5. Нумеричке дескриптивне мере: случај негруписаних података

Серија података се обично описује уз помоћ *нумеричких дескриптивних мера* или *карактеристичних вредности*. Прво посматрајмо тзв. *мере централне тенденције*: *аритметичка средина*, *медијана* и *модус*. У овој глави посматрамо ове мере у случају негруписаних података. *Аритметичку средину негруписаних података* дефинишемо као:

$$\textbf{аритметичка средина} = \frac{\textbf{збир свих вредности}}{\textbf{број вредности}}$$

Ако је у питању основни скуп, онда се ово своди на формулу

$$\mu = \frac{\sum x}{N},$$

а ако је у питању узорак на формулу

$$\overline{x} = \frac{\sum x}{n}.$$

Пример 1. За серију података из примера 1.13 имамо да је

$$\overline{x} = \frac{\sum x}{n} = \frac{138}{10} = 13{,}93.$$

Медијана је средњи члан серије која је уређена претходно у неопадајући низ:

$$x_1, x_2, \ldots, x_n.$$

Она дели низ података на две једнаке, по броју елемената, групе и не мора да одговара вредности неког члана серије. Имамо два случаја:

1) ако је број чланова серије непаран, онда је медијана једнака средњем члану серије, тј. ако је $n = 2k + 1$, онда је медијана једнака x_{k+1};

2) ако је број чланова серије паран, тј. $n = 2k$, онда се по договору за медијану узима вредност

$$\frac{x_{2k} + x_{2k+1}}{2}$$

Пример 2. За серију података из примера 1.13, после сређивања података у неопадајући низ, добијамо низ

5, 9, 11, 12, 13, 15, 17, 18, 19, 21, 21, 21, 22, 23, 24

Како је број елемената серије непаран, то је у нашем случају медијана једнака вредности 18. □

Модус је вредност која се јавља са највећом фреквенцијом у серији података. Серија, с обзиром на то колико има различитих вредности које се појављују са највећом фреквенцијом у њој, може бити *унимодална*, *бимодална*, *мултимодална*, а може и да не постоји уопште модус за дату серију (у случају када су сви подаци серије различити међу собом).

Пример 3. За серију података из примера 1.13 имамо да је модус број 21, а серија је унимодална. □

Односи између аритметичке средине, медијане и модуса зависе од типа криве расподеле фреквенција. У случају променљиве са симетричном кривом расподеле фреквенција вредности које узимају аритметичка средина, медијана и модус се поклапају. У случају када је крива расподеле фреквенција асиметрична удесно [улево] највећа [најмања] је аритметичка средина, а медијана се налази између аритметичке средине и модуса.

Често мере централне тенденције нису довољне у анализи негруписаних података, па се стога користе и *мере дисперзије*. Основне мере дисперзије су: *интервал варијације*, *варијанса* и *стандардна девијација*.

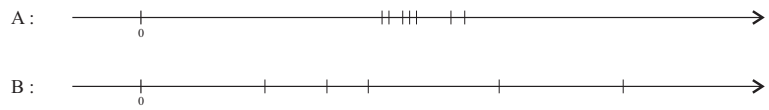
Пример 4. Посматрају се године старости радника две фирме. У случају фирме A добијена је серија

$$45, \quad 38, \quad 39, \quad 35, \quad 47, \quad 40, \quad 36$$

а у случају фирме B добијена је серија

$$18, \quad 52, \quad 70, \quad 33, \quad 27$$

Представимо ове серије одговарајући вредностима на бројној оси:



Обе ове серије имају исту аритметичку средину — 40, али и површни поглед на горњу слику нам говори да се овде ради о две веома различите серије. Стога аритметичка средина у овом случају недовољно добро описује ове групе података. □

Интервал варијације се дефинише следећом релацијом:

интервал варијације = највећа вредност − најмања вредност

Варијанса основног скупа σ^2 и *варијанса узорка* s^2 се рачунају по следећим формулама:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}, \quad s^2 = \frac{\sum(x-\overline{x})^2}{n-1}$$

(у последњој формули се уместо n користи вредност $n-1$ да вредност дисперзије узорка не би била потцењена).

Уместо горњих формула користе се тзв. *радне формуле* за израчунавање варијансе које знатно олакшавају њихову примену и изведене су из њих:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}, \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}.$$

Сада, *стандардну девијацију*, у случају основног скупа и узорка, дефинишемо као

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}.$$

Пример 5. У једном списку најбогатијих људи планете који је објавио Forbes 8. марта 2012. године фигурисала су и имена следећих милијардера (са одговарјућом сумом новца израженом у милијардама долара, које су они поседовали у том тренутку):

милијардер	**богатство**
Carlos Slim Helu	69
Bernard Arnault	41
Amancio Ortega	37,5
Larry Ellison	36
Eike Batista	30
Li-Ka-Shinga	25,5

Одредимо варијансу и стандардну девијацију ових података. Имамо да је

x	x^2
69	4761
41	1681
37,5	1406,25
36	1296
30	900
25,5	650,25

Из ове таблице добијамо да је

$$\sum x = 239, \quad \sum x^2 = 10694{,}5$$

Применом наведених формула добијамо да је

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{10694{,}5 - \frac{(239)^2}{6}}{6-1} = 234{,}87, \quad s = \sqrt{234{,}87} = 15{,}33. \square$$

Приметимо да су вредности σ^2, s^2, σ и s увек ненегативне, а једнаке су све нули једино ако су сви подаци једнаки по величини.

6. Нумеричке дескриптивне мере: случај груписаних података

Посматрајмо сада формуле за израчунавање аритметичке средине, варијансе и стандардне девијације груписаних података.

За аритметичку средину груписаних података користимо формуле

$$\mu = \frac{\sum x'f}{N}, \quad \overline{x} = \frac{\sum x'f}{n}$$

где је x' одговарајућа средина групног интервала, а f његова фреквенција.

Пример 6. Сарадници једне фирме да би дошли на посао троше времена у просеку на пут до 50 минута. Следећом таблицом су дати подаци колико запослени ове фирме троше времена у просеку на пут:

Време проведено у транспорту	**Број запослених**
0–< 10	4
10–< 20	9
20–< 30	6
30–< 40	4
40–< 50	2

На основу ове таблице можемо формирати следећу таблицу:

Време проведено у транспорту	f	x'	$x'f$
0–< 10	4	5	20
10–< 20	9	15	135
20–< 30	6	25	150
30–< 40	4	35	140
40–< 50	2	45	90

Из последње таблице израчунавамо аритметичку средину за дате груписане податке:

$$\mu = \frac{535}{25} = 21{,}40. \square$$

Варијанса груписаних података, у случају основног скупа и узорка, се израчунава, редом, по следећим формулама

$$\sigma^2 = \frac{\sum f(x' - \mu)^2}{N}, \quad s^2 = \frac{\sum f(x' - \overline{x})^2}{n-1}$$

Последње формуле се своде на следеће (тзв. *радне формуле*, које су једноставније за коришћење у пракси):

$$\sigma^2 = \frac{\sum x'^2 f - \frac{(\sum x' f)^2}{N}}{N}, \quad s^2 = \frac{\sum x'^2 f - \frac{(\sum x' f)^2}{n}}{n-1}$$

Стандардна девијација груписаних података се, као и у случају негруписаних података, рачуна, у зависности од тога да ли посматрамо основни скуп или узорак, по следећим формулама

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}$$

Пример 7. Погледајмо претходни пример. Тада можемо формирати следећу таблицу:

Време проведено у транспорту	f	x'	$x'f$	$x'^2 f$
0–< 10	4	5	20	100
10–< 20	9	15	135	2025
20–< 30	6	25	150	3750
30–< 40	4	35	140	4900
40–< 50	2	45	90	4050

Из таблице добијамо да је

$$N = 25, \quad \sum x' f = 535, \quad \sum x'^2 f = 14825,$$

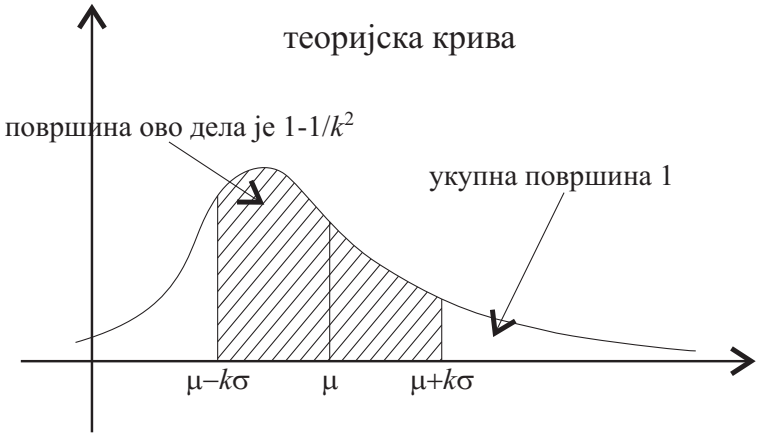
па је

$$\sigma^2 = \frac{\sum x'^2 f - \frac{(\sum x' f)^2}{N}}{N} = 135{,}04,$$

односно,

$$\sigma = \sqrt{135{,}04} = 11{,}62 \,\text{мин.} \,\square$$

Пре него што дамо неке примере коришћења стандардне девијације дајмо једно правило које се често користи у пракси, а заснованост тог правила следи и следећег тврђења, које дајемо без доказа.



Теорема 1. [**Чебишовљева теорема**] *За било који број $k > 1$, најмање $1 - 1/k^2$ вредности (релативна фреквенција) података се налази у опсегу k стандардних девијација од аритметичке средине, односно, $(1 - 1/k^2) \cdot 100\,\%$ података се налази у $k\sigma$-околини тачке μ.*

Пример 8. На пример, ако је $k = 2$, онда се

$$1 - \frac{1}{k^2} = 1 - 0{,}25 = 75\,\%$$

података налази у 2σ-околини тачке μ.

Ако је $k = 3$, онда се

$$1 - \frac{1}{k^2} = 89\,\%$$

података налази у 3σ-околини тачке μ. □

Пример 9. У једној групи од 4000 жена са високим крвним притиском утврђено је да је

$$\mu = 187 \quad \text{и} \quad \sigma = 22.$$

Одредити најмањи проценат жена у овој групи које имају систолни крвни притисак између 143 и 231.

Како је $231 - 187 = 44$, то је $k = 44/22 = 2$, па је тражени проценат једнак $1 - 1/k^2 = 75\,\%$. □

Из теореме Чебишова може се извести тзв. *емпиријско правило*, које тврди да се за нормалну расподелу

1) 68 % вредности се налази у опсегу 1σ

2) 95 % вредности се налази у опсегу 2σ

3) 99,7 % вредности се налази у опсегу 3σ

око аритметичке средине.

Пример 10. У једном узорку од 5000 особа утврђено је да је

$$\overline{x} = 40\,\text{година} \quad \text{и} \quad s = 12.$$

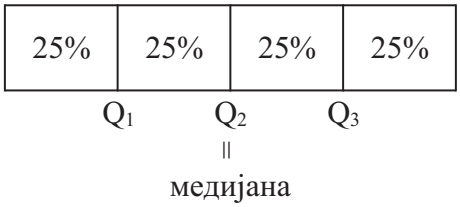
Одредити проценат ове популације који се налази у старосним границама од 16 до 64 године.

Применимо емпиријско правило. Видимо да интервал $[16, 64]$ представља $2s$-околину тачке 40, па је стога најмање 95 % ове групе особа у датом старосном интервалу. □

7. Позиционе мере

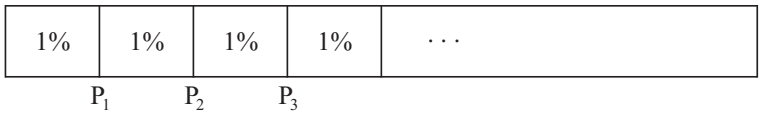
Позиционе мере одређују позицију једне вредности у односу на друге вредности у серији података. Најчешће се користе следеће позиционе мере: *квартил*, *перцентил* и *ранг перцентила*.

Квартили су три дескриптивне мере које деле серију података рангираних по величини на четири једнака дела; означавамо их са Q_1, Q_2 и Q_3. Очигледно да вредност Q_2 је једнака медијани дате серије података.



Често се уочава и тзв. *интерквартилна разлика* која се дефинише на следећи начин

$$\textbf{интерквартилна разлика} = Q_3 - Q_1$$



Перцентили су вредности $P_1, P_2, \ldots, P_{99}$ дефинисане тако да је P_k вредност од које је тачно k % података мање или једнако. Вредност k-тог перцентила P_k се рачуна по следећој приближној формули:

$$P_k = \textbf{вредност } kn/100\textbf{-тог члана у рангираној серији података}$$

Ранг перцентила за одређену вредност x_i је проценат вредности у серији података које су мање од x_i и одређује се по следећој формули:

$$\textbf{ранг перцентила за } x_i = \frac{\textbf{број вредности мањих од } x_i}{\textbf{укупан број података у серији}}$$