Mere varijabiliteta – Zadaci

Deskriptivna statistika

Aleksandar Tomašević

Mart 2020.

Sadržaj

1 Primer

U tabeli su dati podaci o starosnoj strukturi 140 pacijenata zaraženim virusom COVID-19 u jednom gradu.

Tabela 1: Varijabla X (starosna grupa pacijenata) i njihove frekvencije

interval	f
20-29	10
30-39	20
40-49	22
50-59	32
60-69	26
70-79	18
80-89	12

A) Izračunati mere centralne tendencije: aritmetičku sredinu i kvartile

B) Izračunati apsolutne mere varijacije: interval varijacije, interkvartilnu razliku, srednje apsolutno odstupanje i standardnu devijaciju

C) Izračunati relativne mere varijacije: koeficijent oscilacije, koeficijent interkvartilne

razlike i koeficijent varijacije.

1.1 A

Pre početka bilo kakve analize potrebno je pronaći sredinu intervala jer su podaci prikazani u starosnim intervalima.

Tabela 2: Dodata kolona X'

interval	f	X'
20-29	10	24.5
30-39	20	34.5
40-49	22	44.5
50-59	32	54.5
60-69	26	64.5
70-79	18	74.5
80-89	12	84.5

X' je sredina intervala, tj. u ovom slučaju prvog intervala $\frac{20+29}{2} = 24.5$. Širina intervala je 10 (10 različitih godina imamo u rasponu od 20 do 29), tako da je svaka sledeća sredina intervala veća za 10.

Da bismo izračunali aritmetičku sredinu potrebna nam je kolona fX'.

Tabela 3: Dodata kolona fX'

interval	f	X'	fX'
20-29	10	24.5	245
30-39	20	34.5	690
40-49	22	44.5	979
50-59	32	54.5	1744
60-69	26	64.5	1677
70-79	18	74.5	1341
80-89	12	84.5	1014
	140		7690

Aritmetička sredina je:

$$\overline{X} = \frac{7690}{140} = 54.93$$

Prosečna starost pacijenata je 54.93, tj. približno 55 godina.

Da bismo izračunali kvartile potrebna nam je kolona kumulacije.

Tabela 4: Dodata kolona kumulacije

interval	f	X'	fX'	kumulacija
20-29	10	24.5	245	10
30-39	20	34.5	690	30
40-49	22	44.5	979	52
50-59	32	54.5	1744	84
60-69	26	64.5	1677	110
70-79	18	74.5	1341	128
80-89	12	84.5	1014	140
	140		7690	

Drugi kvartil predstavlja središnji podatak u ovoj seriji podataka, odnosno on se nalazi na poziciji $\frac{140}{2} = 70$. Prva kumulacija veća od 70 je 84, što pripada intervalu 50-59. Vrednost medijane, odnosno drugog kvartila dobijamo preko sledeće formule.

$$Q_2 = L_1 + \frac{\frac{\sum f}{2} - f_{kum}}{f_{Q_2}} \ i$$

$$Q_2 = 50 + \frac{70 - 52}{32} 10 = 55.62$$

Prvi kvartil se nalazi na poziciji $\frac{140}{4} = 35$, a prva kumulacija veća od te pozicije je 52, što odgovara intervalu 40-49. Po sličnoj formuli dobijamo vrednost prvog kvartila.

$$Q_1 = 40 + \frac{35 - 30}{22} 10 = 42.27$$

Treći kvartil nalazi se na trećoj četvrtiti, odnosno njegova pozicija je $\frac{3 \times 140}{4} = 105$. Prva kumulacija veća od 105 je 110, što odgovara intervalu 60-69.

$$Q_3 = 60 + \frac{105 - 84}{26} 10 = 68.08$$

1.2 B

Najlakša mera varijacije je interval varijacije. To je u slučaju intervalnih podataka razlika između sredine poslednjeg i prvog intervala.

$$I = X'_{max} - X'_{min} = 84.5 - 24.5 = 60$$

Drugim rečima, interval ili raspon između starosti najmlađih i najstarijih pacijenata iznosi 60 godina.

Na sličan način, pronaćićemo i interkvartilnu razliku, kao razliku između trećeg i prvog kvartila.

$$I_q = Q_3 - Q_1 = 68.08 - 42.27 = 25.81$$

Unutar središnjih 50% podataka, raspon/interval između najmanje i najveće vrednosti je 25.81.

Nešto komplikovanije je izračunavanje srednjeg **apsolutnog odstupanja**.

Formula za intervalne podatke je $S_o = \frac{\sum f|X' - \overline{X}|}{\sum f}$. Ako radimo postupno, potrebno je u tabelu dodati dve kolone: $|X' - \overline{X}|$ i $f|X' - \overline{X}|$.

Tabela 5: Dodata kolona $|X' - \overline{X}|$

interval	f	X'	fX'	kumulacija	$\lvert X' - \overline{X}\rvert$
20-29	10	24.5	245	10	30.43
30-39	20	34.5	690	30	20.43
40-49	22	44.5	979	52	10.43
50-59	32	54.5	1744	84	0.43
60-69	26	64.5	1677	110	9.57
70-79	18	74.5	1341	128	19.57
80-89	12	84.5	1014	140	29.57
	140		7690		

Ovu kolonu dobijamo tako što za svaki red tabele od vrednosti sredine intervala X' oduzmemo aritmetičku sredinu $\overline{X}$ pri čemu uzmimamo apsolutnu vrednost, odnosno ignorišemo minus za negativne vrednosti.

U sledećem koraku, potrebno je da pomnožimo vrednosti dobijene u ovoj koloni sa frekvencijom odgovarjućeg intervala.

Tabela 6: Dodata kolona $f|X' - \overline{X}|$

interval	f	X'	fX'	kumulacija	$\lvert X' - \overline{X}\rvert$	$f\lvert X' - \overline{X}\rvert$
20-29	10	24.5	245	10	30.43	304.29
30-39	20	34.5	690	30	20.43	408.57
40-49	22	44.5	979	52	10.43	229.43
50-59	32	54.5	1744	84	0.43	13.71
60-69	26	64.5	1677	110	9.57	248.86
70-79	18	74.5	1341	128	19.57	352.29
80-89	12	84.5	1014	140	29.57	354.86
	140		7690			1912

Broj koji nam je potreban za izračunavanje srednjeg apsolutnog odstupanja je suma kolone koju smo upravo izračunali, a to je **1912**. Tu vrednost delimo sa sumom frekvencija, odnosno brojem pacijenata, što je u ovom slučaju 140. Samim tim, vrednost srednjeg apsolutnog odstupanja je:

$$S_o = \frac{1912}{140} = 13.66$$

Dakle, prosečna apsolutna udaljenost podataka od aritmetičke sredine je 13.66.

Sledeća vrednost koju treba da izračunamo je **standardna devijacija**, a da bismo nju izračunali potrebna nam je varijansa. Demonstriraćemo dva načina na koji možemo izračunati varijansu.

1.2.1 Prvi način

Prvi način zasniva se na takozvanoj definicionoj formuli varijanse. Varijansa predstavlja *prosečno kvadratno odstupanje vrednosti od aritmetičke sredine* i njena formula je slična formuli za srednje apsolutno odstupanje, s tim što su razlike dignute na kvadrat. Za intervalne podatke ta formula izgleda ovako.

$$\sigma^2 = \frac{\sum f(X' - \overline{X})^2}{\sum f}$$

Ako pogledamo poslednju verziju tabele, videćemo da smo apsolutnu vrednost $X' - \overline{X}$ već izračunali, pa su nam potrebne još dve kolone: $(X' - \overline{X})^2$ i $f(X' - \overline{X})^2$.

Tabela 7: Dodata kolona $(X' - \overline{X})^2$

interval	f	X'	X'f	kumulacija	$\lvert X' - \overline{X}\rvert$	$f\lvert X' - \overline{X}\rvert$	$(X' - \overline{X})^2$
20-29	10	24.5	245	10	30.43	304.29	925.90
30-39	20	34.5	690	30	20.43	408.57	417.33
40-49	22	44.5	979	52	10.43	229.43	108.76
50-59	32	54.5	1744	84	0.43	13.71	0.18
60-69	26	64.5	1677	110	9.57	248.86	91.61
70-79	18	74.5	1341	128	19.57	352.29	383.04
80-89	12	84.5	1014	140	29.57	354.86	874.47
	140		7690			1912	

Novu kolonu dobijamo zapravo tako što ćemo kvadrirati vrednosti 5. kolone. Potrebna je još jedna kolona, koja predstavlja proizvod nove kolone i frekvencija intervala.

Tabela 8: Dodata kolona $f(X' - \overline{X})^2$

int.	f	X'	X'f	kum.	$\lvert X' - \overline{X}\rvert$	$f\lvert X' - \overline{X}\rvert$	$(X' - \overline{X})^2$	$f(X' - \overline{X})^2$
20-29	10	24.5	245	10	30.43	304.29	925.90	9259.0
30-39	20	34.5	690	30	20.43	408.57	417.33	8346.5
40-49	22	44.5	979	52	10.43	229.43	108.76	2392.6

int.	f	X'	X'f	kum.	$\lvert X' - \overline{X} \rvert$	$f\lvert X' - \overline{X} \rvert$	$(X' - \overline{X})^2$	$f(X' - \overline{X})^2$
50-59	32	54.5	1744	84	0.43	13.71	0.18	5.9
60-69	26	64.5	1677	110	9.57	248.86	91.61	2381.9
70-79	18	74.5	1341	128	19.57	352.29	383.04	6894.7
80-89	12	84.5	1014	140	29.57	354.86	874.47	10493.6
	140		7690			1912		39774.3

Potreban nam je zbir ove poslednje kolone i on iznosi 39 774.29. Ako ga ubacimo u formulu za varijansu lako ćemo je izračunati.

$$\sigma^2 = \frac{39774.29}{140} = 284.1$$

Varijansa nema interpretaciju jer je kvadratna mera. Izračunaćemo njen koren, odnosno standardu devijaciju.

$$\sigma = \sqrt{\sigma^2} = \sqrt{284.1} = 16.86$$

Dakle, prosečna udaljenost podataka od aritmetičke sredine iznosi 16.86 ili približno 17 godina. Kasnije ćete videti kako možemo detaljnije interpretirati varijansu.

1.2.2 Drugi način

Drugi način za izračunavanje varijanse zasnovan je na takozvanoj empirijskoj formuli, koju dobijamo uprošćavanjem teorijske formule. Preciznije, ideja je da se varijansa izračuna bez potrebe da računamo **aritmetičku sredinu**. Dakle, ovaj način je ubedljivo najefikasniji kada vam se u zadatku ne traži da izračunate *srednje apsolutno odstupanje* ili *aritmetičku sredinu* (drugi slučaj je ređi). Pored toga, ona zahteva ukupno dodavanje **dve kolone** u odnosu na tabelu koju imate na početku zadatka.

Formula za intervalne podatke izgleda ovako.

$$\sigma^2 = \frac{\sum fX'^2 - \frac{(\sum fX')^2}{\sum f}}{\sum f}$$

Formula izgleda komplikovano, ali kada je rastavimo na osnovne elemente postaje jednostavna.

Potrebne su nam dve kolone u osnovnu na polaznu tabelu:

- fX'
- fX'^2

Prvu kolonu dobijamo (radili smo je ranije) tako što množimo sredinu intervala sa frekvencijom.

Drugu kolonu dobijamo tako što prvu kolonu pomnožimo sa X'.

Tabela 9: Originalna tabela + sredine intervala + dodata jedna kolona

int.	f	X'	fX'
20-29	10	24.5	245
30-39	20	34.5	690
40-49	22	44.5	979
50-59	32	54.5	1744
60-69	26	64.5	1677
70-79	18	74.5	1341
80-89	12	84.5	1014
	140		7690

Jedina kolona koja nam je neophodna nakon dodavanje ove prve izračunava se tako što ćemo pomnožiti vrednosti poslednje dve kolone: 24.5×245, 34.5×690 ...

Tabela 10: Dodata kolona fX'^2

int.	f	X'	fX'	fX'^2
20-29	10	24.5	245	6002.5
30-39	20	34.5	690	23805.0
40-49	22	44.5	979	43565.5
50-59	32	54.5	1744	95048.0
60-69	26	64.5	1677	108166.5
70-79	18	74.5	1341	99904.5
80-89	12	84.5	1014	85683.0
	140		7690	462175.0

Sada kada imamo vrednosti poslednje kolone, imamo sve što nam je potrebno da bismo izračunali varijansu.

Još jednom, formula ima sledeći oblik.

$$\sigma^2 = \frac{\sum fX'^2 - \frac{(\sum fX')^2}{\sum f}}{\sum f}$$

Prvi broj koji nam treba je $\sum fX'^2$ što je zapravo suma poslednje kolone u tabeli odnosno 462175. Drugi broj koji nam treba je kvadrat sume pretposlednje kolone podeljen sa sumom frekvencija.

$$\frac{(\sum fX')^2}{\sum f} = \frac{7690^2}{140} = 422400.7$$

Varijansa onda iznosi:

$$\sigma^2 = \frac{\sum fX'^2 - \frac{(\sum fX')^2}{\sum f}}{\sum f} = \frac{462175 - 422400.7}{140} = 284.1$$

Dobili smo identičan rezultat, međutim u ovom slučaju tabela putem koje smo dobili rezultat je znatno jednostavnija.

1.2.3 Interpretacija standardne devijacije

Rekli smo da je vrednost standardne devijacije približno 17. Standardnu devijaciju pridružujemo aritmetičkoj sredini i često rezultat opisujemo na sledeći način.

$$\overline{X} \pm \sigma \approx 55 \pm 17$$

Dakle, prosečna starost pacijenata je 55 godina, a najtipičnije vrednosti udaljene su 17 godina od proseka (ispod i iznad), tako da se one kreću u intervalu od $\overline{X} - \sigma \approx 55 - 17 = 38$ godina do $\overline{X} + \sigma \approx 55 + 17 = 72$ godine.

Na sličan način, možemo utvrditi koliko je neka vrednost udaljena od aritmetičke sredine (mereno standardnim devijacijama).

Na primer, koliko je neko ko je zaražen i ima 21 godinu udaljen od proseka?

$$\frac{21 - 55}{17} = \frac{34}{17} = 2$$

Pacijent star 21 godinu je udaljen dve standardne devijacije od proseka.

Najčešće kažemo da su vrednosti udaljene najmanje 2 standardne devijacije od proseka **netipične, retke ili ekstremne vrednosti**. U ovom slučaju, retki su pacijenti koji imaju 21 godinu.

Tokom sledećeg semestra videćete kako ovu udaljenost koristimo da bismo izračunali verovatnoću da će (na primer) neko starosti 21. godinu biti zarežen (na osnovu ovih podataka).

1.3 C

Kada smo već izračunali apsolutne mere varijabiliteta, relativne mere se mogu vrlo brzo izračunati.

Koeficijent oscilacije.

$$K_{os} = \frac{I}{\overline{X}} = \frac{60}{54.93} = 1.09$$

Ovo je najmanje koristan koeficijent varijacije i nećemo ga korsiti u daljim zadacima.

Koficijent interkvartilne razlike.

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{25.8}{110.35} = 0.234 = 23.4\%$$

Varijabilitet središnjih 50% podataka iznosi 23.4%. Ovaj koeficijent nam govori o procentualnom varijabilitetu podataka unutar prostora između prvog i trećeg kvartila. Drugim rečima, ovo je procentualni varijabilitet koji ne uključuje netipične vrednosti.

Najpreciznija mera je koeficijent varijacije.

$$V = \frac{\sigma}{\overline{X}} = \frac{15.86}{54.93} = 0.307 = 30.7\%$$

Koeficijent varijacije je 30.7 procenata. Koeficijent variajcije je uvek veći od koeficijenta interkvartilne razlike jer uključuje i netipične vrednosti.

2 Zadatak za vežbanje

U tabeli ispod dati su takođe podaci o pacijentima i njihovoj starosnoj strukturi, ali za drugi grad.

Tabela 11: Varijabla X (starosna grupa pacijenata) i njihove frekvencije

int.	f	X'
20-29	2	24.5
30-39	11	34.5
40-49	31	44.5
50-59	37	54.5
60-69	29	64.5
70-79	3	74.5
80-89	4	84.5

Potrebno je uraditi sledeće:

A. Izračunati aritmetičku sredinu, varijansu i standardnu devijaciju. B. Izračunati koeficijent varijacije. C. Prokomentarisati varijabilitet ovih podataka u odnosu na prethodni zadatak D. Koliko standardnih devijacija je pacijent star 21 godinu udaljen od proseka? E. U kom zadatku je manje verovatno ili ređe da pronađemo pacijenta starog 21. godinu?