1 Materijal - I dio

1.1 Percentilni rang

Prilikom određivanja percentila, mi smo za dati percentil računali konkretnu vrijednost iz uzorka koja datom percentilu odgovara. Sada treba za datu konkretnu vrijednost da odredimo kom pecentilu odgovara. Opisani postupak zovemo određivanjem percentilnog ranga.

U slučaju kada podaci nisu grupisani, percentilni rang neke vrijednosti A računamo po formuli

$$Pr = \frac{L \cdot 100}{n}\,, \tag{1}$$

gdje je L broj vrijednosti koje su u uzorku manje od A i n je ukupan broj elemenata uzorka.

Primjer 1 *Rezultati studentskog takmičenja iz opšte kulture su: 95, 62, 75, 84, 85, 89, 100, 88 i 79. Odrediti percentilni rang za studenta čije je postignuće 89.*

Rješenje: Rješenje ovog zadatka podrazumijeva da se odredi kom percentilu odgovara vrijednost 89.

Sortirajmo uzorak u rastući poredak:
62, 75, 79, 84, 85, 88, 89, 95, 100.

Broj elemenata koji su manji od 89 je $L = 6$, a u uzorku ima $n = 9$ elemenata. Primjenjujući (1), dobijamo

$$Pr = \frac{6 \cdot 100}{9} = 66,7\%\,.$$

Zaključujemo da je $66,7\%$ studenata imalo lošije postignuće od studenta koji je na takmičenju osvojio 89 bodova.

□

Ukoliko su podaci grupisani u intervale, percentilni rang elementa Y računamo na sledeći način:

1. Odredi se interval kome pripada Y.
2. Neka je (a, b) interval koji sadrži Y. Tada se on računa po formuli

$$Pr = \frac{F \cdot 100}{n} + \frac{Y - a}{i}\frac{f \cdot 100}{n}\,, \tag{2}$$

gdje je i širina intervala, n ukupan broj elemenata u uzorku, f frekvenca intervala u kome se nalazi Y, F kumulativna frekvenca intervala koji prethodi intervalu koji sadrži Y.

PRIMJER 2 *Koristeći podatke iz navedene tabele, izračunati percentilni rang za promet 57 500.*

Table 1: Prosječan promet

Promet u hilj. EUR	Broj radnji (f_i)
30-40	2
40,01-50	5
50,01-60	10
60,01-70	12
70,01-80	10
80,01-90	9
90,01-100	2

Rješenje: Stavimo da je $Y = 57,5$. Očigledno je da ostvareni promet pripada intervalu $(50, 60]$. Koristeći formulu (2), dobijamo

$$Pr = \frac{7 \cdot 100}{50} + \frac{57,5 - 50,01}{10} \frac{10 \cdot 100}{50} = 28,98\% \,.$$

Dakle, radnja koja ima promet od 57 500 EUR ima promet veći od $28,98\%$ drugih radnji.

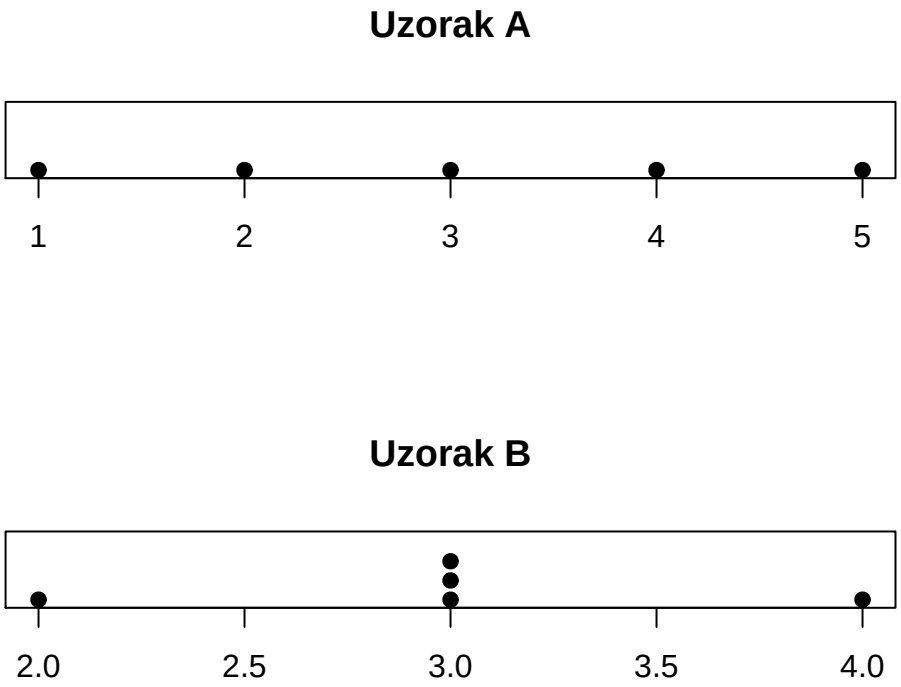
□

2 Mjere varijabiliteta

Mjere centralne tendencije često nisu dovoljne da u potpunosti opišu raspodjelu nekog uzorka. Varijacija nekog uzorka podrazumijeva odstupanje elemenata uzorka od jedne unaprijed određene vrijednosti. Da bi dobili potpunu informaciju o nekom uzorku potrebno je da, pored aritmetičke sredine, imamo i neku mjeru varijacije. U nastavku ćemo proučavati varijaciju od aritmetičke sredine.

Na slici 1 data su dva uzorka koja imaju istu aritmetičku sredinu ($\overline{x} = 3$), ali su različite varijacije. Evidentno je da elementi uzorka A imaju veću

Figure 1: Uzorci različite varijacije.



varijaciju u odnosu na aritmetičku sredinu. Kada bi aritmetička sredina bila dovoljna karakterizacija nekog uzorka, mogli bi da kažemo da uzorci A i B potiču iz iste populacije, što, naravno, nije tačno. Zato se uvode mjere mjere varijacije, koje opisuju koliko se vrijednosti nekog uzorka međusobom razlikuju.

Reprezentativnost neke numeričke karakteristike uzorka (npr. aritmetičke sredine) zavisi od stepena varijabiliteta. Ukoliko je varijabilnost manja, utoliko su vrijednosti obilježja manje odstupaju od aritmetičke sredine i ona je reprezentativnija, a za takav uzorak kažemo da je homogen. Obrnuto, ako je varijabilnost veća, odstupanje pojedinačnih vrijednosti od aritmetičke sredine je veće, pa je reprezentativnost aritmetičke sredine manja. Za takav skup kažemo da je heterogen.

Na primjer, ako imamo informaciju da je prosjek liječenja u jednoj bolnici 8 dana, a u drugoj takođe 8 dana, mogli bi da dođemo do pogrešnog zaključka da je dužina trajanja liječenja jednaka u obije bolnice. Međutim, to može ali ne mora da bude. Znači, da bi smo mogli da poredimo dva ili više uzoraka, pored informacije o aritmetičkoj sredini, moramo da imamo i informaciju o odstupanju pojedinačnih vrijednosti od prosjeka.

Apsolutne mjere disperzije varijabilnosti su:

1. raspon,
2. varijansa ili disperzija ,
3. standardna devijacija,
4. interkvartilni rang.

2.1 Raspon

Najprostija mjera varijacije naziva se raspon. Raspon R se definiše kao razlika najveće i najmanje vrijednosti u uzorku, tj.

$$R = X_{max} - X_{min}\,.$$

Raspon je najprostiji pokazatelj varijabiliteta nekog uzorka. Njime se dobija samo približna informacija o varijabilitetu, jer na njega utiču samo dvije krajnje vrijednosti u uzorku. Ukoliko su obije ili bar jedna krajnja vrijednost ekstremna raspon neće biti prava mjera varijabiliteta. Drugi, isto tako važan nedostatak, jeste što se prilikom izračunavanja raspona ne uzima u obzir broj elemenata u uzorku.

Primjer 3 *Data su dva niza mjera:*

a) *7, 11, 18, 5, 9, 6, 10, 14.*

b) *7, 11, 30, 5, 9, 6, 10, 14.*

Izračunati raspon.

Rješenje: Za prvi niz mjera raspon je $R = 18 - 5 = 13$, dok je raspon za drugi niz mjera $R = 30 - 5 = 25$.

S obzirom da se prethodna dva niza mjera razlikuju samo u maksimalnoj vrijednosti, očigledan je uticaj ekstremnih vrijednosti na raspon.

□

2.2 Interkvartilni rang

Kao što smo vidjeli prilikom definisanja raspona, ideja da se varijabilitet u nekom uzorku mjeri kao razlika maksimalne i minimalne vrijednosti pokazala je određene nedostatke. Postavlja se pitanje da li se može mjera varijabiliteta definisati kao razlika neke dvije vrijednosti na koje ekstremne vrijednosti ne bi imale uticaj. To se postiže uvođenjem interkvartilnog ranga (IQR) koji je jednak razlici trećeg i prvog kvartila, tj.

$$IQR = Q_3 - Q_1 \, .$$

Može se zaključiti da IQR nije podložan uticaju ekstremnih vrijednosti, jer sve jedinice čije su vrijednosti veće od trećeg i manje od prvog kvartila ne učestvuju u njegovom izračunavanju.

Postupak računanja IQR može se svesti na sledeće korake:

1. uzorak se sortira u rastući poredak;
2. odredi se uzoračka medijana (ili drugi kvartil);
3. da bi se odredio prvi kvartil formiramo poduzorak koji se nalazi lijevo od medijane (ne uključujući medijanu). Prvi kvartil će biti medijana tako dobijenog poduzorka.
4. da bi se odredio treći kvartil formiramo poduzorak koji se nalazi desno od medijane (ne uključujući medijanu). Treći kvartil je medijana tako dobijenog poduzorka.

PRIMJER 4 *Jedna osiguravajuća kuća tokom godine isplatila je 18 odšteta vlasnicima automobila koji su učestvovali u saobraćajnim udesima. Visine odštete u eurima su: 675, 991, 346, 237, 211, 233, 189, 119, 370, 141, 467, 195, 100, 735, 802, 618, 180, 165. Odrediti interkvartilni rang.*

Rješenje: Zadatak riješavamo na prethodno opisan način:

1. Podatke sortiramo u rastući poredak:

   100, 119, 141, 165, 180, 189, 195, 211, 233, 237, 346, 370, 467, 618, 675, 735, 802, 991.

2. Ukupno je $n = 18$ opservacija, pa je drugi kvartil, odnosno medijana

$$Q_2 = \frac{233 + 237}{2} = 235.$$

3. Poduzorak koji se nalazi lijevo od medijane je:

   100, 119, 141, 165, 180, 189, 195, 211, 233.

   Prvi kvartil je medijana tako dobijenog poduzorka, odnosno $Q_1 = 180$.

4. Poduzorak koji se nalazi desno od medijane je:

   237, 346, 370, 467, 618, 675, 735, 802, 991.

   Medijana prethodnog poduzorka je $Q_3 = 618$.

Sada je IQR=618-180=438.
□

Jedna od pogodnosti interkvartilnog ranga je da se može koristiti za detekciju ekstremnih vrijednosti. Postupak se sastoji u sledećim koracima:

1. Izračunava se IQR.

2. Određujemo donju i gornju granicu:

$$D = Q_1 - 1,5 \cdot IQR$$

   i

$$G = Q_3 + 1,5 \cdot IQR.$$

3. Ako je vrijednost u uzorku manja od D ili veća od G tada se uzoračka vrijednost smatra ekstremnom vrijednošću.

Primjer 5 *Da li u uzorku*
180, 189, 370, 618, 735, 802, 1 185, 1 414, 1 657, 1 953, 2 332, 2 336, 3 461, 4 668, 6 751, 9 908, 10 034, 21 147
postoje nestandardne opservacije?

Rješenje: Postupajući slično kao prethodnom primjeru, zaključujemo da je IQR =4 668 - 735 = 3 933.

Sada računamo donju vrijednost:

$$D = 735 - 1,5 \cdot 3933 = -5164,5.$$

Gornja vrijednost je

$$G = 4668 + 1,5 \cdot 3933 = 10567,5.$$

Zaključujemo da nijedna vrijednost uzorka nije manja od D, pa ne postoje nestandardno male opservacije. Međutim, vrijednost 21 147 je veća od G i ona predstavlja nestandardno veliku opservaciju.

□

2.3 Disperzija

Interkvartilni rang, kao ni raspon ne uzima u obzir odstupanje svih elemanta nekog uzorka. Ova činjenica se može smatrati još veoma ograničavajućim faktorom. Zato se nameće potreba da konstruišemo takvu mjeru varijabiliteta koja će uzimati u obzir odstupanja svih elemenata uzorka od jedne konkretne vrijednosti. U našem slučaju aritmetička sredina se prirodno nameće kao tražena vrijednost.

Ako su $x_1, x_2, \ldots, x_N$ elementi neke populacije sa aritmetičkom sredinom μ, tada se centralni momenat reda k računa po formuli

$$M_k = \frac{1}{N} \sum_i (x_i - \mu)^k \,. \tag{3}$$

Specijalno, centralni momenat drugog reda zovemo varijansom, odnosno važi

$$\sigma^2 = \frac{1}{N} \sum_i (x_i - \mu)^2 \,. \tag{4}$$

Na jednostavan način prethodna formula se može uprostiti tako da dobijamo

$$\sigma^2 = \frac{\sum_i x_i^2 - N\overline{x}^2}{N} \,. \tag{5}$$

Prethodni izraz je operativniji, pa se češće koristi za računanje varijanse.

A ako su podaci grupisani u intervale, tada se centralni momenat (3) svodi na

$$M_k = \frac{1}{N} \sum_i f_i \left(x_i - \mu\right)^k . \tag{6}$$

S obzirom da je disperzija drugi centralni momenat, pomoću (6) lako dolazimo do izraza računanje disperzije u slučaju grupisanih podataka

$$\sigma^2 = \frac{1}{N} \sum_i f_i \left(x_i - \mu\right)^2 . \tag{7}$$

PRIMJER 6 *Završnu godinu nekog fakulteta pohađa 12 studenata i svi su polagali ispit iz Statistike. Odrediti varijansu ako su rezultati dati Tabelom 2*

Table 2: Broj osvojenih bodova na ispitu

Šifra studenta	Broj bodova
A1	69
A2	58
A3	74
A4	90
A5	55
A6	61
A7	78
A8	84
A9	95
A10	52
A11	59
A12	71

Rješenje: Obilježimo broj bodova svakog studenta sa x_i, gdje je $i = 1, 2, \ldots, 12$. Da bi primijenili formulu (5) formiramo sledeću radnu tabelu:

Sada lako dobijamo

$$\sigma^2 = \frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{N}}{N} = \frac{61838 - \frac{846^2}{12}}{12} = 182,92 .$$

Table 3: Radna tabela sa postupkom izračunavanja varijanse

Šifra studenta	Broj bodova (x_i)	x_i^2
A1	69	4761
A2	58	3364
A3	74	5476
A4	90	8100
A5	55	3025
A6	61	3721
A7	78	6084
A8	84	7056
A9	95	9025
A10	52	2704
A11	59	3481
A12	71	5041
	$\sum_i x_i = 846$	$\sum_i x_i^2 = 61838$

□

Izraz (7) se može uprostiti tako da se dobije sledeća operativnija formula

$$\sigma^2 = \frac{\sum_i f_i\, x_i^2}{N} - \mu^2\,. \tag{8}$$

Neka je $x_1, x_2, \ldots, x_n$ uzorak sa aritmetičkom sredinom $\overline{x}$. Tada se uzorački centralni momenat reda k definiše kao

$$m_k = \frac{1}{n}\sum_i (x_i - \overline{x})^k\,. \tag{9}$$

Uzoračka disperzija predstavlja uzorački centralni momenat drugog reda, s tim što se suma kvadrata odstupanja svih elemenata uzorka od aritmetičke sredine dijeli sa $n-1$, tj.

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2\,. \tag{10}$$

Kao i u slučaju populacione varijanse prethodna formula se može uprostiti na sledeći način

$$s^2 = \frac{\sum_i x_i^2 - n\overline{x}^2}{n-1}\,. \tag{11}$$

Može se postaviti pitanje zašto se prilikom računja populacione varijanse odgovarajuća suma kvadrata odstupanja dijeli sa N, a u slučaju uzoračke sa $n-1$. U praksi se statističko zaključivanje izvodi na bazi uzorka, iz razloga što je prikupljanje podataka od svih elemenata neke populacije vrlo često nemoguće (zbog ograničenja resursa). Zato je populacionu varijansu potrebno na najbolji mogući način procijenti. Naime, u teorijskoj statistici se kaže da je (10) ocjena populacione varijanse (4). Da bi neka ocjena bila preciznija, ona mora da zadovolji i neke osobine. To je i razlog što se u (10) suma kvadrata odstupanja dijeli sa $n-1$, a ne sa n kako bi bilo očekivano. Osobine ocjena izlaze van okvira ovog kursa. Više o ovoj temi može se naći u XXX. Veličinu $n-1$ zovemo broj stepeni slobode. Mi ćemo ovdje pokušati da damo intuitivnu interpretaciju broja stepeni slobode. Naime, polazimo od činjenice da je $\sum_i (x_i - \overline{x}) = 0$. Ako imamo poznat $n-1$ element uzorka i aritmetičku sredinu, tada se $n-$ ti element uzorka mora izračunati tako da važi prethodni uslov. Na primjer, ako imamo uzorak od 3 elementa i poznato je $x_1 = 4$, $x_2 = 7$ i $\overline{x} = 11$. Tada element x_3 određujemo iz uslova

$$\begin{aligned} \frac{x_1 + x_2 + x_3}{3} &= \overline{x} \\ \frac{4 + 7 + x_3}{3} &= 11 \\ 11 + x_3 &= 33 \\ x_3 &= 22\,. \end{aligned}$$

Vidimo da $n-1$ element ima "slobodu" da uzme bilo koju vrijednost, dok $n-$ ti element tu slobodu nema. Dakle, uzorak je potpuno određen ako imamo poznatu $n-1$ opseravciju i aritmetičku sredinu $\overline{x}$.

PRIMJER 7 *Koristeći rezultate iz prethodnog primjera, izračunati uzoračku varijansu ako su u uzorak izabrani studenti A2, A4, A6, A8, A10 i A12.*

Rješenje: Za izračunavanje uzoračke varijanse koristićemo formulu (11). U tom cilju formiramo Tabelu 4

Sada je

$$s^2 = \frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}}{n-1} = \frac{29986 - \frac{416^2}{6}}{5} = 228,67\,.$$

□

U slučaju kada je uzorak grupisan u intervale, uzoračka centralni momenat reda k je

$$m_k = \frac{1}{n} \sum_i f_i \, (x_i - \overline{x})^k \,. \tag{12}$$

Table 4: Radna tabela sa postupkom izračunavanja uzoračke varijanse

Šifra studenta	Broj bodova (x_i)	x_i^2
A2	58	3364
A4	90	8100
A6	61	3721
A8	84	7056
A10	52	2704
A12	71	5041
	$\sum_i x_i = 416$	$\sum_i x_i^2 = 29986$

dok je uzoračka varijansa u slučaju grupisanih podataka

$$s^2 = \frac{1}{n-1} \sum_i f_i \, (x_i - \overline{x})^2 \, .$$

Prethodna formula se može uprostiti

$$s^2 = \frac{1}{n-1} \left(\sum_i f_i \, x_i^2 - n \, \overline{x}^2 \right) \, . \tag{13}$$

PRIMJER 8 *Trideset učenika jednog odjeljenja ocijenjeno je na kraju školske godine iz fizike na sledeći način:*

Table 5: Ocjene iz fizike

Ocjena	5	4	3	2	1	$\sum$
Frekvenca	4	8	9	6	3	30

Odrediti varijabilitet.

Rješenje: Smatraćemo da su učenici iz odabranog odjeljenja uzorak na kome se sprovodi neko istraživanje. Zato koristimo formulu (13). Formiramo radnu Tabelu 6

Kao što je pokazano ranije aritmetička sredina je $\overline{x} = \frac{94}{30} = 3,13$. Sada je

$$s^2 = \frac{1}{n-1} \left(\sum_i f_i \, x_i^2 - n \, \overline{x}^2 \right) = \frac{1}{29} \left(336 - 30 \cdot 3,13^2 \right) = 1,43 \, .$$

Table 6: Radna tabela sa primjerom izračunavanja varijanse kod podataka datih u obliku frekvence

Ocjena (x_i)	Frekvenca(f_i)	x_i^2	$f_i\, x_i$	$f_i\, x_i^2$
5	4	25	20	100
4	8	16	32	128
3	9	9	27	81
2	6	4	12	21
1	3	1	3	3
		$\sum_i x_i^2 = 55$	$\sum_i f_i\, x_i = 94$	$\sum_i f_i\, x_i^2 = 336$

□

Ukoliko su bilo uzorački podaci dati u obliku intervala, tada se populaciona odnosno uzoračka varijansa računa tako što se u (8) odnosno (13), umjesto x_i, stavi sredina intervala x_i'.

2.4 Standardna devijacija

Ako pogledamo primjere u kojima smo računali varijansu, možemo da zaključimo da je varijansa izražena u kvadratima mjernih jedinica (bodovi na kvadrat u Primjeru 8). Ovo se može smatrati značajnim nedostatkom, jer se na taj način povećava i mjera varijabiliteta. Zato je prirodno da se računa kvadratni korijen iz varijanse. Pozitivnu vrijednost kvadratnog korijena iz varijanse zovemo standardnom devijacijom. Dakle, populaciona standardna devijacija je $\sigma = +\sqrt{\sigma^2}$, dok je uzoračka $s = +\sqrt{s^2}$.

Primjer 9 **a)** *Izračunati standardnu devijaciju koja odgovara populaciji iz Primjera 6.*

b) *Izračunati standardnu devijaciju koja odgovara uzorku iz Primjera 8.*

Rješenje: Na osnovu definicije standardne devijacije dobijamo:

a)

$$\sigma = \sqrt{\frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{N}}{N}} = \sqrt{182,92} = 13,52\,.$$

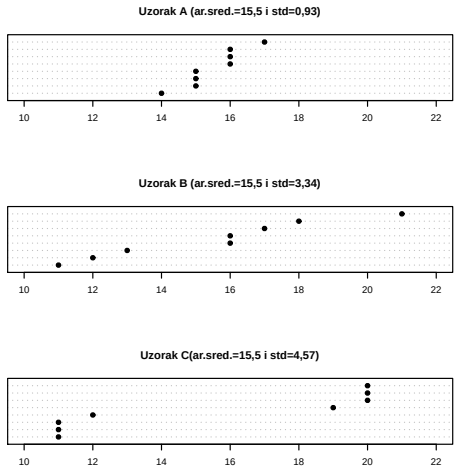
b)

$$s = \sqrt{\frac{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}}{n-1}} = \sqrt{228,67} = 15,12\,.$$

□

Na Slici 2 prikazana su tri različita uzorka sa istom aritmetičkom sredinom i različitim standardnom devijacijom. U slučaju kada je standardna devijacija najmanja (Uzorak A) sve vrijednosti uzorka su koncentrisane oko aritmetičke sredine. Porast standardne devijacije, dovodi do heterogenosti uzorka, odnosno do većeg odstupanja od aritmetičke sredine (Uzorci B i C).

Figure 2: Uticaj standardne devijacije.



2.5 Koeficijent varijacije

Mjere varijacije koje smo do sada izučavali izražene su istim jedinicama kojima je izražen i uzorak. Postavlja se pitanje kako upoređivati varijabilitet uzoraka koji se mjere različitim jedinicama mjere. Slično pitanje možemo da postavimo i u slučaju upoređivanja varijabiliteta uzoraka koji imaju istu jedinicu mjere ali različite aritmetičke sredine. Odgovori na prethodna pitanja motivišu uvođenje relativnih mjera varijabiliteta od kojih je najpoznatiji koeficijent varijacije ili skraćeno CV.

Populacioni koeficijent varijacije definiše se kao odnos populacione standardne devijacije i populacione aritmetičke sredine, tj.

$$CV = \frac{\sigma}{\mu}\,.$$

Analogno se definiše i uzoračka standardna devijacija

$$CV = \frac{s}{\overline{x}}\,.$$

Preporuka je da se pri upoređivanju varijabiliteta dva ili više uzoraka koristi koeficijent varijacije.

Prisustvo varijabiliteta u uzorku možemo da shvatimo na na sledeći način:

1. Heterogenost podataka znači da će raspon, interkvartilni rang, varijansa, standardna devijacija i koeficijent varijacije biti veći.

2. Homogenost podataka znači da će raspon, interkvartilni rang, varijansa, standardna devijacija i koeficijent varijacije biti manji.

3. Jednakost svih elemenata populacije ili uzorka znači odsustvo varijabiliteta, što povlači da su sve mjere varijabiliteta jednake 0.