
$\sum$	1000

Rješenje:

$$\bar{x} = 2.47, \qquad s^2 = 1.2443$$

Standardno odstupanje ili devijacija

- Varijansa
  - mjeri se u kvadratima originalnih jedinica (podatak u cm, varijansa u cm2)
  - nema jasno tumačenje
- Korisnije su mjere izražene u jedinicama jednakim originalnim
- Standardno odstupanje ili devijacija, s, uzorka od n podataka se računa kao kvadratni korijen varijanse

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

Tumačenje standardne devijacije

- Cilj – odrediti intervale u koje upada najveći dio podataka. Razmatraju se intervali oblika

  $x_{avg} \pm k*s$, k = 1,2,3

- Ukoliko imamo zvonoliku raspodjelu podataka sa aritmetičkom sredinom x_{avg}, te standardnom devijacijom s, tada je procenat podataka unutar intervala $x_{avg} \pm k*s$, k = 1,2,3 sljedeći:
  - $x_{avg} \pm s$ – obično izmedu 60 i 80%. Procenat će biti blizu 70% za simetrične raspodjele, a oko 90% za izrazito asimetrične
  - $x_{avg} \pm 2s$ – oko 95%. Procenat će biti veći (blizu 100%) za izrazito asimetrične raspodjele
  - $x_{avg} \pm 3s$ – blizu 100%.

Tumačenje standardne devijacije 2

- Čebiševljeva teorema - procenat od ukupnog broja podataka unutar intervala $x_{avg}\pm ks$, pri čemu je k konstanta, barem $1-1/k^2$
  - Za proizvoljni skup podataka s aritmetičkom sredinom x_{avg}, te standardnom devijacijom s, procenat ukupnog broja podataka unutar intervala $x_{avg}\pm 2s$ je barem 75%, $x_{avg}\pm 3s$ je barem 89%
  - Prednost – primjenljiv je na bilo koji skup podataka
  - Nedostatak – konzervativan je, u smislu da daje samo donju ocjenu stvarnog procenta

Dokaz

- Dokazali da je procenat broja podataka van intervala $x_{avg} \pm ks$ manji od $1/k^2$, te je stoga procenat od ukupnog broja podataka unutar tog intervala barem $1-1/k^2$

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum (x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|>ks\}} (x_i - \bar{x})^2 + \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|\leqslant ks\}} (x_i \\
&> \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|>ks\}} k^2 s^2 \\
&= \frac{1}{n-1} B_k\, k^2 s^2,
\end{aligned}
$$

gdje B_k označuje broj podataka sa svojstvom da je $|x - \bar{x}| >$
Stoga je

$$\frac{B_k}{n-1} < \frac{1}{k^2}.$$

Mjere položaja

- opisuju položaj podatka u odnosu na preostale podatke
- Za k $\in$[0,100] definišemo k-ti percentil, P_k, kao vrijednost sa svojstvom da je k% podataka manje ili jednako njemu, a (100−k)% podataka veće ili jednako od njega
  - 50-ti percentil – medijan, M
  - 25-ti percentil – donji kvartil, Q1
  - 75-ti percentil – gornji kvartil, Q3

- Za manje skupove podataka često je teško naći vrijednost koja premašuje, npr. tačno 25% podataka. ({4,5,8})

Procedura pronalaženja percentila

- Uredite podatke po veličini, od najmanjeg ka najvišem
- Izračunajte (n +1)/4 i zaokružite na najbližu cjelobrojnu vrijednost r (ukoliko je tačno između dva cijela broja, zaokružite na veći)
- Podatak na rednom mjestu r je donji kvartil
- Izraˇcunajte ¾*(n +1) i zaokružite na najbližu cjelobrojnu vrijednost r (zaokružite na manji ako je potrebno). Podatak na rednom mjestu r je gornji kvartil
- Za k-ti percentil izračunajte k/100*(n +1) i zaokružite na najbližu cjelobrojnu vrijednost r. Podatak na rednom mjestu r je P_k

Primjer 6

- Nađite medijan, donji i gornji kvartil, 90-ti percentil za dati skup podataka.

Q_1 : $\frac{1}{4}(n+1) = \frac{26}{4} = 6.5 \simeq 7$

Stoga je Q_1 podatak na 7. mjestu, tj.

$Q_1 = x_7 = 650.$

M: n- neparan

$\frac{n+1}{2} = \frac{26}{2} = 13$

M je podatak na 13. mjestu, tj.

$M = x_{13} = 760.$

Q_3 : $\frac{3}{4}(n+1) = 3 \cdot \frac{26}{4} = 19.5 \simeq 19$

Q_3 je podatak na 19. mjestu, tj.

$Q_3 = x_{19} = 950.$

P_{90} : $\frac{9}{10}(n+1) = 23.4 \sim 23.$

$P_{90} = x_{23} = 1120.$

660	595	1060	500	630
899	1295	749	820	843
710	950	720	575	760
1090	770	682	1016	650
425	367	1480	945	1120

z varijabla

- z varijabla (obilježje) definiše se kao količnik odstupanja od srednje vrijednosti i standardne devijacije, tj. $z_i = (x_i - x_{avg})*s$

- Negativna vrijednost znači da je podatak manji od srednje vrijednosti, a pozitivna da je veći od nje

- Ako se podatak nalazi unutar intervala $x_{avg} \pm ks$, važi

$$(x_i \in [\bar{x} - ks, \bar{x} + ks]) \iff (|z_i| \leqslant k)$$

Metode određivanja izuzetaka

- Izuzetak - podatak koji po veličini odskače od ostatka populacije

- Takve podatke nalazimo iz niza razloga
  - Podatak je pogrešno izmjeren ili zapisan
  - Podatak pripada drugoj populaciji
  - Podatak je ispravan, ali opisuje veoma rijedak događaj

- Efikasna metoda za određivanje izuzetaka je z varijabla

Primjer 7

- Cijena nekretnina u uzorku ima x_{avg} = 1064050, i standardu devijaciju s = 854414. Jedan podatak iznosi 11460000. Je li ovo izuzetak?
- Rješenje
  - Izračunajmo vrijednost z varijable z_i = (11460000−1064050)/854414 = 12.17
  - Čebiševljeva teorema kaže da bi skoro svi podaci trebali upasti u interval x_{avg}±3s, dakle njihovo z obilježje mora po apsolutnoj vrijednosti biti manje od 3
  - S obzirom da je vrijednost z varijable od 12.17 malo vjerovatna, riječ je o izuzetku

Dijagram pravougaonika

660	595	1060	500	630
899	1295	749	820	843
710	950	720	575	760
1090	770	682	1016	650
425	367	1480	945	1120

- Drugi način određivanja izuzetaka
- Zasniva se na interkvartilnom rasponu, IQ = Q3 −Q1
- Primjer 8. Napravite dijagram pravougaonika (box-plot) za date podatke.

Dijagram pravougaonika 2

- Q1 = 650, Q3 = 950, IQ = 950−650 = 300
- Nacrtajte realnu osu, na njoj naznačite kvartile
- Nacrtajte pravougaonik nad nacrtanom osom, s donjim tjemenima u tačkama Q1 i Q3. Pravougaonik prepolovite uspravnom crtom na mjestu medijana
- Označite prave koje leže za 1.5*IQ lijevo od donjeg kvartila – donja granica, i za 1.5*IQ desno od gornjeg kvartila – gornja granica
- Označite najmanji podatak između donje granice i kvartila. Povucite liniju od pravougaonika do tog podatka. Slično, povucite liniju sa desne strane pravougaonika do najvećeg podatka između kvartila i gornje granice
- Podaci koji se nalaze izvan granica su izuzeci. Njih označite posebnom oznakom

Zadatak

- Kocku smo bacali 20 puta i zabilježili rezultate:

  6 3 3 6 3 5 6 1 4 6

  3 5 5 2 2 2 2 3 2 3

- Odredite aritmetičku sredinu, mod i medijan uzorka
- Odredite varijansu i standardnu devijaciju uzorka
- Odredite raspon uzorka
- Odredite donji i gornji kvartil, i interkvartilni raspon uzorka
- Nacrtajte dijagram pravougaonika

Dijagram pravouganika 3

- Koliko se podataka nalazi u pravouganiku?
- Koliko se podataka nalazi izmedju donje i gornje granice?
- Koji podaci su izuzeci?