Mere varijabilnosti- disperzije

Osnovna karakteristika vrednosti jednog istog obeležja je, da ta vrednost varira od jedne do druge statističke jedinice osnovnog skupa. Te vrednosti, mere centralne tendencije, a pre svega aritmetička sredina, sažimaju u jednu brojčanu vrednost. Ta brojčana vrednost treba da bude reprezentativna za sve vrednosti. Njena reprezentativnost zavisi od stepena varijabilnosti pojedinačnih vrednosti u odnosu na centralnu vrednost, konkretno u odnosu na aritmetičku sredinu. Ukoliko je varijabilnost manja, utoliko su vrednosti obeležja sabijenije oko aritmetičke sredine (manje odstupaju) i ona je reprezentativnija, a za takav skup kažemo da je homogen. Obrnuto, ako je varijabilnost veća, odstupanje pojedinačnih vrednosti od aritmetičke sredine je veće, a reprezentativnost aritmetičke sredine je manja i za takav skup kažemo da je heterogen.

Sa druge strane, ako imamo informaciju da je prosek lečenja u jednoj bolnici 8 dana, a u drugoj takođe 8 dana, mogao bi da navede na pogrešan zaključak da je dužina trajanja lečenja kod pojedinih slučajeva u većini jednaka u obe bolnice. Drugim rečima, da su rasporedi dužine trajanja lečenja po pacijentu, jednaki u obe bolnice. Međutim, to može ali ne mora da bude tako. Znači, da bi smo mogli da poredimo dve ili više serija, pored informacije o prosečnoj vrednosti, moramo da imamo i informaciju o odstupanju pojedinačnih vrednosti od proseka.

Sledi zaključak da mere varijabilnosti zapravo ukazuju na reprezentativnost mera centralne tendencije. Manja mera varijabilnosti ukayuje na veću repreyentativnost srednje vrednosti i obrnuto. Mere varijabilnosti nas opredeljuju koju od meara centralne tendencije trba da koristimo, aritmetičku sredinu (ukoliko je skup homogen), ili medijanu (ukoliko je skup heterogen).

Varijabilnost, disperzija, odstupanje pojedinačnih vrednosti ispitivanog obeležja u odnosu na prosek merimo tzv. merama varijabilnosti ili disperzije. One mogu da budu, prema brojčanom izrazu apsolutne i relativne.

Apsolutne mere disperzije

Apsolutne mere disperzije varijabilnosti su:

1) Interval varijacije;

2) Interkvartilna razlika;

3) Varijansa i

4) Standardna devijacija.

Relativne mere varijabilnosti su:

1) Koeficijent varijacije i

2) Standardizovano odstupanje ili z-vrednost.

Interval varijacije

Interval varijacije je gruba i orjentaciona mera varijacije i predstavlja razliku između maksimalne i minimalne vrednosti serije. Izračunava se po formuli:

Iv=Xmax-Xmin·

Iv-interval varijacije

Primer 1: U prethodnom primeru, o dužini lečenja u dve bolnice, najduže lečenje jednog bolesnika je iznosilo 22 dana, a najkraće leenje je bilo 6 dana. U drugoj bolnici najduže lečenje je bilo 18 dana, a najkraće5dana. Na osnovu oviih podataka:

Iv =Xmax-Xmin = 22-6=16 prva bolnica
Iv = Xmax-Xmin= 18-5=13 druga bolnica

Na osnovu dobijenih vrednosti za interval varijacije zaključujemo:

1. U prvoj bolnici su obe ekstremne vrednosti udaljenije od centralne vrednosti serije nego u drugoj bolnici;
2. Više se slučajeva u drugoj bolnici po dužini lečenja grupiše oko proseka u odnosu na prvu bolnicu. Manji interval varijabilnosti koincidira sa većom grupisanošću članova serije oko centralne vrednosti;
3. Što je veći interval varijacije to je veća varijabilnost pojedinačnih vrednosti oko proseka, to je prosek manje reprezentativan i obrnuto, manja vrednost, manja varijabilnost, veća sabijenost, veća reprezentativnost proseka.

Nedostaci:

a) Uzima u obzir samo dve vrednosti, odnosno, samo dva člana serije, sa najvećom i najmanjom vrednošću i

b) Obzirom da se radi o ekstremnim vrednostima to one mogu da budu veoma udaljene od osnovne koncentracije ostalih vrednosti serije.

Interkvartilna razlika

Primer 2: Izmerena je telesna masa 11 novorođenčadi i dobijene vrednosti su sređene po veličini:

N	1	2	3	4	5	6	7	8	9	10	11
x	2,8	3,2	3,4	3,6	3,7	3,8	4,0	4,4	4,6	4,8	5,0
			Q_1			Q_2			Q_3		

Medijanu ovog statističkog skupa predstavlja telesna masa šestog novorođenčeta pa je Me = 3,8 kg. Kao što nam je već poznato, medijana deli niz na 50% vrednosti manje od medijane i na 50% vrednosti veće od medijane. Ako svaku od ove dve polovine podelimo na još po pola, dobijamo četiri jednaka dela statističke serije od kojih svaki sadrži po 25% vrednosti serije. Ovako ustrojeni delovi serije nazivaju se kvartilima (Q), četvrtinama.

Kako jedan kvartil predstavlja četvrtinu serije, to se njihova mesta u nizu vrednosti izračunavaju na sledeći način:

1. Mesto prvog kvartila $Q_1 = \frac{N+1}{4} = \frac{12}{4} = 3$. Težina trćeg novorođenčeta predstavlja vrednost prvog kvartila (Q_1), pa je njegova vrednost Q_1=3,4 kg.

   U intervalu od 2,8 (minimalna vrednost) do 3,4 kg nalazi se 25% svih vrednosti serije i ovo su 25% najmanjih vrednosti od ukupnog broja vrednosti (od 100% vrednosti, od svih vrednosti).

2. Mesto drugog kvartila $Q_2 = \frac{N+1}{2} = \frac{12}{2} = 6$. Telesna masa 6, novorođenčeta predstavlja vrednost drugog kvartila, pa je Q_2=3,8 kg. Medijana je ustvari drugi kvartil serije pa je Me = Q_2. U intervalu između drugog (medijane) i prvog kvartila nalazi se 25% vrednosti, koje su veće od vrednosti intervala prvog kvartila, ali su manje od svih ostalih vrednosti. U našem primeru ovaj interval je između 3,4 i 3,8 kg.

3. Mesto trećeg kvartila $Q_3 = \frac{3N+1}{4} = \frac{36}{4} = 9$. Telesna masa devetog novorođenčeta predstavlja vrednost trećeg kvartila pa je Q_3=4,6 kg. U intervalu između Q_2 (medijane) i Q_3 nalaze se 25% vrednosti serije veće od prethodnih vrednosti ali manje od ostalih 25% vrednosti serije. Za naš primer to je interval od 3,8 kg (Q_2=Me) do 4,6 kg = Q_3.

4. Vrednost četvrtog intervala predstavlja maksimalnu vrednost serije, pa je u našem primeru Q_4=5,0 kg.

Osnovni zaključak je: Između prvog (Q_1) i trećeg (Q_3) kvartila nalaze se 50% svih vrednosti serije, a van ovog intervala ostaju još 50% vrednosti, od kojih 25% manjih od Q_1 (ekstremno najmanje vrednsoti) i 25% vrednosti veće od Q_3 (ekstremno najveće vrednosti).

Zato se razlika između trećeg i prvog kvartila uzima kao mera varijabilnosti jer ova mera, za razliku od intervala varijacije isključuje ekstremno male i ekstremno velike vrednosti.[1]

Dakle, interkvartilnu razliku, možemo da definišemo kao distancu između prvog i trećeg kvartila, pa je formula za njegovo izračunavanje:

$$Ig = Q_3 - Q_1 .$$

Za naš primer:

$Iq = Q_3\text{-}Q_1 = 4{,}2 - 3{,}2 = 1{,}2$ kg

U intervalu između 3,4 kg i 4,6 kg, nalazi se telesna masa 50% novorođenčadi, odnosno u intervalu od 1,2 kg. Na ovaj način smo izbegli merenje disperzije na osnovu ekstremnih vrednosti.

Zaključivanje o stepenu varijabilnosti je isto kao i kod intervala varijacije: Što je interkvartilna razlika manja to je varijabilnost manja, a sabijenost vrednosti oko centra veća i obrnuto.

Interkvartilna razlika i interval varijacije mogu da se upoređuju i ako je interkvartilna razlika znatno manja od intervala varijacije, to znači da na krajevima sređene serije, postoje ekstremno niske i ekstremno visoke vrednosti. U našim primeru za telesnu masu novorođenčadi:

a) $Iv = X_{max} - X_{min} = 5{,}0 - 2{,}8 = 2{,}2$ kg

b) $Iq = Q_3\text{-}Q_1 = 4{,}6 - 3{,}4 = 1{,}2$ kg

c) $Iv/Iq = 2{,}2/1{,}2 = 1{,}8$

Interval varijacije je skoro dva puta veći od interkvartilne razlike, što znači da postoje novorođenčadi, između 25% gore i dole koja imaju ekstremne vrednosti u odnosu na prosek, pa je statistička serija (uzorak) heterogena.

Za sve izvedene radnje i konstatacije uslov je i da je serija sređena po veličini vrednosti od najmanje do najveće ili obrnuto. Drugo, izneti primeri se odnose samo na osnovnu seriju (prost statistički niz) i sa neparnim brojem podataka u njemu.

Varijansa i standardna devijacija

Dok interval varijacije obuhvata samo dve vrednosti serije, a interkvartilna razlika 50% vrednosti, dotle varijansa i standardna devijacija obuhvataju distancu svih vrednosti u odnosu na prosek (centar), odnosno u odnosu na aritmetičku sredinu.

Kako je zbir odstupanja svih članova serije od aritmetičke sredine jednak nuli, to nije moguće izračunati prosek odstupanja.[2] Da bi se izbegla 0, pristupilo se kvadriranju razlika pojedinačnih vrednosti od aritmetičke sredine, i iz njihovog zbira je izračunato prosečno kvadratno odstupanje.

Prosečno kvadratno odstupanje svih vrednosti serije od aritmetičke sredine, izračunato gore definisanim postupkom predstavlja Varijansu čija je matematička definicija:

$$Варијанса - SD^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n} \qquad SD^2 = \frac{\sum X^2}{n} - \overline{X}^2$$

Dok interval varijacije obuhvata samo dve vrednosti serije, a interkvartilna razlika 50% vrednosti, dotle varijansa i standardna devijacija obuhvataju distancu svih vrednosti u odnosu na prosek (centar), odnosno u odnosu na aritmetičku sredinu.Za statističku seriju koja nije sređena u vidu distribucije frekvencije.

[2] Може да се израчуна и апсолутно просечно одступање од просека, када се занемаре плус и минус вредности, али оно не омогућава даље статистичке операције, па овде као такво није ни обрађено.

Primer 3: Uzmimo za primer telesne mase 11 novorođenčadi iz prethodnog podpoglavlja: 2,9 3,0 3,2 3,4 3,5 3,7 3,9 4,1 4,2 4,5 4,7 kg.

Rešenje: Radi izračunavanja varijanse po navedenim formulama treba konstruisati radnu tabelu

N	X	$X - \overline{X}$	$\left(X - \overline{X}\right)^2$	X^2
1	2,9	2,9 - 3,73 = -0,83	0,69	8,41
2	3,0	3,0 - 3,73 = -0,73	0,54	9,00
3	3,2	3,2 - 3,73 = -0,53	0,28	10,24
4	3,4	3,4 - 3,73 = -0,33	0,11	11,56
5	3,5	3,5 - 3,73 = -0,23	0,053	12,25
6	3,7	3,7 - 3,73 = -0,03	0,0009	13,69
7	3,9	3,9 - 3,73 = +0,17	0,029	15,21
8	4,1	4,1 – 3,73 = +0,37	0,14	16,81
9	4,2	4,2 - 3,73 = +0,47	0,22	17,64
10	4,5	4,5 - 3,73 = +0,77	0,59	20,25
11	4,7	4,7 - 3,73 = +0,97	0,94	22,09
Σ	41,1	0,00	3,59	157,15

Postupak

1. Prvo izračunamo aritmetičku sredinu:

$$\overline{X} = \frac{\sum X}{n} = \frac{41,1}{11} = 3,73kg$$

Izračunamo razlike između svake vrednosti i vrednosti aritmetičke sredine: x-3,73 kg (treća kolona u radnoj tabeli). Obavezno treba proveriti da li je suma razlika jednaka 0. Ako nije, onda postoji greška u izračunavanju aritmetičke sredine. Zanemarljiva razlika, kao u našem primeru može da se javi zbog zaokruživanja decimala.

2. Dobijene razlike kvadriramo i kvadrate saberemo i na taj način dobijemo ukupnu sumu kvadratnog odstupanja:

$\Sigma x - x^{-2}$ (četvrta kolona u radnoj tabeli)

4. Na osnovu podataka iz radne tabele, a na osnovu formule, dobijamo:

$$SD^2 = \frac{\sum \left(X - X^2\right)^2}{n} = \frac{3,59}{11} = 0,33kg$$

Po drugoj formuli, koja se obično i naziva radnom formulom za varijansu, znatno je jednostavnije izračunati varijansu, a vrednost je ista. Potrebno je u radnoj tabeli formirati kolonu za kvadrate svake vrednosti i dobijene kvadrate sabrati Σx^2.

Kako je za naš primer $\Sigma x^2 = 157{,}15$ (peta kolona u radnoj tabeli) to je:

$$SD^2 = \frac{\Sigma x^2}{n} - \bar{x} = \frac{157{,}15}{11} - 13{,}91 = 14{,}24 - 13{,}91$$

$$SD^2 = 0{,}33 \text{кг}$$

I ovim postupkom je dobijena ista vrednost za varijansu, pa ga zbog jednostavnosti treba primenjivati u praksi.

Odmah treba uočiti, da je vrednost varijanse dobijena na drugom stepenu, odnosno 0,33 kg^2, što je apsurdna vrednost i nepogodna za poređenje.

Da bi se izbegla apsurdnost, pristupilo se vađenju korena iz dobijene vrednosti za varijansu, i tako se dobilo prosečno odstupanje u istim mernim jedinicama u kojima je izražena i vrednost posmatranog obeležja. Tako se ustvari došlo do Standardne devijacije, kao apsolutne mere varijabilnosti, pa je:

Standardna devijacija kvadratni koren iz varijanse, a formule za njeno izračunavanje:

$$SD^2 = \sqrt{SD^2} = \sqrt{\frac{\Sigma x - \bar{x}^2}{n}} \quad \text{или SD} = \sqrt{\frac{\Sigma x^2}{n} - x^{-2}}$$

Standardna devijacija za naš primer bi imala vrednost:

$$SD = \sqrt{SD^2} = \sqrt{0{,}33} = 0{,}57 \text{ кг}$$

Odnosno SD = 570 grama.

Zaključivanje: Ukoliko je vrednost standardne devijacije manja, to je sabijenost vrednosti oko aritmetičke sredine veća, pa je i njena reprezentativnost za seriju (uzorak ili osnovni skup) veća, i obrnuto. Veća vrednost standardne devijacije - veća varijabilnost i sve ostalo što sledi iz toga.

Nedostatak: Standardna devijacija omogućava poređenje između varijabilnosti dve serije ako su vrednosti date u istim mernim jedinicama i ako su aritmetičke sredine serija međusobno jednake. Međutim, i pored ovog nedostatka, kao što ćemo videti, u statističkoj metodologiji pored aritmetičke sredine, standardna devijacija je odigrala najznačajniju ulogu.

Varijansa i standardna devijacija za distribuciju frekvencije, kao i aritmetička sredina, zahtevaju složenije matematičke postupke, mada u principu postoji potpuna analogija sa izračunavanjem kod proste statističke serije.

a) Varijansa i standardna devijacija za osnovnu distribuciju frekvencije (bez grupnih intervala) izračunavaju se po formulama:

$$SD^2 = \frac{\sum f\left(X - \overline{X}\right)^2}{\sum f} \; ili \; \frac{\sum fX^2}{\sum f}$$

a podsetimo se: $\bar{x} = \frac{\Sigma fx}{\Sigma f}$,

U praksi treba koristiti drugu formulu za koju je sadržaj radne tabele:

x	f	fx	fx^2
x_1	f_1	$f_1 x_1$	$(f_1 x_1)x_1 = f_1 x_1^2$
x_2	f_2	$f_2 x_2$	$(f_2 x_2)x_2 = f_2 x_2^2$
x_3	f_3	$f_3 x_3$	$(f_3 x_3)x_3 = f_3 x_3^2$
$\therefore$	$\therefore$	$\therefore$	$\therefore$
x_n	f_n	$f_n x_n$	$(f_n x_n)x_n = f_n x_n^2$
Σ	Σf	Σfx	Σfx^2

Formule za izračunavanje varijanse i standardne devijacije bile bi:

$$SD^2 = \frac{\Sigma f \bar{x}_i^2}{\Sigma f} - \bar{x}^2 \quad i \quad SD = \sqrt{SD^2} = \sqrt{\frac{\Sigma f \bar{x}_i^2}{\Sigma f} - \bar{x}^2} \; .$$

Podsetimo se:

1. Aritmetička sredina grupnog intervala (x_i) izračunava se tako što se početna (donja, manja) vrednost i završna vrednost (gornja, veća) intervala saberu, pa se dobijeni zbir podeli sa 2.
2. Formula aritmetičke sredine za distribuciju frekvencija sa grupnim intervalima je:

$$\bar{x} = \frac{\Sigma f \bar{x}_i}{\Sigma f} .$$

Relativne mere disperzije

Osnovni nedostatak apsolutnih mera varijabilnosti, pa i standardne devijacije kao najrelevantnije, je u tome što se njihove vrednosti moraju da iskazuju u mernim jedinicama u kojima je iskazano posmatrano obeležje, pa nije moguće poređenje varijabilnosti dve serije sa različitim mernim jedinicama. Ovaj problem je razrešen relativnim merama varijabilnosti (1) Koeficijentom varijacije i (2) z -vrednošću.

Koeficijent varijacije

Koeficijent varijacije (*Cv*) je odnos (količnik) između standardne devijacije i aritmetičke sredine. Obično se iskazuje u procentima, pa je njegova formula:

$$C_v = \frac{SD}{\bar{x}} \cdot 100 .$$

Primer 4: Za statistički niz telesnih masa 11 novorođenčadi, kod koga je $\bar{x}$ =3,73 kg, a SD=0,57 kg, koeficijent varijacije je:

$$Cv = = \frac{0{,}57}{3{,}73} * 100 \qquad = 15{,}28\%$$

Primer 5: Prosečna telesna dužina istih 11 novorođenčadi iznosila je $\bar{x}$ =50 cm i SD=10 cm. Da li je dužina novorođenčadi varijabilnija od telesne mase pri rođenju?

$$C_v = \frac{SD}{\bar{x}} \cdot 100 = \frac{10}{50} \cdot 100 = \frac{500}{50} = 10\% .$$

Zaključivanje: Što je relativna vrednost koeficijenta varijabilnosti manja, to je i varijabilnost manja, sabijenost oko proseka veća, njegova reprezentativnost veća. Ustvari, što se tiče koeficijenta varijacije, postoji pravilo po kome ako je relativna vrednost koeficijenta

manja od 30%, statistički niz, (uzorak, osnovni skup) može se smatrati homogenim, a aritmetička sredina, reprezentativnom centralnom vrednošću.

$C_v < 30\%$ - homogeni skup

$C_v > 30\%$ heterogeni skup

Prema ovom pravilu, i telesna masa i telesna dužina 11 novorođenčadi predstavljaju homogen uzorak, ($Cv \leq 15{,}28\% < 30\%$ i $Cv = 10\% < 30\%$), pri čemu je telesna dužina znatno homogenija, odnosno znatno je manja prosečna varijabilnost pojedinačnih vrednosti u odnosu na sopstvenu aritmetičku sredinu.

Standardizovano odstupanje

Sve dosadašnje mere varijabilnosti, su mere zajedničkog (ukupnog) odstupanja svih vrednosti od sopstvenog proseka. Međutim, varijacija se može ocenjivati i sa gledišta individualnih podataka, odnosno svake vrednosti pojedinačno. Tako, možemo da postavimo pitanje koliko odstupa telesna masa novorođenčeta od 3 kg, od aritmetičke sredine $\bar{x}$ =3,73 kg (primer 11 novorođenčadi). U apsolutnom iznosu ovo odstupanje je:

$x - \bar{x} = 3 - 3{,}73 = -0{,}73$

Odgovor na pitanje da li je ova razlika velika ili mala, daje upoređenje razlike sa standardnom devijacijom istog niza, čija je vrednost: SD=0,57 kg. Stavljanjem ove dve vrednosti u odnos dobija se:

$$\frac{0{,}73}{0{,}57} = 1{,}28$$

Vrednost od 1,28 govori da je telesna masa novorođenčeta od 4 kg udaljena od aritmetičke sredine 1,28 standardnih devijacija. Znak minus (-) znači, da je ova vrednost manja od aritmetičke sredine ($x=3 < \bar{x} =3{,}73$) i da se na grafičkom prikazu nalazi levo od aritmetičke sredine.

Odstupanje svake individualne vrednosti serije od aritmetičke sredine te serije, kada se izrazi u jedinicama standardne devijacije, predstavlja standardizovano odstupanje ili Z-vrednost kako se ono najčešće iskazuje u statistici. Izračunava se po formuli:

$$z - vrednost = \frac{x - \bar{x}}{SD}.$$

Standardizovano odstupanje predstavlja opšte sredstvo ocene odstupanja individualnih podataka od aritmetičke sredine i njihovog raspoređivanja oko aritmetičke sredine. Z-vrednost ima svoju posebnu ulogu, kod tzv. standardizovanog normalnog rasporeda, o čemu će biti reči u posebnom poglavlju.

Mere varijabiliteta-izračunavanje u MS Excelu

Interval varijacije:

Najjednostavnija i najgrublja mera varijabilnosti i predstavlja razliku minimalne i maksimalne vrednosti: X_{max}-X_{min}.

U programu MS Excel nemamo direktnu funkciju za izračunavanje intervala varijacije, ali ga veoma lako izračunavamo koristeći se formulom:

=MAX(raspon podataka) – MIN(raspon podataka).

Varijansa i standardna devijacija:

Varijansa predstavlja prosečno kvadratno odstupanje svih vrednosti od aritmetičke sredine, a njenim korenovanjem dobijamo standardnu devijaciju. Standarna devijacija predstavlja najvažniju meru varijabilnosti, pa samim tim i najčešće korišćenu pri prikazivanju rezultata.

U programu MS Excel varijansu iz negrupisanih podataka izračunavamo pomoću funkcije:

=VARP(raspon podataka)

Za izračunavanje standardne devjacije iz negrupisanih podataka koristimo funkciju:

=STDEVP(raspon podataka)

Inerkvartilna razlika:

Interkvartilna razlik razliku trećeg i prvog kvartila. U programu MS Excel se izračunava prema funkciji:

=QUARTILE(raspon podataka;3) - QUARTILE(raspon podataka;1)

Koeficijent varijacije:

Koeficijent varijacije, ili Cv, predstavlja relativnu meru varijabilnosti, i služi nam za procenu varijabilnosti nekog statističkog niza, poređenjem standardne devijacije i aritmetičke sredine. To je vrednost koja nam omogućava poređenje varijabilnosti dva statistička niza, koja nemaju iste srednje vrednosti. Ukoliko je Cv<30%, smatra se da je skup homogen, i suprotno, ukoliko je Cv>30%, smatramo da je skup heterogen.

U programu MS Excel ne postoji funkcija za direktno izračunavanje koeficijenta varijacije, već se izračunava upisivanjem obrasca:

=STDEVP(raspon podataka)/AVERAGE(raspon podataka)*100