



# Бизнис статистика

---

## Предавање 3: Дескриптивни статистики



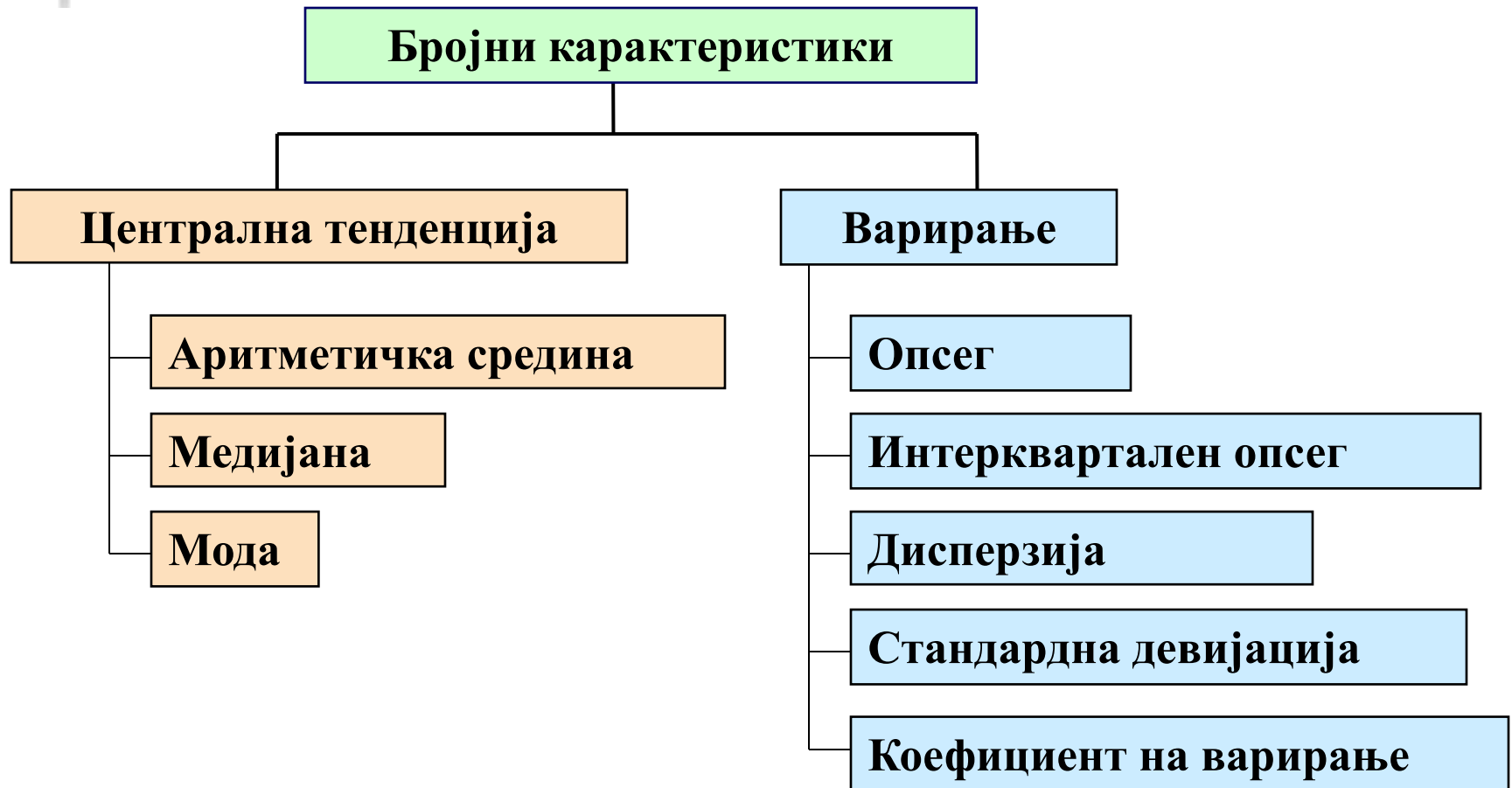
# Вовед

---

- Во претходното предавање, видовме како од необработените (сирови) податоци се добиваат табели на фреквенција, хистограми и други визуелни прикази.
- И покрај тоа што ваквите графички техники им овозможуваат на истражувачите да донесат некакви општи заклучоци за обликот и варирањето на податоците, поцелосно разбирање на податоците може да се постигне ако за податоците се пресметаат одредени бројни карактеристики кои се вредности на таканаречените дескриптивни статистики.
- Во ова предавање ќе бидат воведени такви статистички техники за опишување на податоците, кои може да се поделат на две големи групи: мерки на централна тенденција, мерки на варијабилност (расејување) на податоците.



# Бројни карактеристики на податоците





# Мерки на централна тенденција

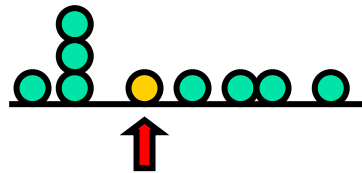
## Централна тенденција

Просек

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

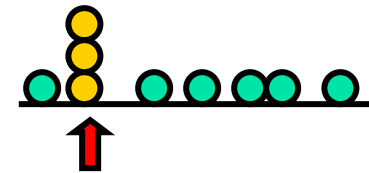
Аритметичка  
средина

Медијана



Средина на  
подредени  
вредности

Мода



Најчесто  
набљудувана  
вредност



# Аритметичка средина (Просек)

- Аритметичката средина (просек) е најчеста мерка за централна тенденција
- За примерок со големина  $n$ , просекот на примерокот се дефинира со:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

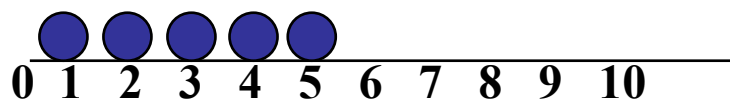
Набљудувани  
вредности

Големина на примерок



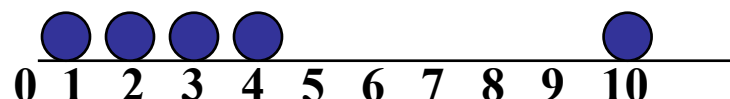
# Аритметичка средина

- Аритметичката средина е најпозната мерка за централна тенденција.
- Недостаток на оваа мерка за централна тенденција е тоа што е многу осетлива на екстремни вредности. Тоа се вредности кои значително отстапуваат од повеќето други вредности.
- Така, во примерот подолу, ако само еден податок (вредноста 5), се замени со екстремна вредност 10, аритметичката средина се зголемува за 1. А ако бројот 5 се замени со 100, тогаш аритметичката средина ќе биде 22 и е за 19 поголема од првобитниот случај.



**Просек = 3**

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



**Просек = 4**

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$



# Медијана

- Медијаната на примерокот е број што стои на средина на подредениот примерок. Приближно 50% од податоците во подредениот примерок се лево од него и приближно 50% од податоците се десно од него.
- Ако бројот на податоци е непарен, медијана е податокот во средината.
- Ако бројот на податоци е парен, тогаш во средината има два податока, па медијаната е просек на тие два податока.

■ Всушност,

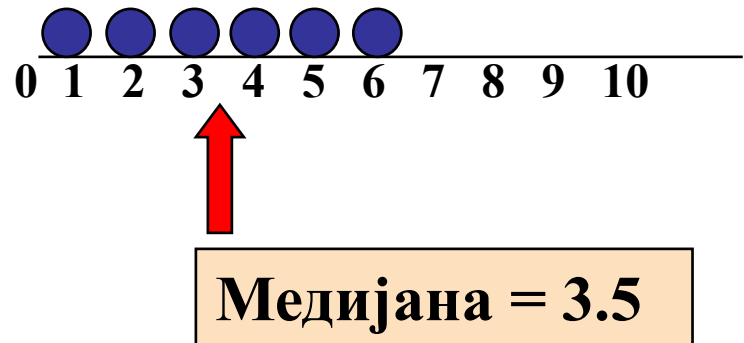
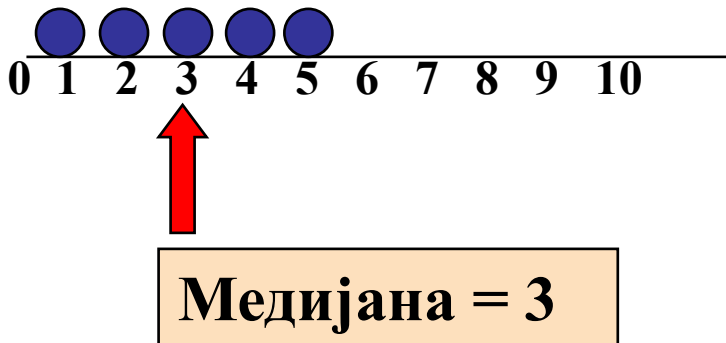
$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{ако } n \text{ е непарен} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{ако } n \text{ е парен} \end{cases},$$

каде  $x_{(k)}$  го означува  $k$ -тиот елемент во подредениот примерок.



# Медијана

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{ако } n \text{ е непарен} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{ако } n \text{ е парен} \end{cases},$$

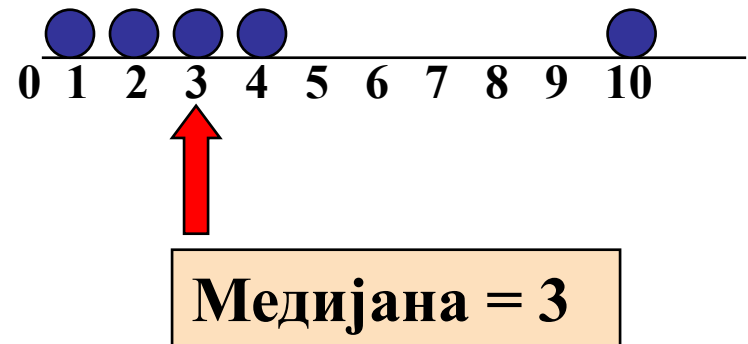
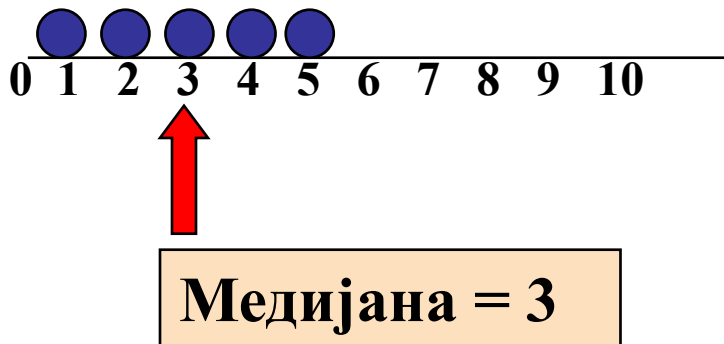






# Медијана

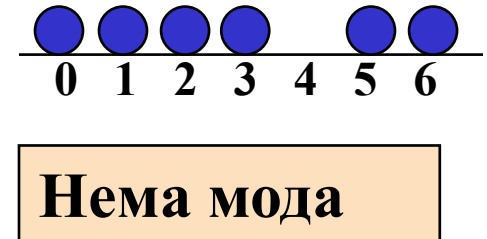
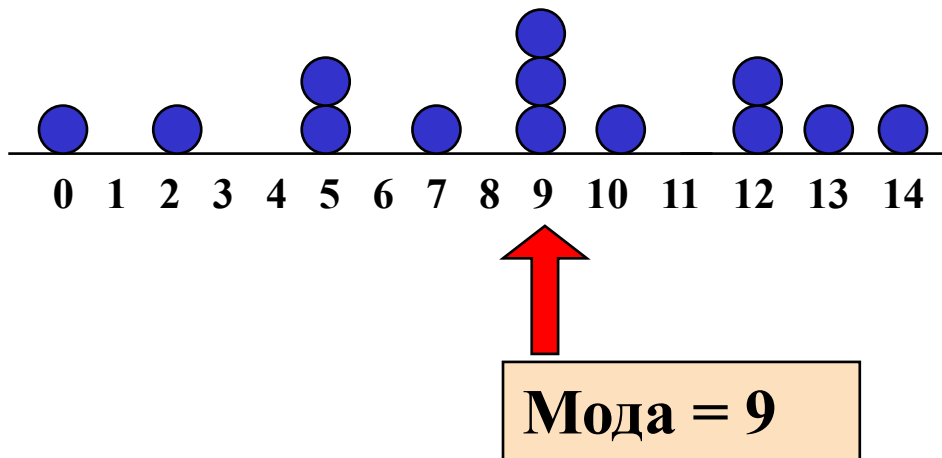
- Медијаната не е осетлива на екстремни вредности.
- Како што може да се види во примерот подолу, ако само еден податок (вредноста 5), се замени со екстремна вредност 10, медијаната не се менува.





# Мода

- Модата е мерка за централна тенденција.
- Тоа е вредност од примерокот кој има најголема честота.
- На оваа вредност не и влијаат екстремни вредности.
- Мода може да се пресмета и за нумерички и за категоришки податоци.
- Еден примерок може да нема мода, но може да има и повеќе моди.





# Која е „најдобрата“ мерка за централна тенденција?

- Вообичаено се користи просекот, освен кога постојат екстремни вредности (аутлаери).
- Во тој случај, често се користи медијаната, бидејќи таа не е сензитивна на екстремни вредности.
- Просекот главно се преферира за нумерички податоци, но не и за категориски податоци.
- Да претпоставиме дека во една мала група студенти има 2 машки (кои ќе се лабелираат со 1) и 3 женски (кои ќе се лабелираат со 2) студенти. Аритметичката средина на лабелите ќе биде  $(1+1+2+2+2)/5=1.6$  и е практично бесмислена. Од друга страна, модата е 2 што значи дека во групата има повеќе женски отколку машки студенти.



## Пример:

- Во табелата дадени се 10-те најголеми производители на автомобили во светот и бројот на возила произведени од секоја во последната година.

Производител	Производство (во милиони)
Toyota Motor Corp.	9.37
General Motors	8.90
Volkswagen AG	6.19
Ford Motor Co.	5.96
Hyundai-Kia Automotive Group	3.96
Honda Motor Co. Ltd.	3.83
Nissan Motor Co.	3.68
PSA/Peugeot- Citroen SA	3.43
Chrysler LLC	2.68
Fiat S.p.A.	2.62

- Просек

$$\bar{x} = \frac{9.37+8.90+6.19+5.96+3.96+3.83+3.68+3.43+2.68+2.62}{10}$$
$$= 5.062$$

- Медијана (средна вредност на двата броја на средината од подредениот примерок)

$$\frac{3.96 + 3.83}{2} = 3.895$$



# Перцентили

- Перцентилите ги делат подредените податоци во 100 дела со еднаква големина (секој содржи 1% од податоците).
- Постојат 99 перцентили, секој одговара на една од 99-те прегради за да се подели групата на податоци во 100 делови.
- $n$ -тиот перцентил е вредноста така што приближно  $n$  проценти од податоците се лево од таа вредност во подредениот примерок и приближно  $(100 - n)$  проценти се десно.
- На пример, 87-от перцентил ( $P_{87}$ ) е вредноста за која 87% од податоците се лево и не повеќе од 13% се десно од таа вредноста.



# Перцентили

$P$ -тиот перцентил се пресметува со следната постапка.

1. Податоците се подредуваат во неопаѓачки редослед.
2. Позицијата на перцентилот во подредениот примерок се пресметува со:

$$i = \frac{P}{100} \cdot n$$

$P$  = перцентилот што го бараме

$n$  = обемот на примерокот.

3.  $P$ -тиот перцентил зависи од тоа дали  $i$  е цел број или не е цел број.
  - Ако  $i$  е цел број,  $P$ -тиот перцентил е просекот на вредноста на  $i$ -тата позиција и вредноста на  $(i+1)$ -та позиција.
  - Ако  $i$  не е цел број,  $P$ -тиот перцентил е вредноста која се наоѓа на позицијата  $[i+1]$ , каде  $[i+1]$  е целиот дел од бројот  $i+1$ .



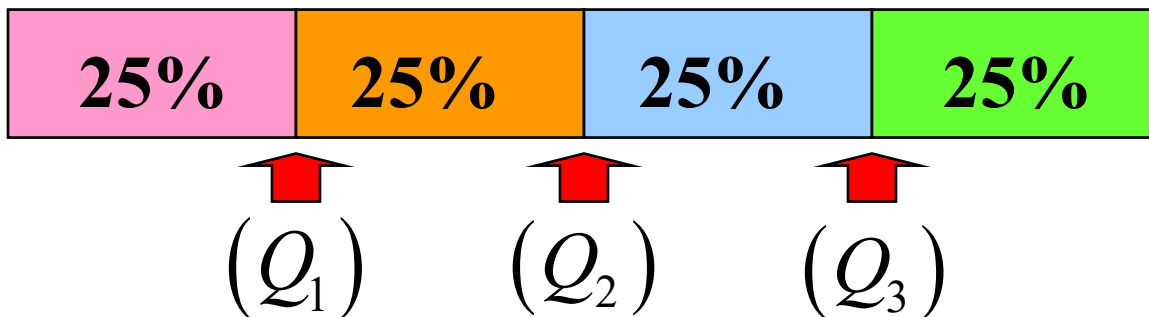
# Квартили

- Во статистиката, *квартили* (*квартал*) се вид на перцентили кои го делат примерокот на четири дела, или четвртини, со приближно еднаква големина.
- За да се пресметаат кварталите, податоците мора да бидат подредени во неопаѓачки редослед.
- Трите главни квартали се следните:
  - Првиот квартил  $Q_1 = P_{25}$  и тоа е вредност таква што приближно 25% од податоците во подредениот примерок се лево од него, а приближно 75% се десно. Познат е и како долен квартил.
  - Вториот квартил  $Q_2 = P_{50}$  е медијана на примерокот, така што приближно 50% од податоците во подредениот примерок се лево од него и приближно 50% се десно.
  - Третиот квартил  $Q_3 = P_{75}$  и тоа е вредност таква што приближно 75% од податоците во примерокот се лево од него и приближно 25% се десно. Познат е и како горен квартил.



# Квартили

- Подредените податоци се разделени во 4 квартили



**Подредени податоци: 11   12   13   16   16   17   18   21   22**

Локацијата на  $Q_1$  :  $i = \frac{25}{100} \cdot 9 = 2.25$ .

$$Q_1 = x_{(\lfloor 2.25+1 \rfloor)} = x_{(3)} = 13$$

Притоа,  $x_{(k)}$  го означува  $k$ -тиот елемент во подредениот примерок.





## Пример

Да се определат кварталите за податоците од примерот со производители на автомобили.

Производител	Производство (во милиони)
Fiat S.p.A.	2.62
Chrysler LLC	2.68
PSA/Peugeot-Citreon SA	3.43
Nissan Motor Co.	3.68
Honda Motor Co. Ltd.	3.83
Hyundai-Kia Automotive Group	3.96
Ford Motor Co.	5.96
Volkswagen AG	6.19
General Motors	8.90
Toyota Motor Corp.	9.37

- Прв квартал

$$i = \frac{25}{100} \cdot 10 = 2.5, \text{ па } Q_1 = x_{([2.5+1])} = x_{(3)} = 3.43$$

- Втор квартал

Бројот на елементи во примерокот е парен, па медијаната е просек на двата средни (5-тиот и 6-тиот) елементи во подредениот примерок.

$$Q_2 = \frac{3.96 + 3.83}{2} = 3.895 \text{ (= медијаната)}$$

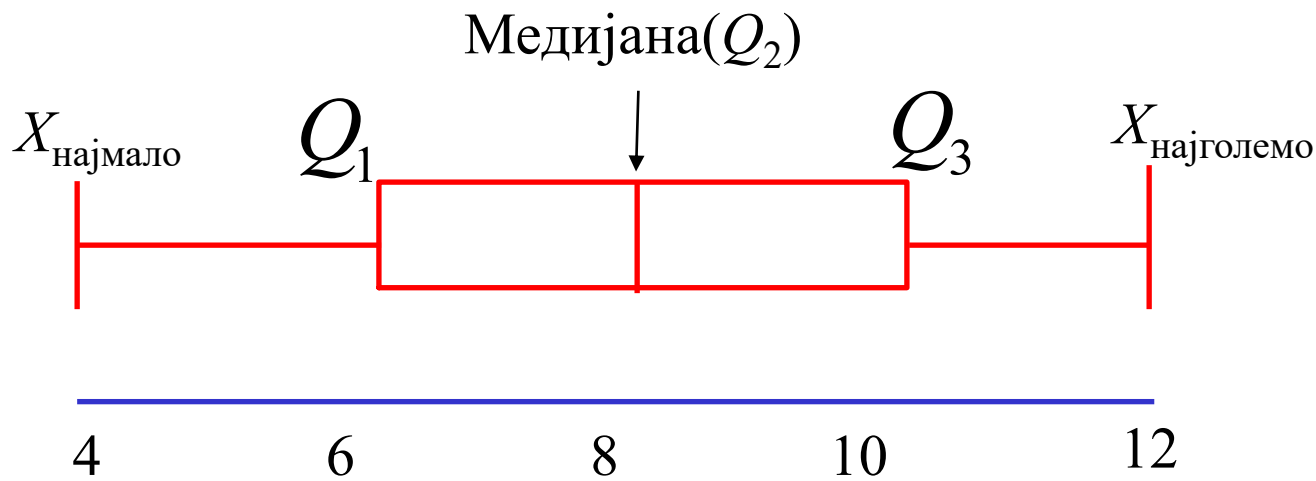
- Трет квартал

$$i = \frac{75}{100} \cdot 10 = 7.5, \text{ па } Q_3 = x_{([7.5+1])} = x_{(8)} = 6.19.$$



# Графички приказ на 5-те карактеристични броеви

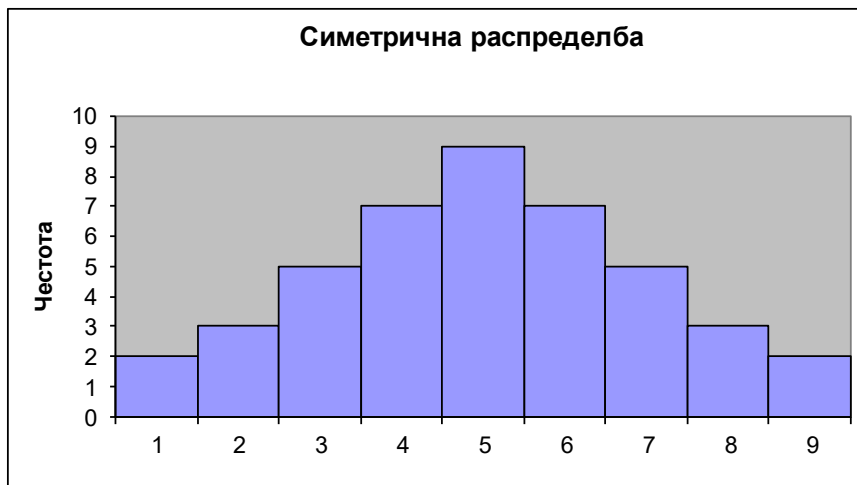
- Box-and-whisker приказ
  - Графички приказ на 5-те карактеристични броеви





# Облик на распределбата

- Велиме дека обликот на распределбата е **симетричен** ако набљудувањата се балансирани, односно рамномерно распределени околу центарот.

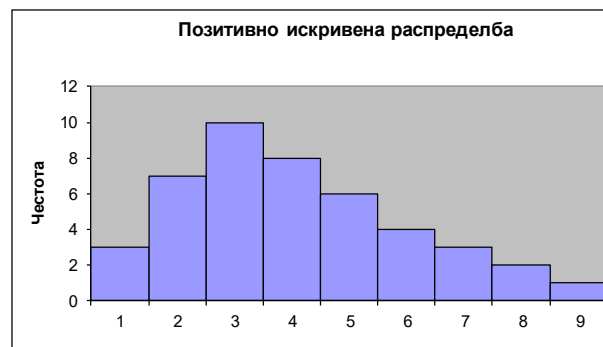




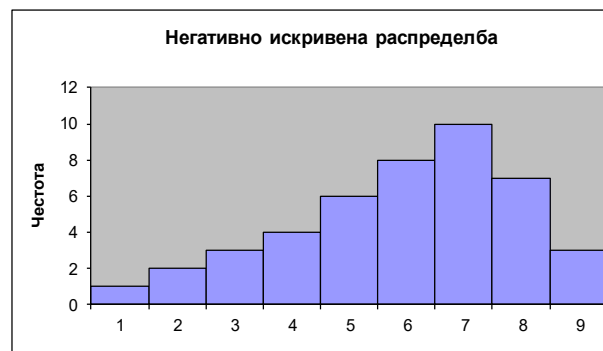
# Облик на распределбата

- Велиме дека обликот на распределбата е **искосен** (искривен) ако набљудувањата не се рамномерно распределени околу центарот.

Позитивно искосената распределба (искосена на десно) има продолжена опашка на десната страна (во насока на позитивните вредности).



Негативно искосената распределба (искосена на лево) има опашка на левата страна (во насока на негативните вредности).



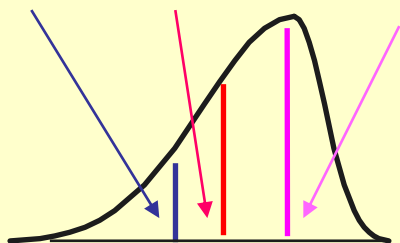


# Облик на распределба

- Опишува како се распределени податоците.
  - Симетричен или искосен

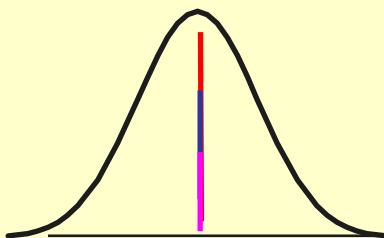
## Лево искосена

Просек < Медијана < Мода



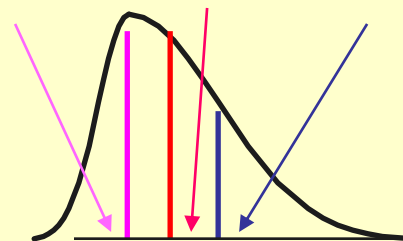
## Симетрична

Просек = Медијана = Мода



## Десно искосена

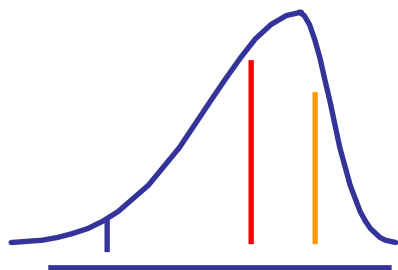
Мода < Медијана < Просек





# Облик на распределба и Box-and-Whisker приказ

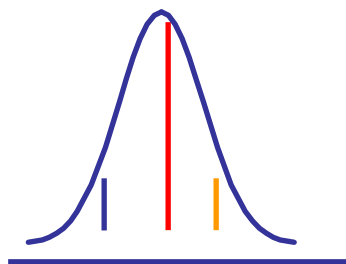
Лево искосен



$Q_1$   $Q_2$   $Q_3$



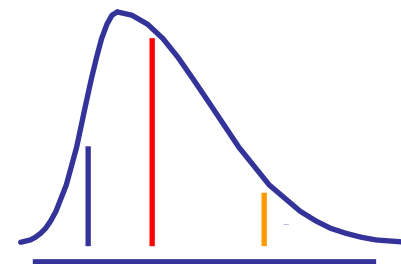
Симетричен



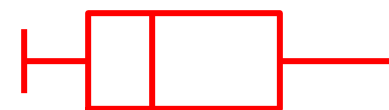
$Q_1$   $Q_2$   $Q_3$



Десно искосен

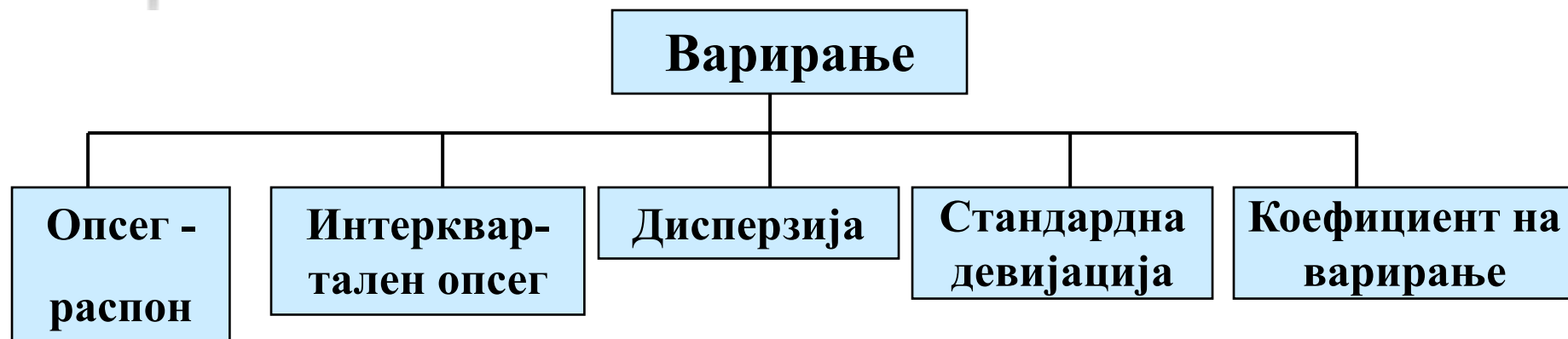


$Q_1$   $Q_2$   $Q_3$

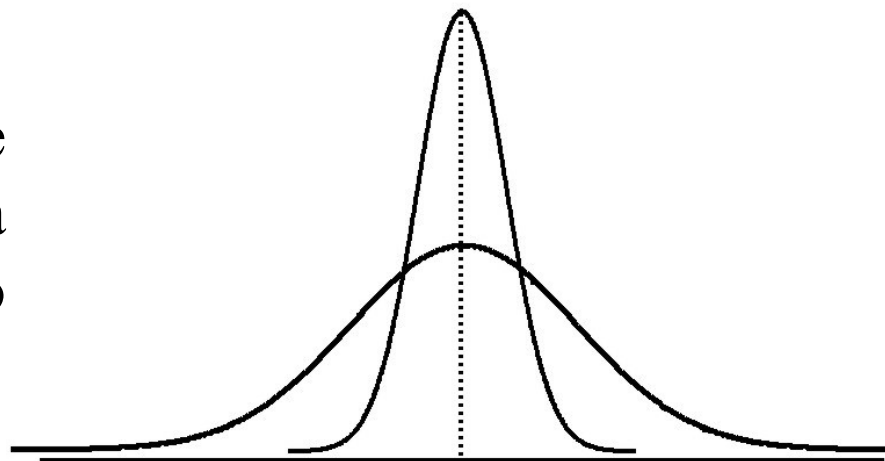




# Мерки на варирање



- Мерките на варирање даваат информации за распространетоста односно варирањето на податоците.



ист центар,  
различно варирање

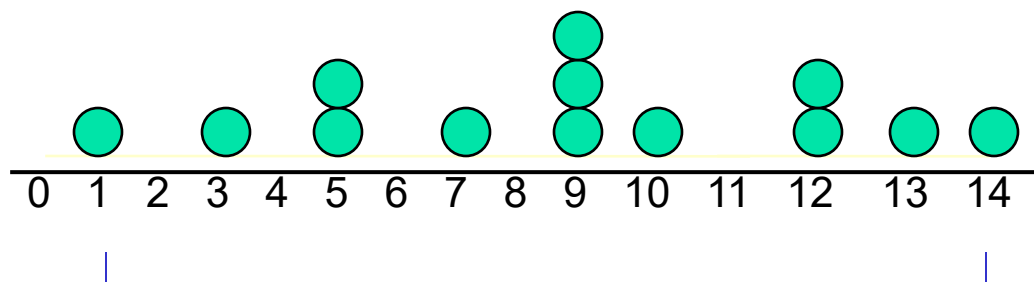


## Опсег (или распон или ранг)

- Ова е наједноставна мерка на варирање.
- Опсегот (т.е. рангот) се пресметува како разлика помеѓу најголемата и најмалата набљудувана вредност, т.е. со:

$$\text{Опсег} = x_{\text{најголем}} - x_{\text{најмал}}$$

**Пример:**



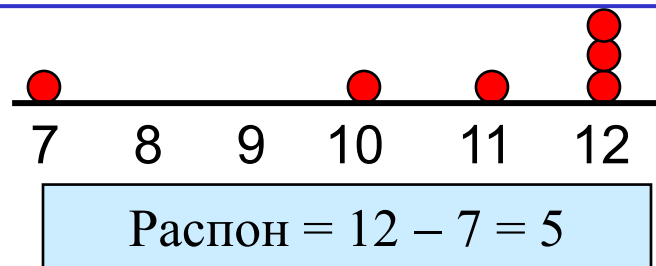
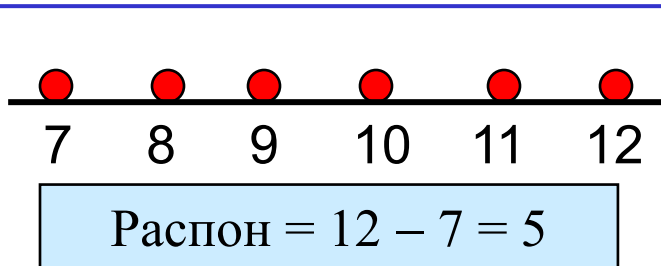
$$\text{Распон} = 14 - 1 = 13$$





## Недостатоци на опсегот

- Ја игнорира распределбата на податоците



- Осетлив на екстремни вредности

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

$$\text{Опсег} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Опсег} = 120 - 1 = 119$$



# Интерквартален распон

- Интеркварталниот распон ( $IQR$ ) е мерка на варирање на податоците.
- Познат е и како среден распон
  - Распон на средните 50% од податоците во подредениот примерок.
- Се пресметува како разлика меѓу третиот и првиот квартал.
- На интеркварталниот распон, не му влијаат екстремните вредности.

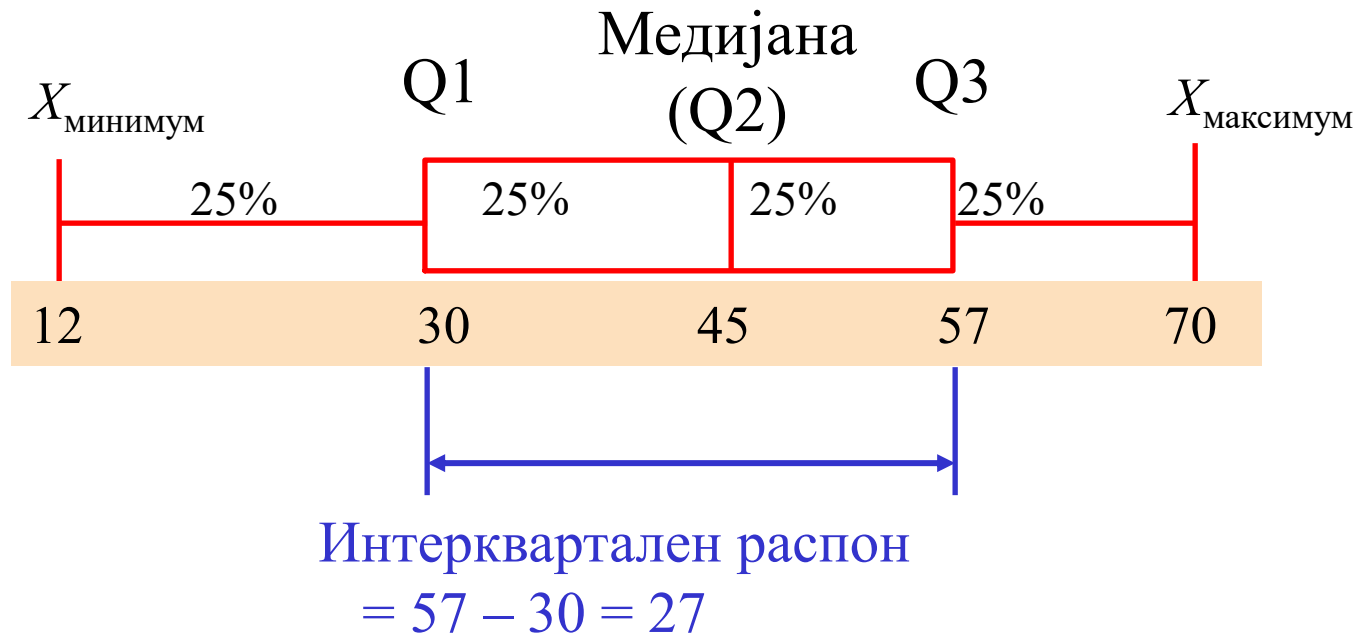
**Подредени податоци: 11 12 13 16 16 17 17 18 21**

$$IQR = Q_3 - Q_1 = 17 - 13 = 4$$



# Интерквартален распон

Пример:





## Пример

Да се определат распонот и интеркварталниот распон за податоците од примерот со производители на автомобили.

Производител	Производство (во милиони)
Fiat S.p.A.	2.62
Chrysler LLC	2.68
PSA/Peugeot-Citreon SA	3.43
Nissan Motor Co.	3.68
Honda Motor Co. Ltd.	3.83
Hyundai-Kia Automotive Group	3.96
Ford Motor Co.	5.96
Volkswagen AG	6.19
General Motors	8.90
Toyota Motor Corp.	9.37

- Претходно ги определивме квантилите

$$Q_1 = 3.43, \quad Q_3 = 6.19$$

- Распон

$$9.37 - 2.62 = 6.75$$

- Интерквартален распон

$$Q_3 - Q_1 = 6.19 - 3.43 = 2.76$$



# Екстремни вредности

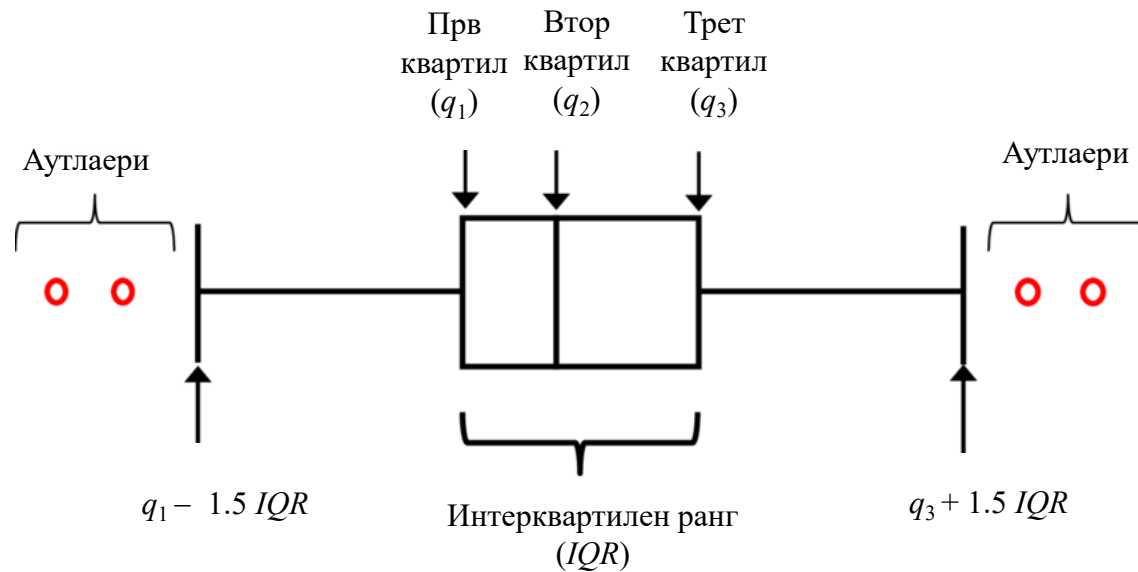
- Интеркварталниот ранг ( $IQR$ ) на податоците се користи за определување на екстремни вредности (аутлаери) од податоците.
- Како што претходно спомнавме, екстремни вредности се оние кои значајно отстапуваат од другите вредности во примерокот.
- Прашањето е како тие може да се определат?
- Еден податок ќе сметаме дека е екстреман, т.е. значајно отстапува од другите, ако е на растојание поголемо од  $1.5 \cdot IQR$  лево од првиот, односно десно од третиот квартал, т.е. податок кој не е во интервалот

$$(Q_1 - 1.5 IQR, Q_3 + 1.5 IQR).$$



# Екстремни вредности

$$(Q_1 - 1.5 IQR, Q_3 + 1.5 IQR)$$





## Пример

Да се најдат квантилите за следниот примерок и да се провери дали има екстремни вредности.

70 36 43 69 82 48 34 62 35 15

59 139 46 37 42 30 55 56 36 82

38 89 54 25 35 24 22 9 56 19

**Решение:** Прво треба да се подредат податоците

9 15 19 22 24 25 30 34 35 35

36 36 37 38 42 43 46 48 54 55

56 56 59 62 69 70 82 82 89 139



## Пример: продолжение

**Решение:** Прво треба да се подредат податоците

9 15 19 22 24 25 30 **34** 35 35  
36 36 37 38 42 43 46 48 54 55  
56 56 **59** 62 69 70 82 82 89 139

- $n = 30$  е парен број, па медијана е  
 $Me = (x_{(15)} + x_{(16)})/2 = (42+43)/2 = 42.5$ .

- I квартил.

$$i = (25/100) \cdot 30 = 7.5$$

- $Q_1 = P_{25} = x_{([7.5+1])} = x_{(8)} = 34$

па  $Q_1$  е 8-миот елемент од подредениот примерок, т.е.  $Q_1=34$ .

- III квартил.

$$i = (75/100) \cdot 30 = 22.5$$

- $Q_3 = P_{75} = x_{([22.5+1])} = x_{(23)} = 59$

па  $Q_3$  е 23-миот елемент од подредениот примерок, т.е.  $Q_3=59$ .





## Пример: продолжение

Дали во примерокот од примерот има екстремни вредности?

За да се определат екстремните вредности (ако постојат), најпрво се пресметува *интеркварталното растојание*:

$$IQR = Q_3 - Q_1 = 59 - 34 = 25.$$

Интеркварталниот ранг ( $IQR$ ) на податоците се користи за определување на вредности кои отстапуваат (аутлаери) или екстремни вредности. Имено, за еден податок се смета дека е екстреман, т.е. значајно отстапува од другите ако е на растојание поголемо од  $1,5 IQR$  лево од првиот, односно десно од третиот квартал.

$$Q_1 - 1.5(IQR) = 34 - 37.5 = -3.5$$

$$Q_3 + 1.5(IQR) = 59 + 37.5 = 96.5$$

Екстремни вредности се вредностите кои се надвор од интервалот  $(-3.5, 96.5)$ . Во примерокот има една таква вредност, т.е. 139, па таа е екстремна вредност.



# Дисперзија (варијанса) на примерок

- Дисперзија на примерок се дефинира со:

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

- Во литература се сретнува и следната формула за дисперзија на примерок:

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

која понатаму ќе ја нарекуваме коригирана дисперзија на примерокот.

- Притоа,

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



# Стандардна девијација на примерок

- Највообичаена мерка за варијација.
- Го покажува варирањето на податоците во примерокот околу просекот на примерокот.
- Се мери во истата единица како и податоците.
- ***Стандардната девијација на примерок*** се дефинира како корен од дисперзијата на примерокот, т.е.

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



# Пресметување на стандардна девијација

Податоци

$(x_i)$  :

10 12 14 15 17 18 18 24

$n = 8$

Просек =  $\bar{x} = 16$

$$s = \sqrt{\frac{(10 - \bar{x})^2 + (12 - \bar{x})^2 + (14 - \bar{x})^2 + \dots + (24 - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \dots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{126}{7}}$$

=

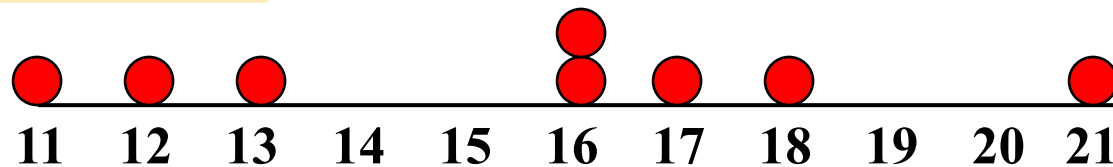
4.2426



Мерка за расејување околу  
просекот

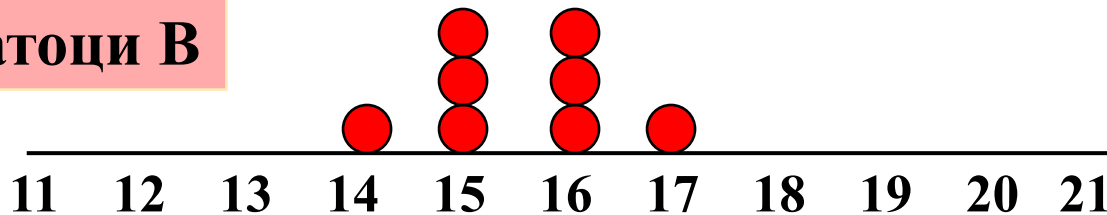
# Споредба на стандардни девијации

Податоци А



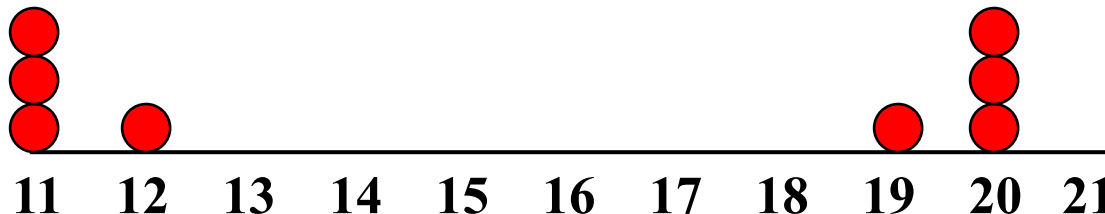
Просек = 15.5  
 $s = 3.338$

Податоци В



Просек = 15.5  
 $s = 0.9258$

Податоци С



Просек = 15.5  
 $s = 4.57$



## Пример

Да се определи дисперзија и стандардна девијација за податоците од примерот со производители на автомобили.

Производител	Производство (во милиони)
Fiat S.p.A.	2.62
Chrysler LLC	2.68
PSA/Peugeot-Citreon SA	3.43
Nissan Motor Co.	3.68
Honda Motor Co. Ltd.	3.83
Hyundai-Kia Automotive Group	3.96
Ford Motor Co.	5.96
Volkswagen AG	6.19
General Motors	8.90
Toyota Motor Corp.	9.37

- Претходно го определивме просекот

$$\bar{x} = 5.062$$

- Дисперзија

$$s^2 = \frac{(2.62^2 + 2.68^2 + 3.43^2 + \dots + 9.37^2) - 10 \cdot (5.062)^2}{10 - 1}$$

$$= 6.0345$$

- Стандардна девијација

$$s = \sqrt{s^2} = \sqrt{6.0345} = 2.4565$$



## Пример

Пресметувањето на просекот и дисперзијата може да се направи и табеларно.

Производител	$x_i$	$x_i^2$
Fiat S.p.A.	2.62	6.86
Chrysler LLC	2.68	7.18
PSA/Peugeot-Citreon SA	3.43	11.76
Nissan Motor Co.	3.68	13.54
Honda Motor Co. Ltd.	3.83	14.67
Hyundai-Kia Automotive Group	3.96	15.68
Ford Motor Co.	5.96	35.52
Volkswagen AG	6.19	38.32
General Motors	8.90	79.21
Toyota Motor Corp.	9.37	87.80
<b>вкупно</b>	<b>50.62</b>	<b>310.55</b>

$$\bar{x} = \frac{50.62}{10} = 5.062$$

$$s^2 = \frac{310.55 - 10 \cdot (5.062)^2}{10 - 1} = 6.0345$$

$$s = \sqrt{s^2} = \sqrt{6.0345} = 2.4565$$



# Коефициент на варирање

- Коефициентот на варирање е мерка на релативното варирање на податоците.
- Се прикажува во проценти (%)
- Покажува релативно варирање (споредба) на стандардната девијација во однос на просекот.
- Се користи за споредба на две или повеќе групи податоци мерени во различни единици.

$$CV = \left( \frac{s}{\bar{x}} \right) 100\%$$





# Споредба на коефициентите на варирање


- Производ А:
  - Просечна цена минатата година = 50 ден.
  - Стандардна девијација = 5 ден.

- Производ В:
  - Просечна цена минатата година = 100 ден
  - Стандардна девијација = 5 ден

- Коефициент на варирање:

- Производ А:  $CV = \frac{s}{\bar{x}} = \left( \frac{5 \text{ ден}}{50 \text{ ден}} \right) 100\% = 10\%$

- Производ В:  $CV = \frac{s}{\bar{x}} = \left( \frac{5 \text{ ден}}{100 \text{ ден}} \right) 100\% = 5\%$



# Позиција на еден податок во однос на групата

- Ако сакаме да ја определиме релативната позицијата на една вредност од примерокот во однос на целата група, мериме колку стандардни девијации отстапува податокот од просекот на целиот примерок.
- Релативната позиција се определува со следната формула

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Ако  $z_i < 0$ , тогаш податокот  $x_i$  е помал од просекот, ако  $z_i > 0$  тогаш податокот е поголем од просекот.



# Пресметување на релативната позиција

Податоци

$(x_i)$  : 10 12 14 15 17 18 18 24

$n = 8$  просек:  $\bar{x} = 16$ , стандардна дев.  $s = 4.2$

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{x_i - 16}{4.2}$$

$$z_1 = \frac{10 - 16}{4.2} = \frac{-6}{4.2} = -1.43$$

$$z_3 = \frac{14 - 16}{4.2} = \frac{-2}{4.2} = -0.48$$

$$z_5 = \frac{17 - 16}{4.2} = \frac{1}{4.2} = 0.24$$

$$z_7 = \frac{18 - 16}{4.2} = \frac{2}{4.2} = 0.48$$

$$z_2 = \frac{12 - 16}{4.2} = \frac{-4}{4.2} = -0.95$$

$$z_4 = \frac{15 - 16}{4.2} = \frac{-1}{4.2} = -0.24$$

$$z_6 = \frac{18 - 16}{4.2} = \frac{2}{4.2} = 0.48$$

$$z_8 = \frac{24 - 16}{4.2} = \frac{8}{4.2} = 1.9$$



# Коефициент на корелација

- Ја мери јачината на линеарната врска меѓу две квантитативни променливи.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Изразот за  $r$  може да се запише и во некој од следните облици:

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X \cdot s_Y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{(n-1)s_X \cdot s_Y} \end{aligned}$$



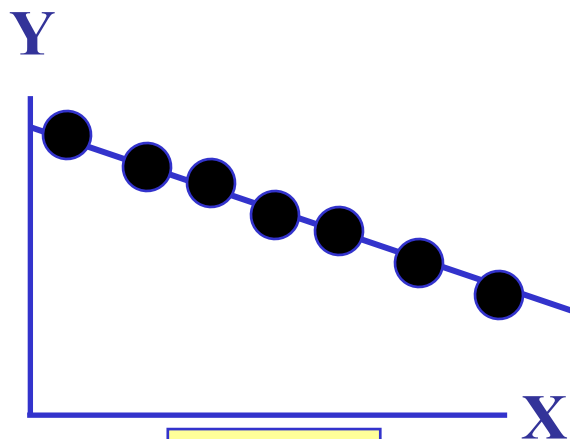
# Својства на коефициентот на корелација

---

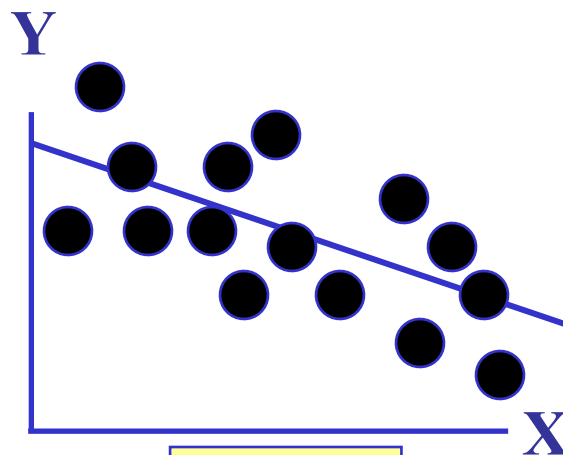
- Коефициентот на корелација е слободен од мерни единици.
- Вредностите на овој коефициент се помеѓу  $-1$  и  $1$ .
- Колку е поблиску до  $-1$ , толку посилна негативна линеарна поврзаност.
- Поблиску до  $1$ , посилна линеарна позитивна поврзаност.
- Поблиску до  $0$ , послаба линеарна поврзаност.



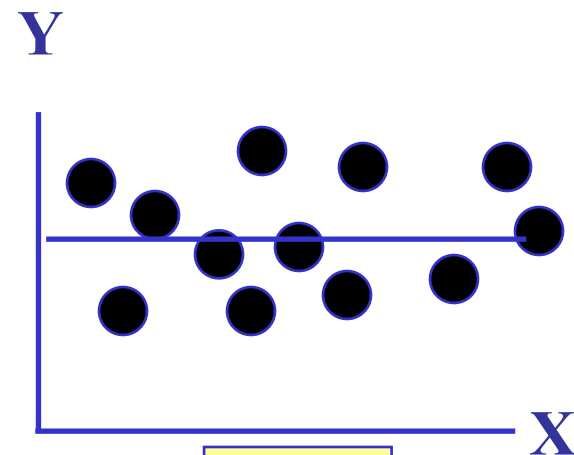
# Дијаграм на расејување за различни вредности на коефициентот на корелација



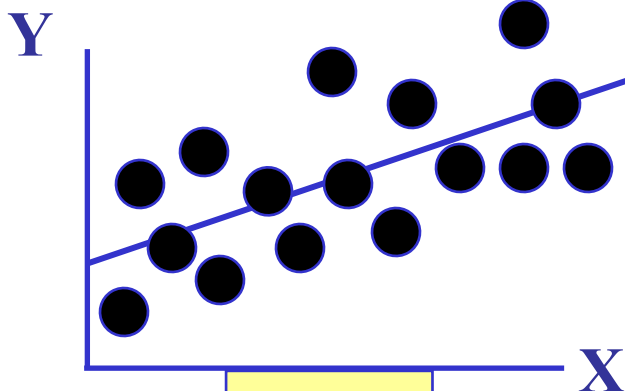
$$r = -1$$



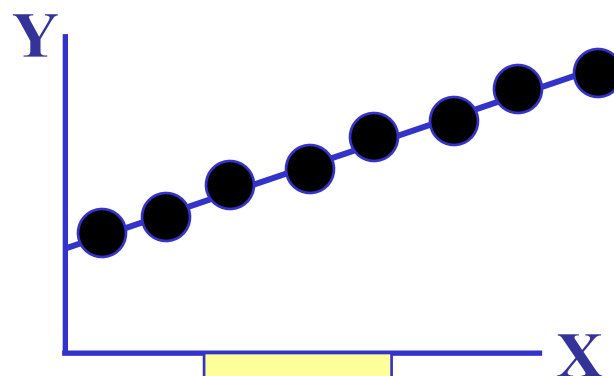
$$r = -.6$$



$$r = 0$$



$$r = .6$$



$$r = 1$$



# Тежински просек

- Тежински просек на множество податоци се дефинира со

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i},$$

каде  $w_i$  е тежината на  $i$  тогто набљудување.

- Се користи кога податоците имаат различна важност (тежина).



# Тежински просек

- Пример: Сакате да купите нов фотоапарат, при што одлучувате според следниот систем на одлучување:
  - Квалитет на слика 50%
  - Животен век на батеријата 30%
  - Опсег на зумирање 20%
- Се двоумите помеѓу **Sony** и **Canon**. Кој од двата модела ќе го изберете, ако знаете дека:
  - **Sony** добил оценка 8 за квалитет на слика, 6 за животен век на батеријата и 7 за опсег на зумирање
  - **Canon** добил оценка 9 за квалитет на слика, 4 за животен век на батеријата и 6 за опсег на зумирање
- **Sony**:  $(0.5 \cdot 8 + 0.3 \cdot 6 + 0.2 \cdot 7) / (0.5 + 0.3 + 0.2) = 7.2$
- **Canon**:  $(0.5 \cdot 9 + 0.3 \cdot 4 + 0.2 \cdot 6) / (0.5 + 0.3 + 0.2) = 6.9$
- Бидејќи **Sony** има повисока просечна оценка според вашите критериуми, изборот е **Sony**.





## Пресметување за групирани податоци

Да претпоставиме дека множеството на податоци содржи  $k$  различни вредности  $m_1, m_2, \dots, m_k$ , кои се појавуваат со честоти  $f_1, f_2, \dots, f_k$ .

- За примерок со обем  $n$ , просекот е

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i m_i, \quad \text{каде} \quad n = \sum_{i=1}^k f_i.$$

- За примерок со обем  $n$ , дисперзијата е:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (m_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_i f_i m_i^2 - n \bar{x}^2 \right).$$



## Пример

- Земен е примерок од 16 студенти од ФИНКИ и добиени се податоци за бројот на положени испити од претходниот семестар: 3,5,2,4,0,1,3,5,2,3,2,3,3,2,4,1. Да се пресмета просек и дисперзија.
- Прво податоците ги групираме.

Бр. на пол. испити ( $m_i$ )	Честота ( $f_i$ )
0	1
1	2
2	4
3	5
4	2
5	2
<b>вкупно</b>	<b>16</b>

$$\bar{x} = \frac{0 \cdot 1 + 1 \cdot 2 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 2 + 5 \cdot 2}{16} = 2.6875$$

$$\begin{aligned} s^2 &= \frac{1}{16-1} (1 \cdot (0-2.6875)^2 + 2 \cdot (1-2.6875)^2 + \\ &\quad + 4 \cdot (2-2.6875)^2 + 5 \cdot (3-2.6875)^2 + \\ &\quad + 2 \cdot (4-2.6875)^2 + 2 \cdot (5-2.6875)^2) = \\ &= 1.9625 \end{aligned}$$



# Апроксимации за податоци групирани во интервали

Да претпоставиме дека податоците се групирани во  $k$  интервали и нека  $m_1, m_2, \dots, m_k$ , се средните точки на интервалите, а  $f_1, f_2, \dots, f_k$  честотите.

- За примерок со обем  $n$ , просекот е:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i m_i, \quad \text{каде} \quad n = \sum_{i=1}^k f_i.$$

- За примерок со обем  $n$ , дисперзијата е:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (m_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_i f_i m_i^2 - n \bar{x}^2 \right).$$



## Пример

За податоците групирани во следната табела да се пресмета просек и дисперзија.

Инт.	Фрек.	средна точка
[1, 3)	4	2
[3, 5)	12	4
[5, 7)	13	6
[7, 9)	19	8
[9, 11)	7	10
[11, 13]	5	12
<b>вкупно</b>	<b>60</b>	

$$\bar{x} = \frac{4 \cdot 2 + 12 \cdot 4 + 13 \cdot 6 + 19 \cdot 8 + 7 \cdot 10 + 5 \cdot 12}{60} =$$

$$= 6.93$$

$$s^2 = \frac{1}{60-1} (4 \cdot (2-6.93)^2 + 12 \cdot (4-6.93)^2 + \\ + 13 \cdot (6-6.93)^2 + 19 \cdot (8-6.93)^2 + \\ + 7 \cdot (10-6.93)^2 + 5 \cdot (12-6.93)^2) = \\ = 7.2498$$



## Пример

Пресметките и во овој случај може да се табелираат.

Инт.	$f_i$	$m_i$	$f_i m_i$	$f_i m_i^2$
[1, 3)	4	2	8	16
[3, 5)	12	4	48	192
[5, 7)	13	6	78	468
[7, 9)	19	8	152	1216
[9, 11)	7	10	70	700
[11, 13]	5	12	60	720
<b>вкупно</b>	<b>60</b>		<b>416</b>	<b>3312</b>

$$\bar{x} = \frac{416}{60} = 6.93$$

$$s^2 = \frac{3312 - 60 \cdot 6.93^2}{60 - 1} = 7.2498$$



## Етички аспекти

---

Нумеричките дескриптивни карактеристики:

- Треба да ги документираат и добрите и лошите резултати,
- Треба да бидат презентирани на фер, објективен и неутрален начин,
- Не смее да се користат несоодветни збирни мерки за да се искриват фактите.

