



Trainer: Bondor Cosmina-Ioana, PhD, MsC

Descriptive statistics

Quantitative Variables



ALWAYS



SEEK



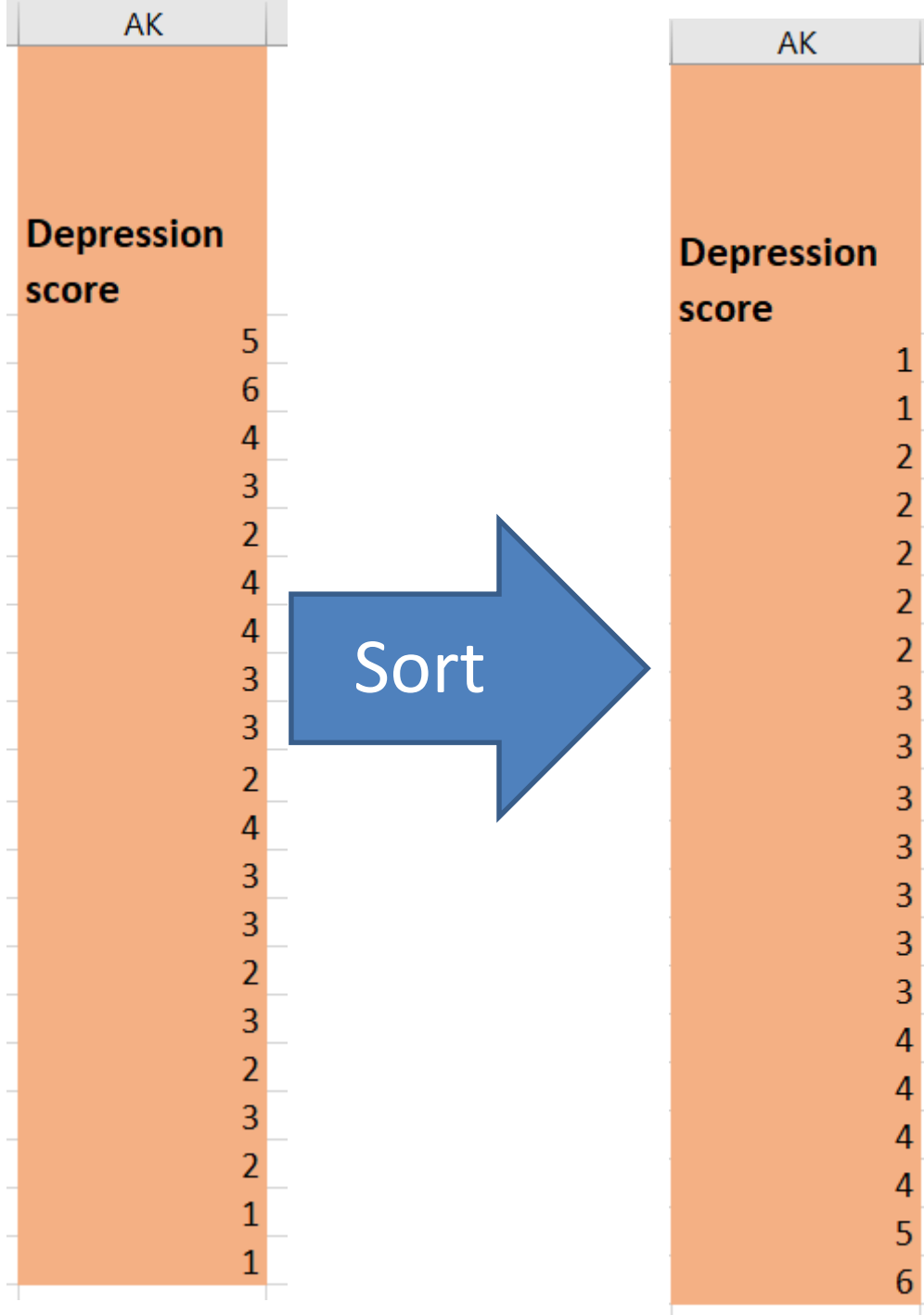
KNOWLEDGE

Objective: Measures

- Measures of location
 - Quartiles
 - Percentiles
 - Measures of central tendency
 - Arithmetic mean
 - Median
 - Mode
- Measures of spread
 - Variances
 - Standard Deviation
 - Coefficient of variation
 - Standard error

- Measures of shape
 - Skewness
 - Kurtosis
 - Measures of association
- Lecture I3
- Correlation
 - Covariance

Measures for ordinal variables



Measures for ordinal variables

Minimum = 1

Maximum = 6

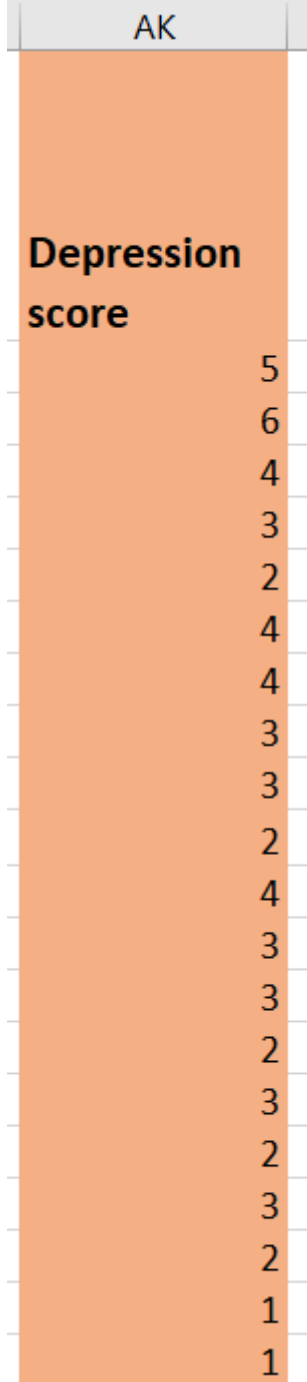
Range = Maximum – Minimum =
= 6 - 1 = 5

Summarizing Ordinal Data

Minimum – the smallest score

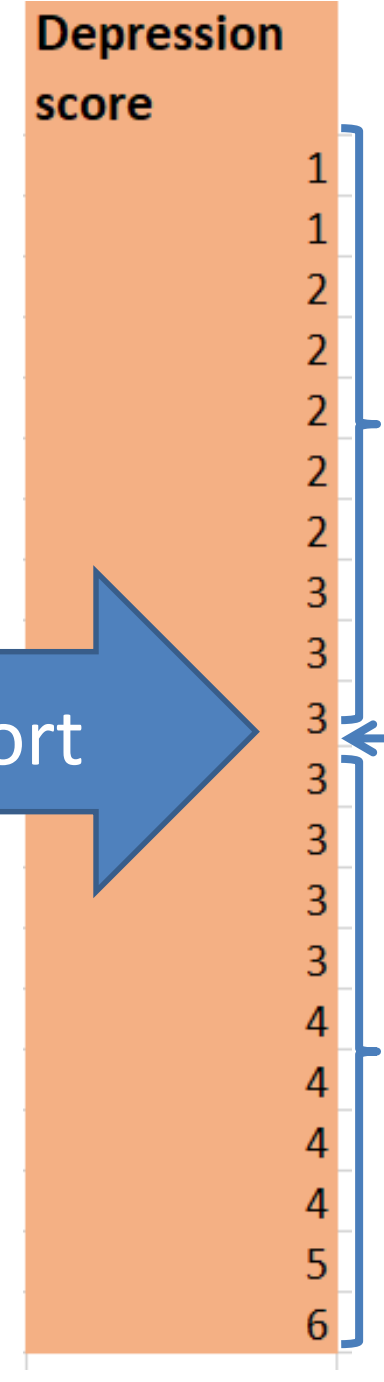
Maximum – the biggest score

Range – the difference between maximum and minimum score



20 •

Sort



10

10

middle

Median $= (3+3)/2 = 3$

Summarizing Ordinal Data

Median – the middle of observation,
the point at which half the observations are smaller and half are larger.

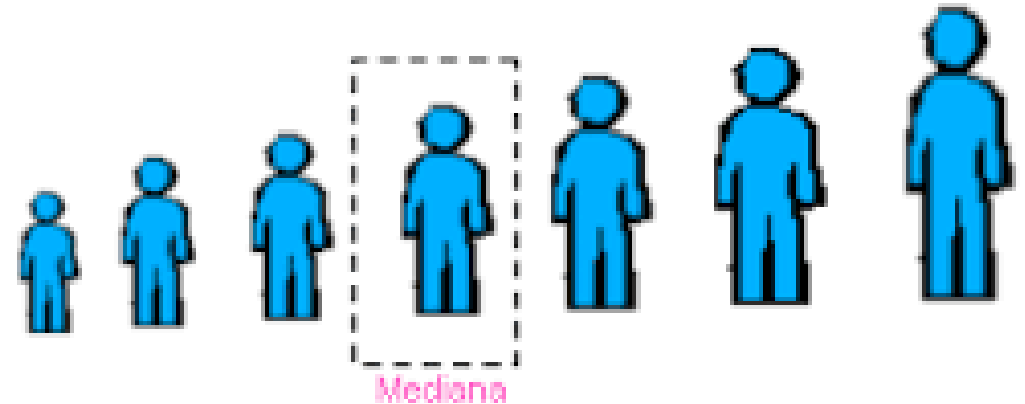
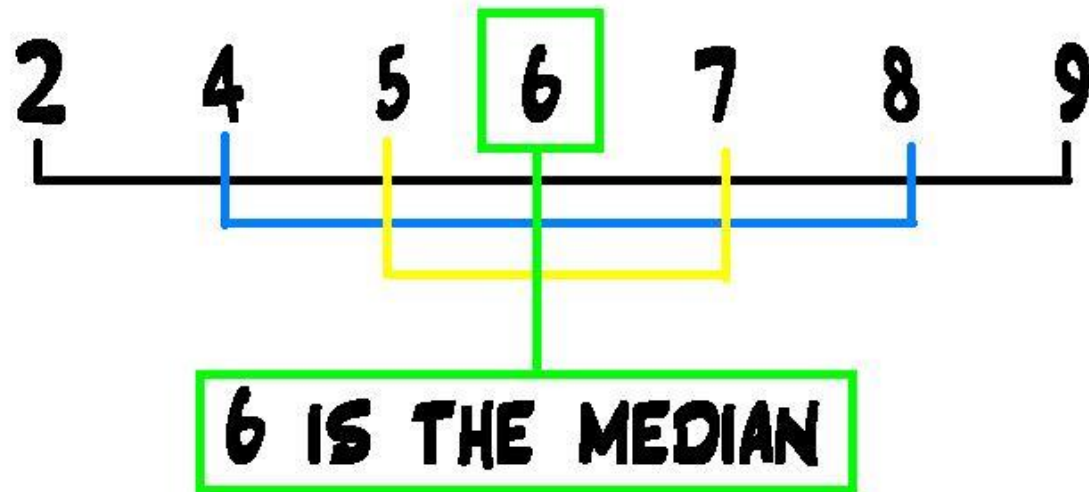
Calculation:

1. Arrange the observations from smallest to largest
2. Count in to find the middle value. The median is the middle value for an odd number of observations; it is defined as the mean of the two middle values for an even number of observations.

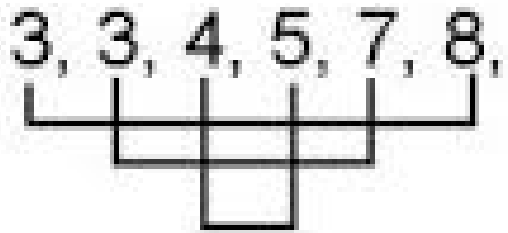
ODD number of observation

Given numbers: 8, 7, 4, 6, 5, 2, 9

Sort them than choose:



EVEN number of observation



$$4 + 5 = 9$$

$$9 / 2 = 4.5$$

4.5 = Median

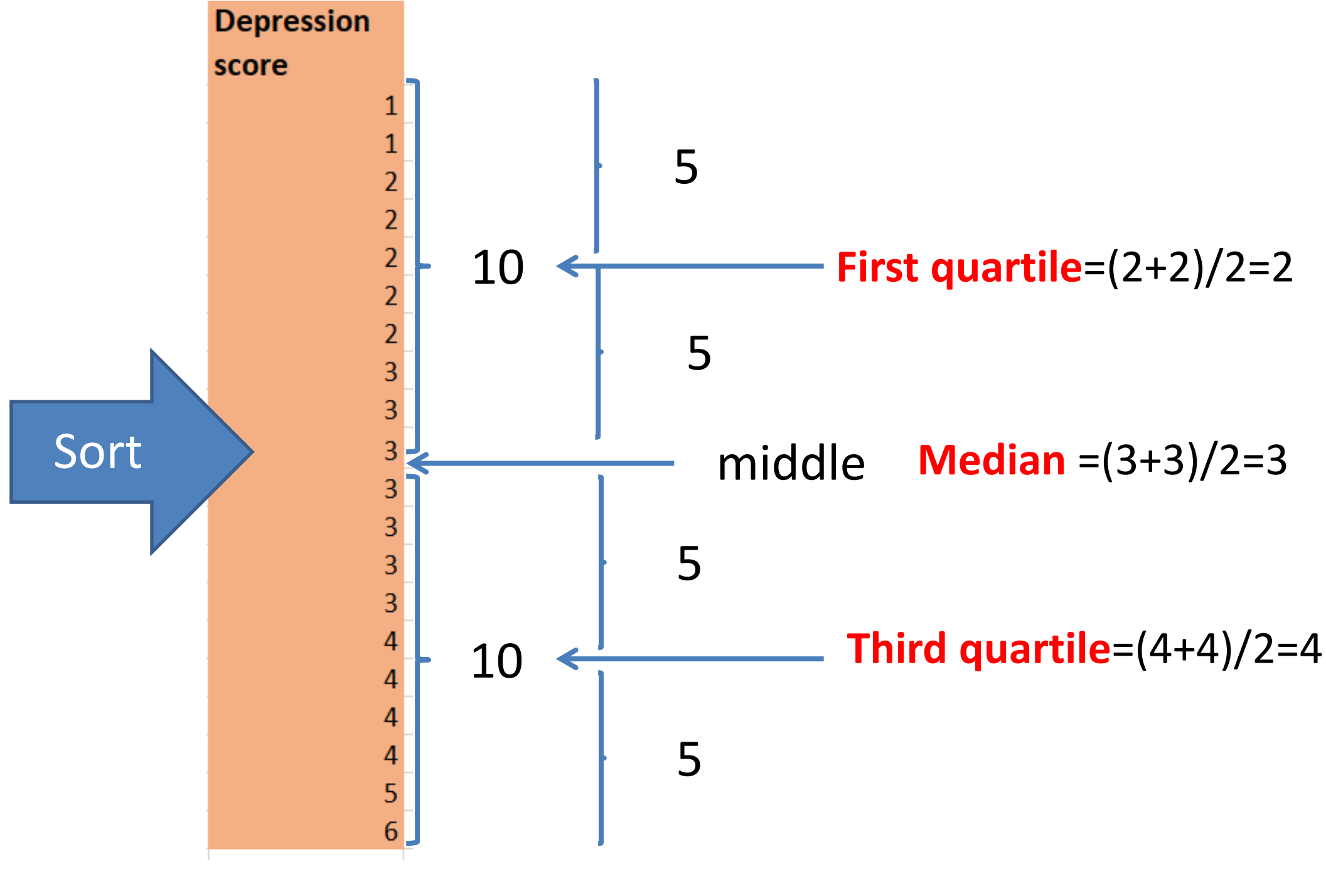
$$\frac{MED + IAN}{2}$$

Quartile/Percentile

Divide the data in two parts:

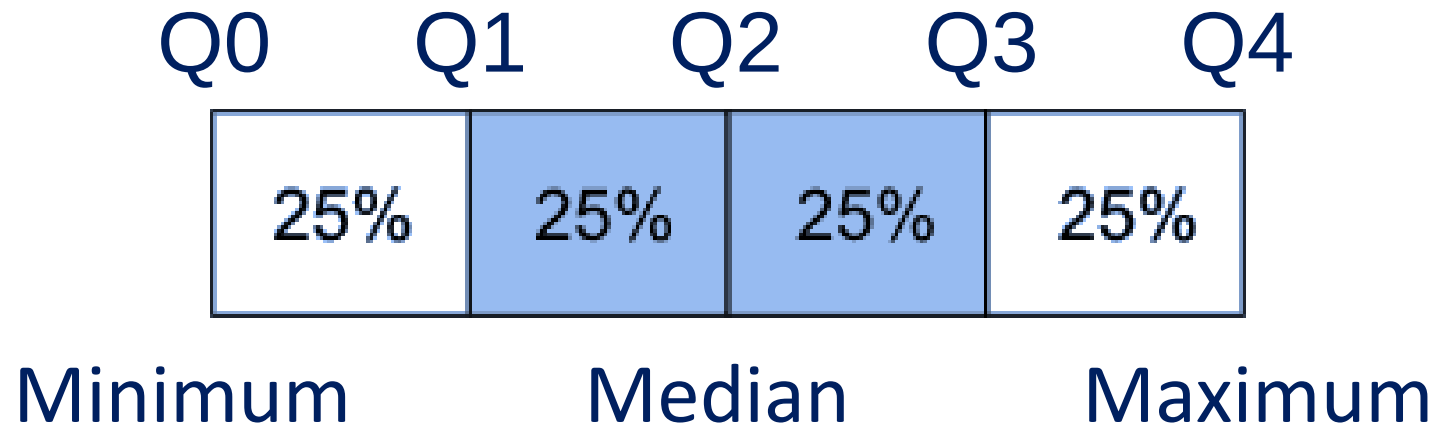
- X% less than ($<$) Quartile/Percentile
- 100%-X% higher than ($>$) Quartile/Percentile
- Example: 25th percentile=10
- 25% of data less than 10; 75% of data higher than 10
- 5 Quartiles
- 101 Percentiles

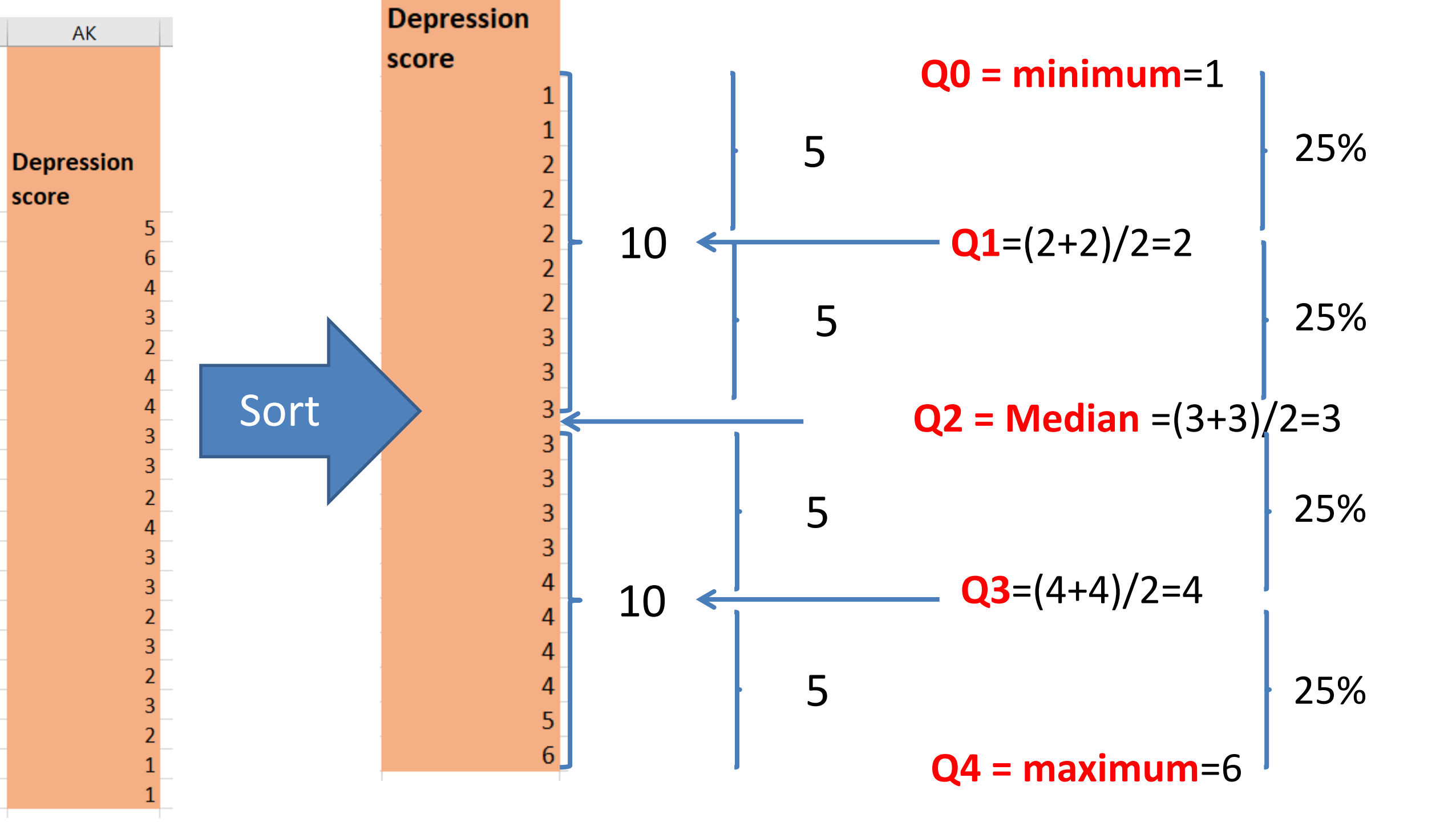
AK	Depression score
	5
	6
	4
	3
	2
	4
	4
	3
	3
	2
	4
	3
	3
	2
	3
	2
	3
	2
	1
	1



5 Quartiles

- Obs. Second quartile (Q2) = Median
- Q = quartile



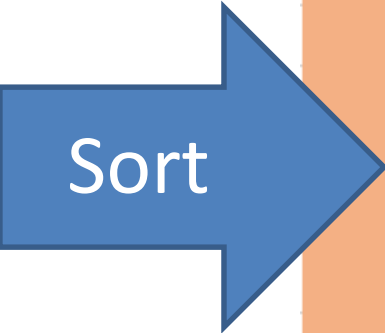


Other measures

- **Interquartile Range** - the difference between the third and first quartiles.

$$\text{IQR} = Q3 - Q1$$

AK	
Depression score	
5	
6	
4	
3	
2	
4	
4	
3	
3	
3	
2	
4	
3	
3	
3	
2	
4	
4	
3	
2	
3	
2	
1	
1	



Depression score	
1	
1	
2	
2	
2	
2	
2	
3	
3	
3	
3	
3	
3	
3	
4	
4	
4	
4	
4	
5	
6	

1
1
2
2
2
2
2
3
3
3
3
3
3
4
4
4
4
4
5
6

10

10

5

5

5

5

Q0 = 1

Q1 = $(2+2)/2 = 2$

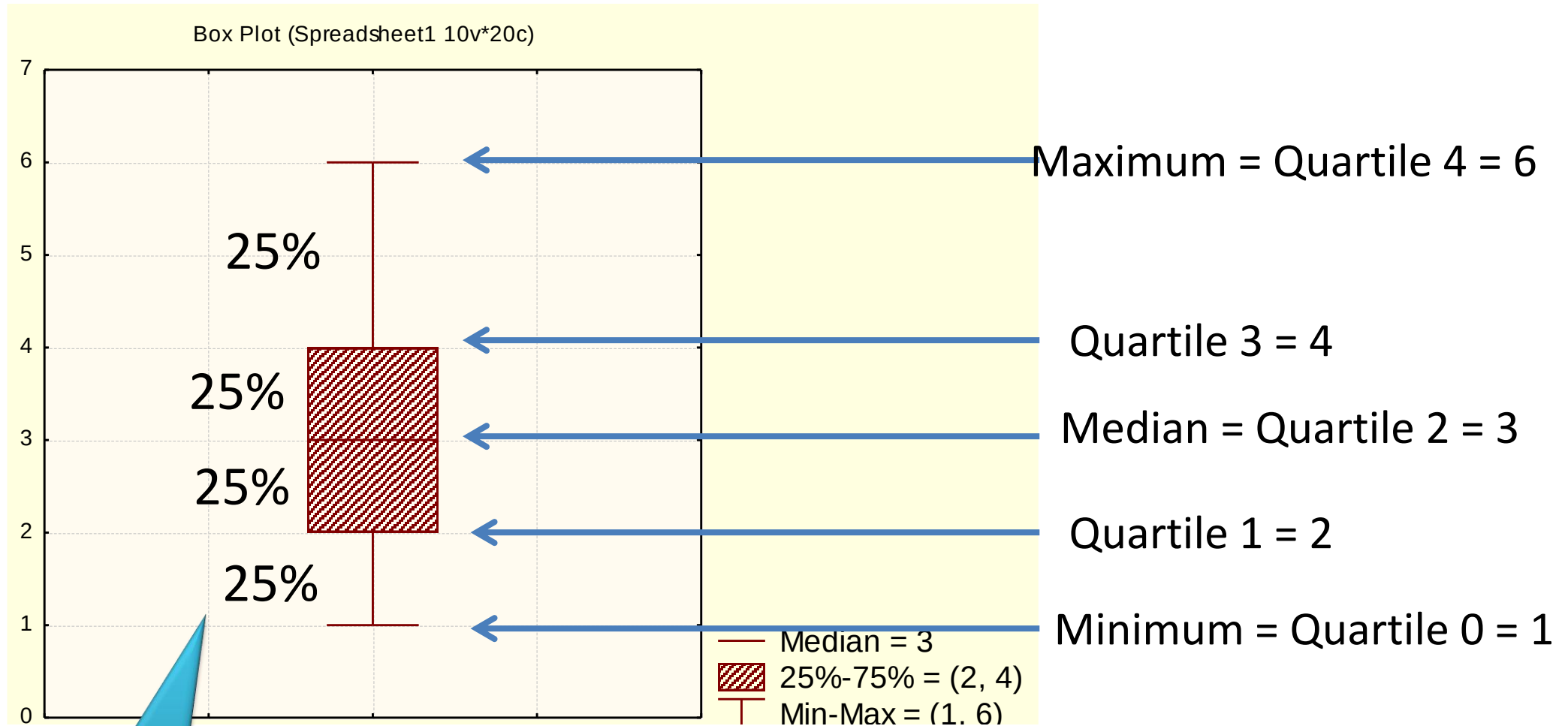
Q2 = $(3+3)/2 = 3$

Q3 = $(4+4)/2 = 4$

Q4 = 6

IQR = $4 - 2 = 2$

Chart – Box-whiskers Plot



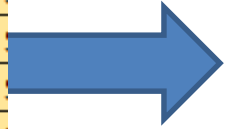
Same amount of data in each interval

Mode

- The most popular value! = Mode
- Marks for info exam for group 2: 10, 7, 5, 6, 10, 9, 8, 9, 10
- Most popular: 10 = mode

Example 2. MELD Score Descriptive measures

C	
MELD score	
5	1
1	5
6	5
9	6
11	6
8	7
25	8
30	8
32	9
14	11
8	12
7	14
6	16
5	19
22	22
25	25
29	25
12	29
16	30
19	32



Sort



Minimum = 1

Maximum = 32

Range = $32 - 1 = 31$

Median = $(11 + 12)/2 = 11.5$

Cuartile 1 = $(6 + 7)/2 = 6.5$

Cuartile 3 = $(22 + 25)/2 = 47/2 = 23.5$

IQR = $23.5 - 6.5 = 17$

Mode - none

C
MELD score
5
1
6
9
11
8
25
30
32
14
8
7
6
5
22
25
29
12
16
19



Sort



MELD score
1
5
5
6
6
7
8
8
9
11
12
14
16
19
22
25
25
29
30
32

10

10

5

5

5

5

Q0 = minimum=1

Q1 = $(6+7)/2=6.5$

Q2 = Median = $(11+12)/2=11.5$

Q3 = $(22+25)/2=23.5$

Q4 = maximum=32

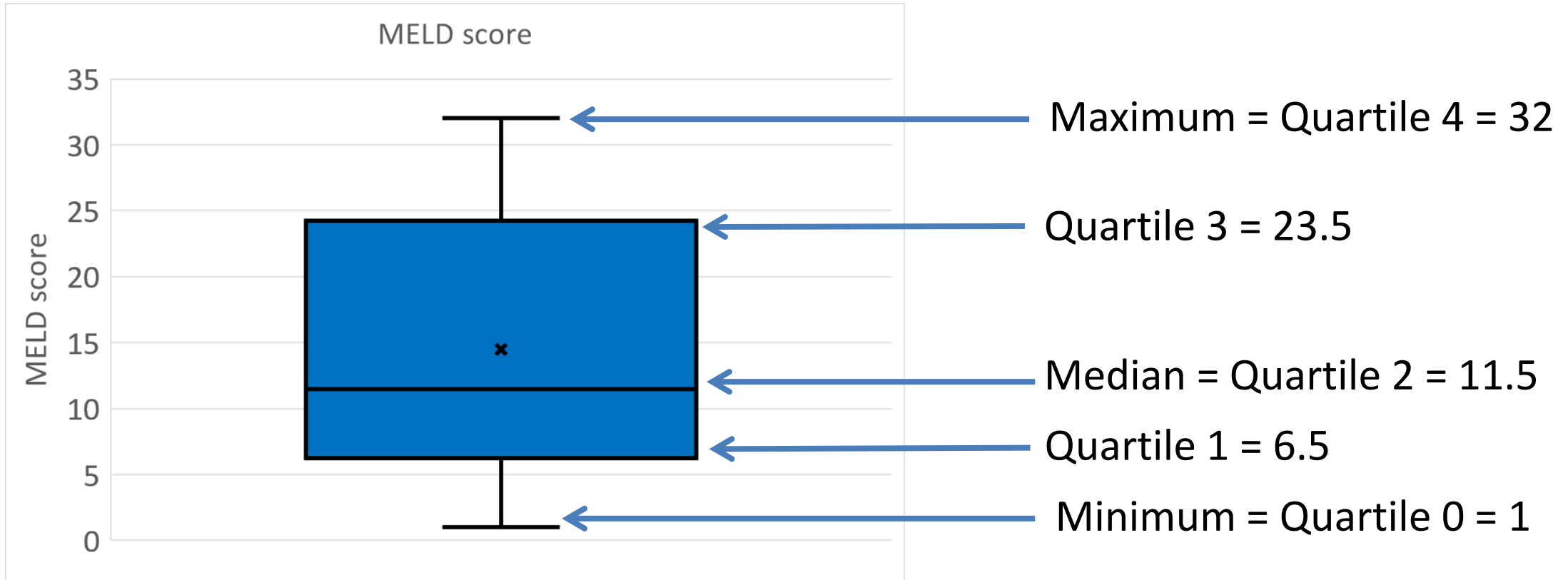
25%

25%

25%

25%

Ordinal Data – Chart – Box-whiskers Plot



Relationship between a qualitative variable
and a score (ordinal variable)

B	C
Urban/Rural	MELD score
Urban	5
Rural	1
Urban	6
Urban	9
Urban	11
Rural	8
Urban	25
Urban	30
Rural	32
Rural	14
Urban	8
Urban	7
Urban	6
Rural	5
Rural	22
Urban	25
Urban	29
Urban	12
Rural	16
Rural	19

Measures



Sort by Urban/Rural
than by MELD



Urban/Rural	MELD score
Rural	1
Rural	5
Rural	8
Rural	14
Rural	16
Rural	19
Rural	22
Rural	32
Urban	5
Urban	6
Urban	6
Urban	7
Urban	8
Urban	9
Urban	11
Urban	12
Urban	25
Urban	25
Urban	29
Urban	30

Rural
subgroup

Urban
subgroup

Measures

Urban/Rural	MELD score
Rural	1
Rural	5
Rural	8
Rural	14
Rural	16
Rural	19
Rural	22
Rural	32
Urban	5
Urban	6
Urban	6
Urban	7
Urban	8
Urban	9
Urban	11
Urban	12
Urban	25
Urban	25
Urban	29
Urban	30

Urban group

number of people = 12

Minimum = 5

Maximum = 30

Range = $30 - 5 = 25$

Median = $(9 + 11)/2 = 10$

25th percentile = Q1 = $(6 + 7)/2 = 6.5$

75th percentile = Q3 = $(15 + 25)/2 = 20$

IQR = $20 - 6.5 = 13.5$

Mode = none

Rural group

number of people = 8

Minimum = 1

Maximum = 32

Range = $32 - 1 = 31$

Median = $(14 + 16)/2 = 15$

25th percentile = Q1 = $(5 + 8)/2 = 6.5$

75th percentile = Q3 = $(19 + 22)/2 = 20.5$

IQR = $20.5 - 6.5 = 14$

Mode - none

Compare MELD score at urban group versus rural group

Chart – Box-whiskers plot

Urban/Rural	MELD score
Rural	1
Rural	5
Rural	8
Rural	14
Rural	16
Rural	19
Rural	22
Rural	32
Urban	5
Urban	6
Urban	6
Urban	7
Urban	8
Urban	9
Urban	11
Urban	12
Urban	25
Urban	25
Urban	29
Urban	30

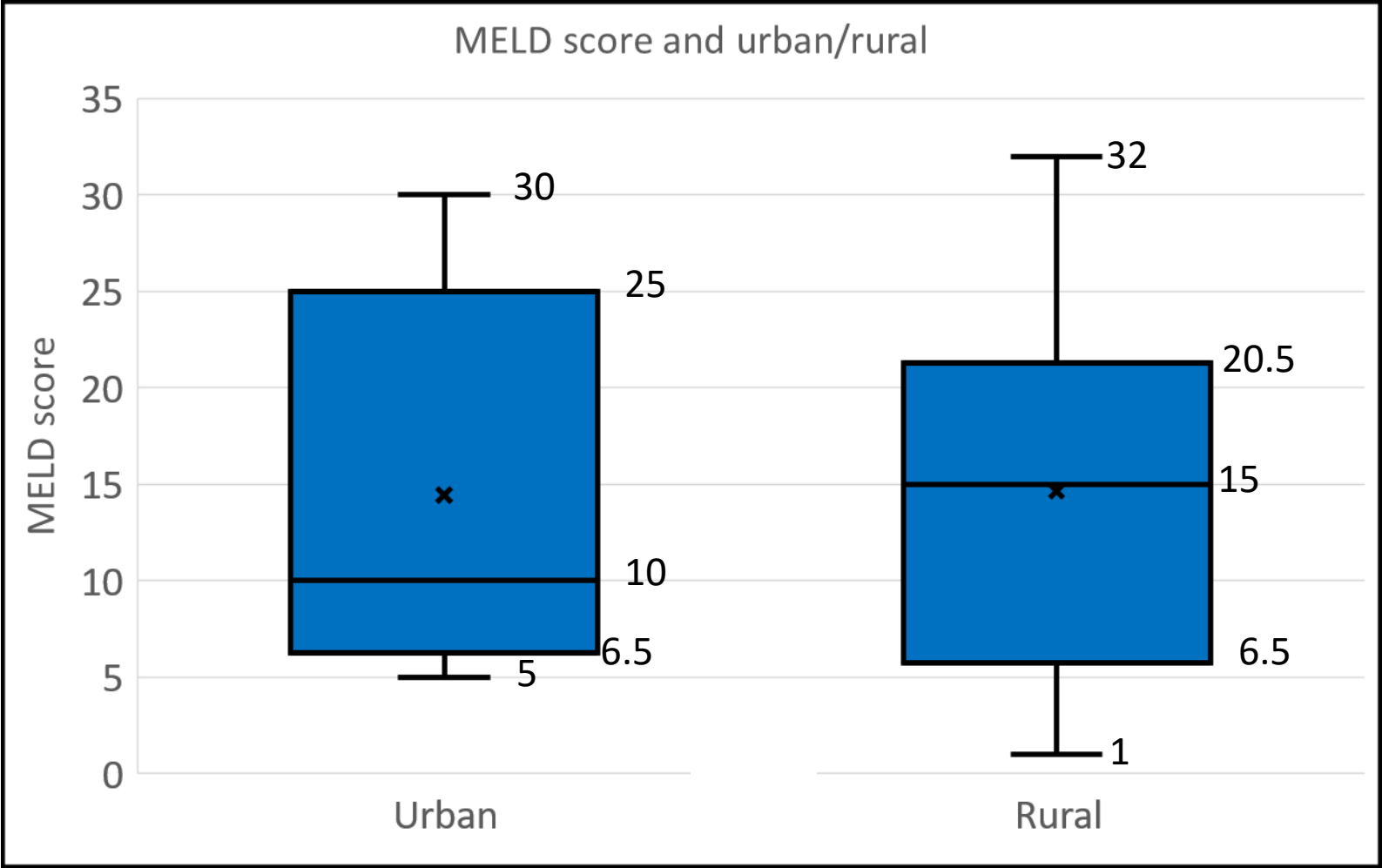
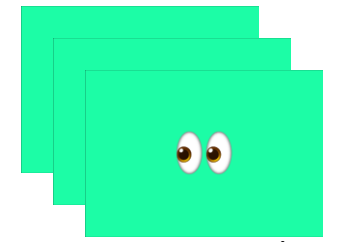


Figure. MELD score distribution in urban versus rural

Summarizing Ordinal Variables



<https://www.wooclap.com/XZLYSC>

One variable

- Parameters
 - Range
 - Median
 - Quartile 0-4
 - Interquartile range
 - Mode
- Chart
 - Box-Whiskers Plot
 - Column

One dichotomical variable and one ordinal variable

- Parameters on each subgroup
 - Range
 - Median
 - Quartile 0-4
 - Interquartile range
 - Mode
- Chart
 - Box-Whiskers Plot
 - Column/Bar
 - Staked Column/Bar
 - 100% Stacked Column/Bar



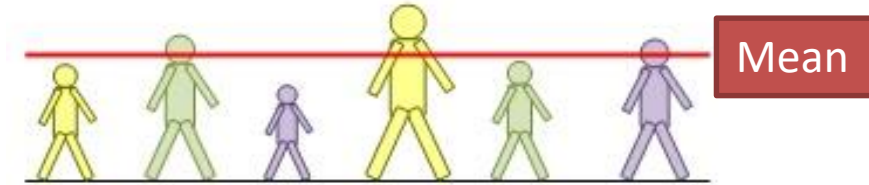
Measures for quantitative variables

Objectives

- Summarizing Quantitative (Numerical) Data
 - Descriptive statistics
 - Charts



Arithmetic mean



Arithmetic mean $\bar{X}$ (the average of the observations)

$$\bar{X} = \frac{1}{n} \sum X$$

Σ - sum

X - the individual observations

n - number of observations

Calculation: add the observations to obtain the sum and then divide by the number of observations.

Example

In 2011, 6 medical doctors earn in a month 2400, 2600, 2900, 2900, 3200, 3400 Euro.

$$\bar{X} = \frac{2400+2600+2900+2900+3200+3400}{6} = 2900 \text{ Euro}$$

The average income is 2900 Euro

Summarize quantitative data with

- **0th Quartile** = minimum
- **First Quartile**
- **Second Quartile** = median
- **3rd Quartile**
- **4th Quartile** = maximum
- Mode
- Interquartile range

Arithmetic mean is a measure of the middle

Median is a measure of the middle

Mean or median?



We can use **arithmetic mean** when data are **numerical without outliers**

Median can be use any time, but because the arithmetic mean is easy to understand ...

We use **median** only when the data are numerical with outliers or data are ordinal

Medical doctors' income in 2011

2400€ 2600 €



2900 € 2900 €



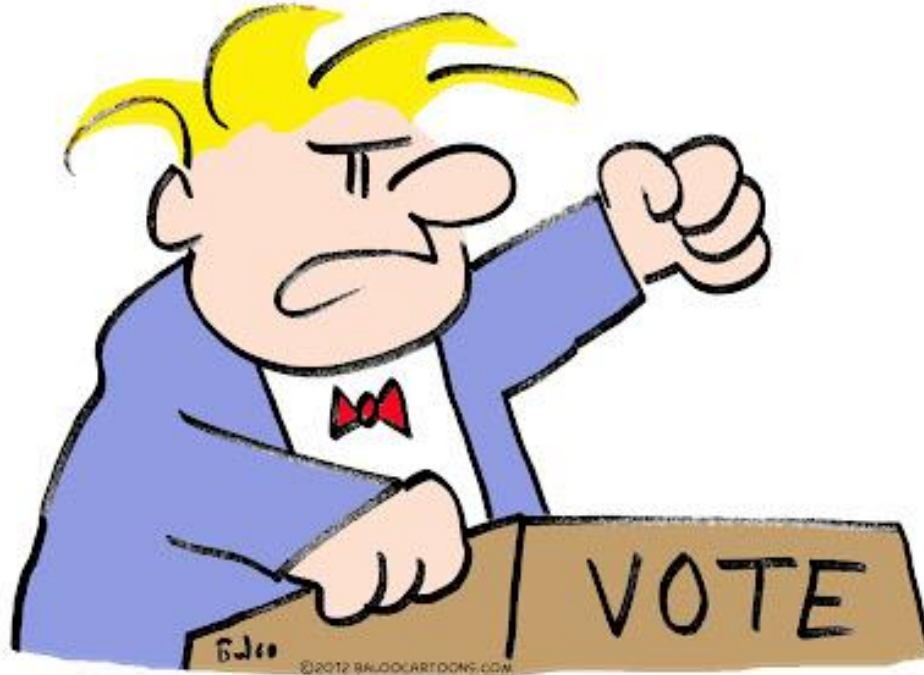
3200 € 3400 €



Arithmetic mean = 2900 €

Average income of medical doctors is 2900 €

Some political changes



"Desperate times call for
desperate politicians!"

After 12 months politician announce that:

Medical doctors' income near **double** in 12 months

2400€ 2600 €



2900 € 2900 €



3200 €



19000 €



Arithmetic mean = **5500 €**
was 2900 €

- Conclusion: outliers **can** influence the mean

Outlier 19000 €



Median

2400€ 2600 €



2900 €



2900 €

3200 €



3400 €



Median = 2900 €

Medical doctors income **after 12 months**

2400€ 2600 €



2900 €



2900 €

3200 €



19000 €



Median = **2900 €**

was 2900 €

- Conclusion: outliers **cannot** influence the median

Outlier or extreme case

19000 €



Mean is a measure of the middle?

We use **mean** if data are **numerical without outliers**

We use **median** if data are **numerical with outliers** or **data are ordinal**



Measures of Spread

Series 1	Series 2	Series 3
1	1	1
1	44	8
2	45	11
3	46	14
5	48	28
6	48	30
6	49	37
7	50	48
93	50	52
94	51	62
94	52	70
95	52	72
97	54	84
98	55	91
98	55	92
100	100	100
Sum800	Sum800	Sum800

Why we need them when we want to describe data?

Number of observations in each set = 16

Average of each set = 50

Median of each set = 50

Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

The distribution

Frequency classes		
1-20	values between 1 to 20	
21-40		
41-60		
61-80		
81-100		
Total		

minimum = 1
maximum = 100
we divide the interval 1 - 100
in 5 equal subintervals
= 5 “classes” with the length
of 20

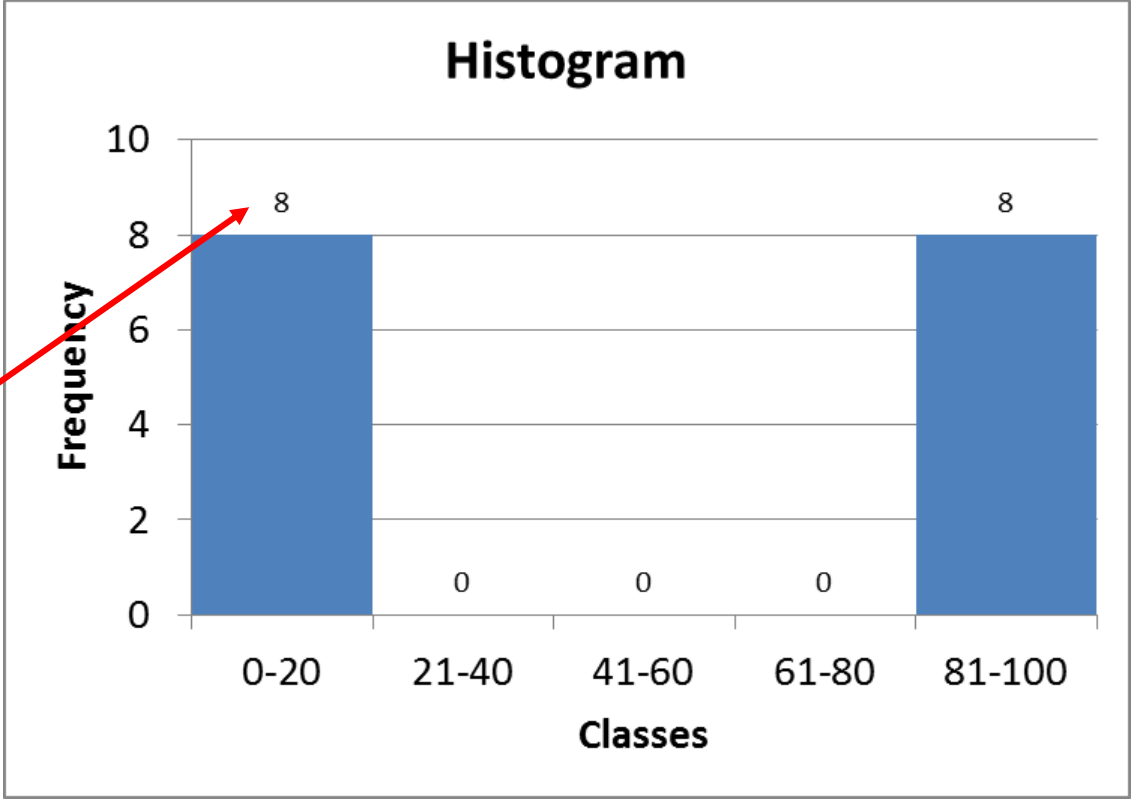
Series 1

1
1
2
3
5
6
6
7
93
94
94
95
97
98
98
100

The distribution

8 values between 0 to 20

Series 1 – Classes	Absolute Frequency	Relative Frequency %
0-20	8	50
21-40	0	0
41-60	0	0
61-80	0	0
81-100	8	50
Total	16	100



Series 1

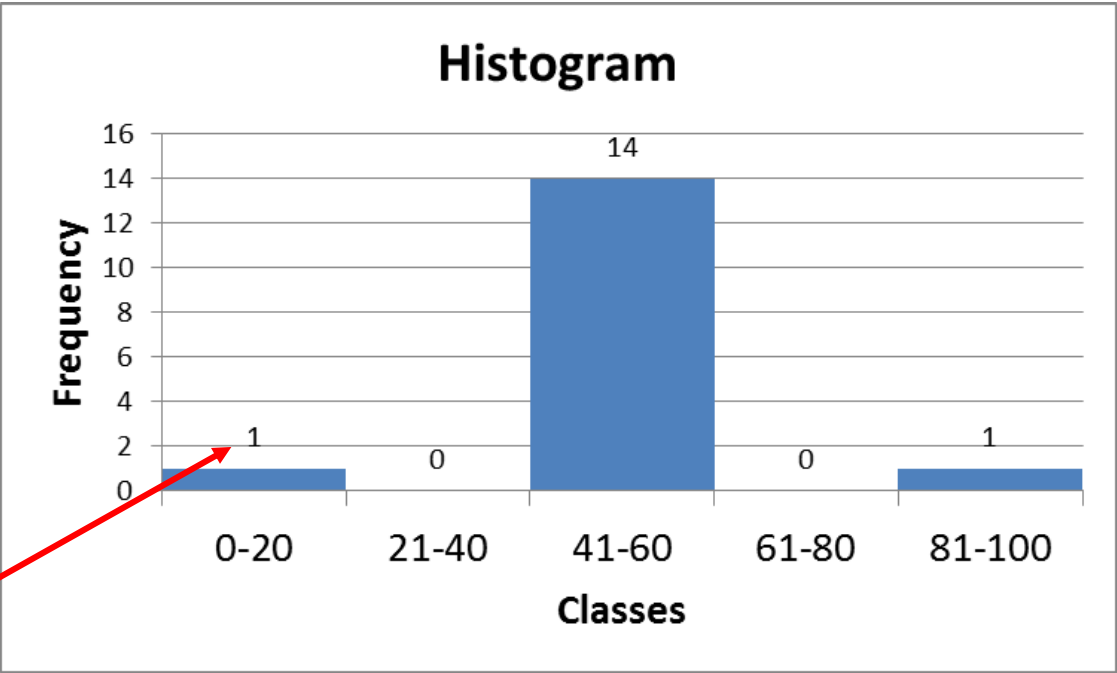
Series 2

The distribution

1 value between 0 to 20

44
45
46
48
48
49
50
50
51
52
52
54
55
55
100

Series 2 – Classes	Absolute Frequency	Relative Frequency %
0-20	1	6.25
21-40	0	0
41-60	14	87.50
61-80	0	0
81-100	1	6.25
Total	16	100



Series 2

Series 3

The distribution

1 } 2 values between 0 to 20
11

24

29

36

41

45

49

51

55

59

64

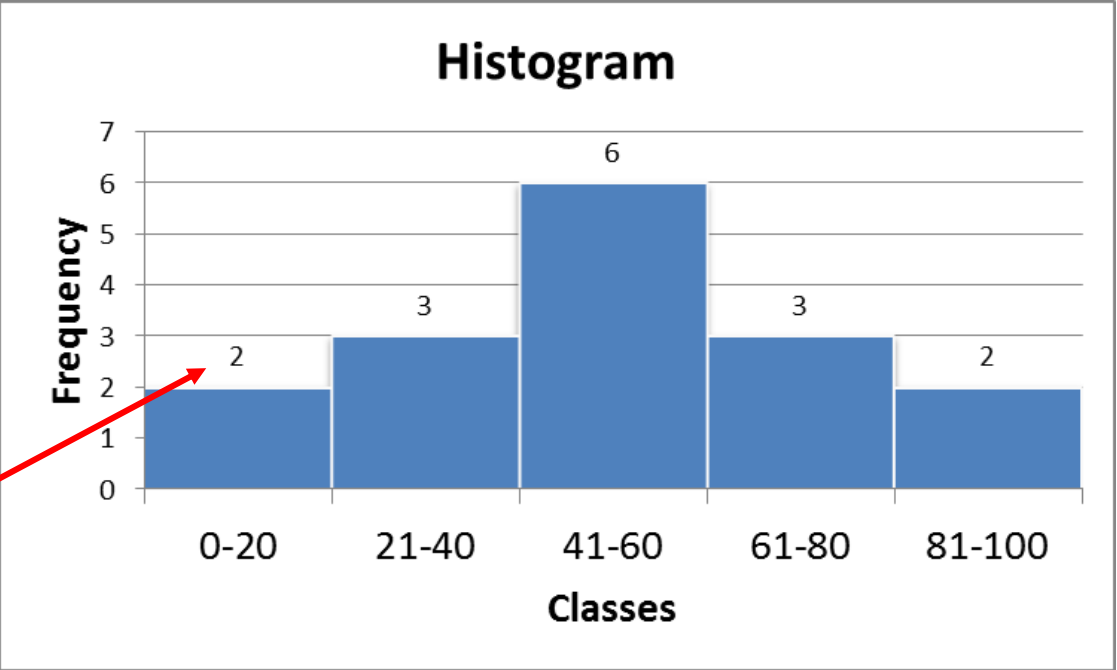
71

76

88

100

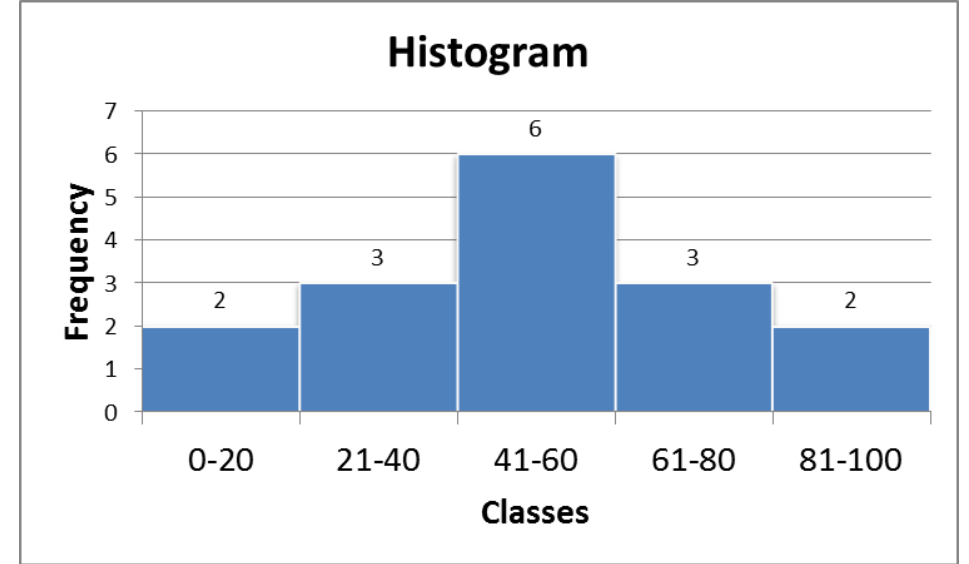
Series 3 – Classes	Absolute Frequency	Relative Frequency %
0-20	2	12.50
21-40	3	18.75
41-60	6	37.50
61-80	3	18.75
81-100	2	12.50
Total	16	100



Series 3

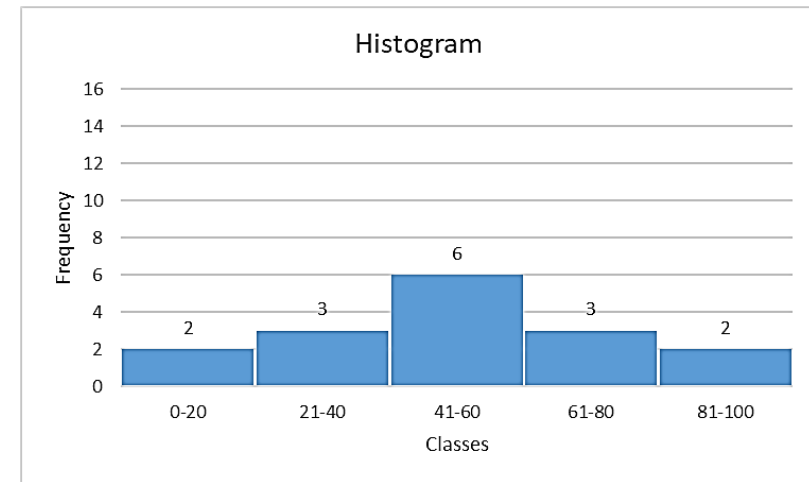
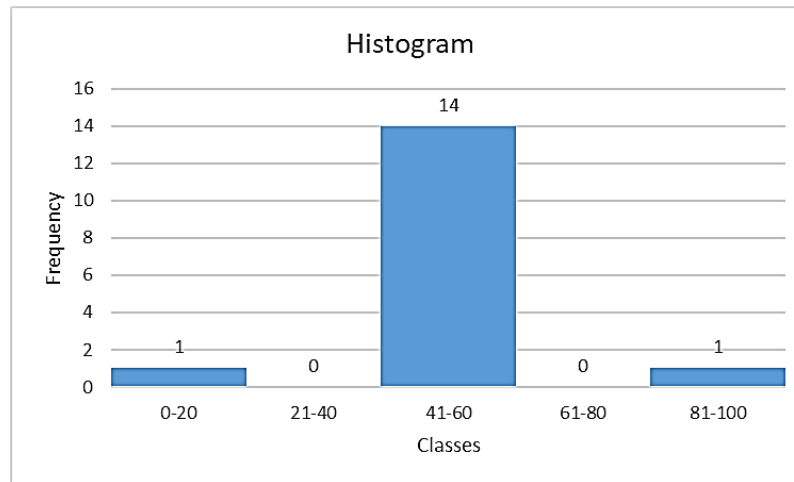
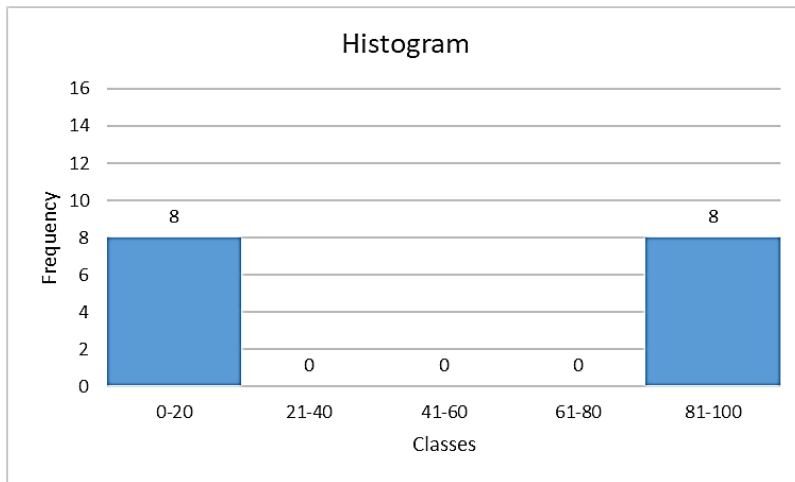
Histogram

- a graphical representation
 - the distribution of quantitative **continuous** variables
 - columns
 - show the frequency of values within ranges (bins)
 - column height = the number of values in that particular interval
 - **must have**
 - **equal** column width = equal intervals (to compare the frequencies)
 - **no gap (space)** between the columns (intervals are adjacent and cover a continuous range of values)



Histogram

- useful to see the shape and the spread of data values



the same range on Oy
equal intervals on Ox

we can compare the distributions

Measures of Spread

The Standard Deviation – the average of the deviation (departure) from the mean

Formula for population (all observation):

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

Σ means "sum of,"

X = each value in the data set,

$\bar{X}$ = mean of all values in the data set,

N = number of values in the data set.

Formula for subset of population (sample)

- The Standard Deviation of a **sample** (with correction for medical research)

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_N - \bar{X})^2}{N - 1}}$$

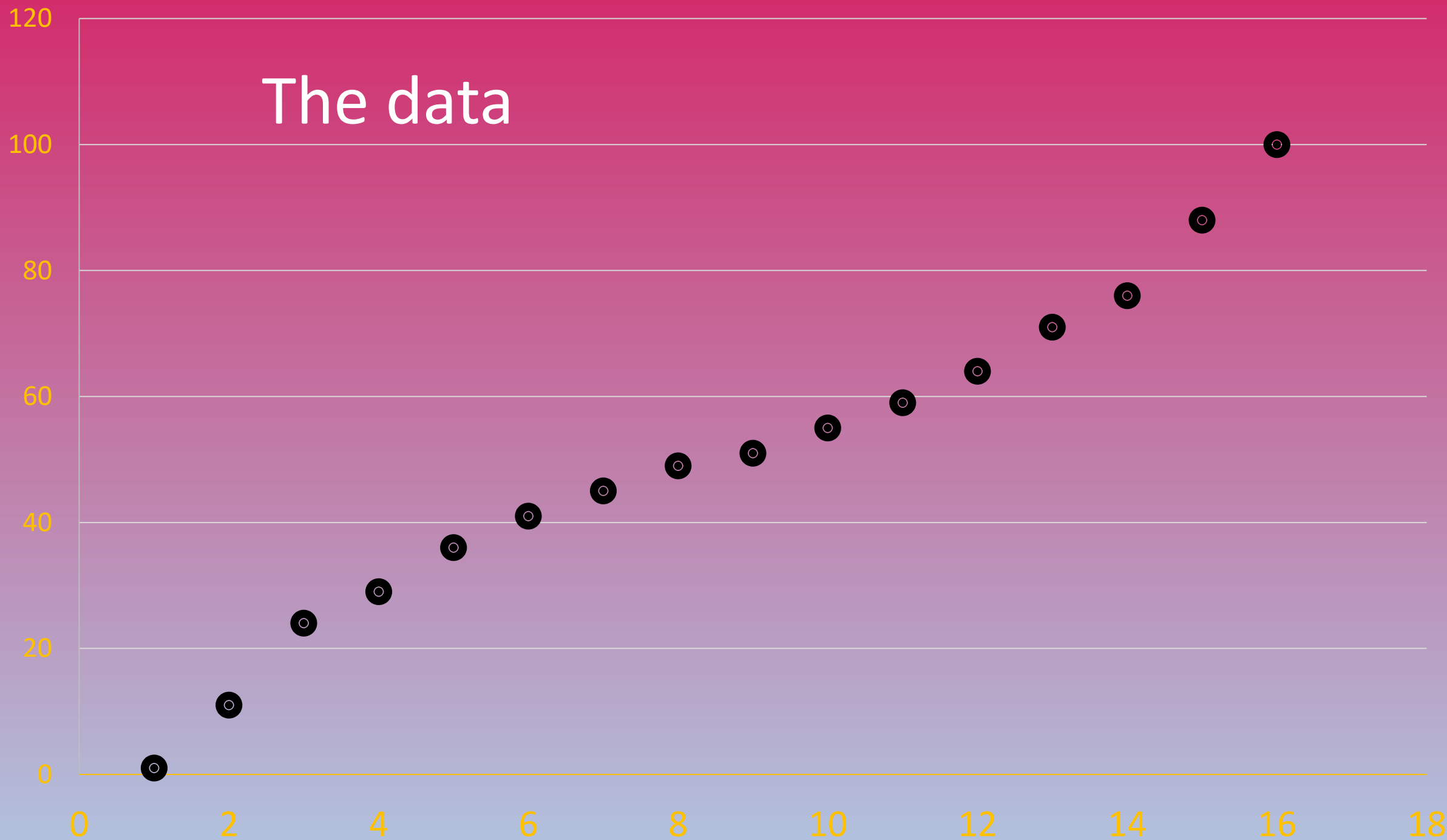
Where

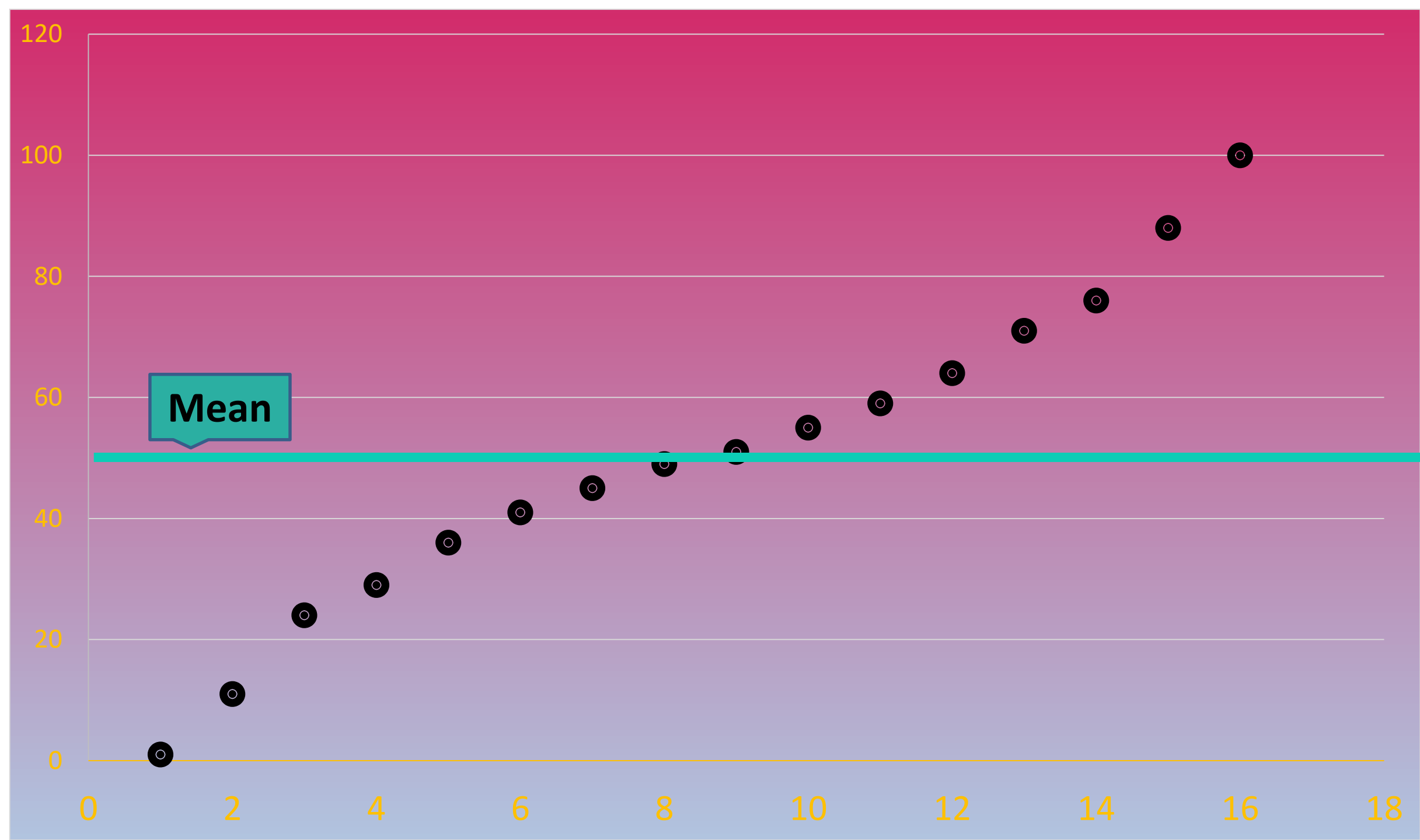
N - total number of observation

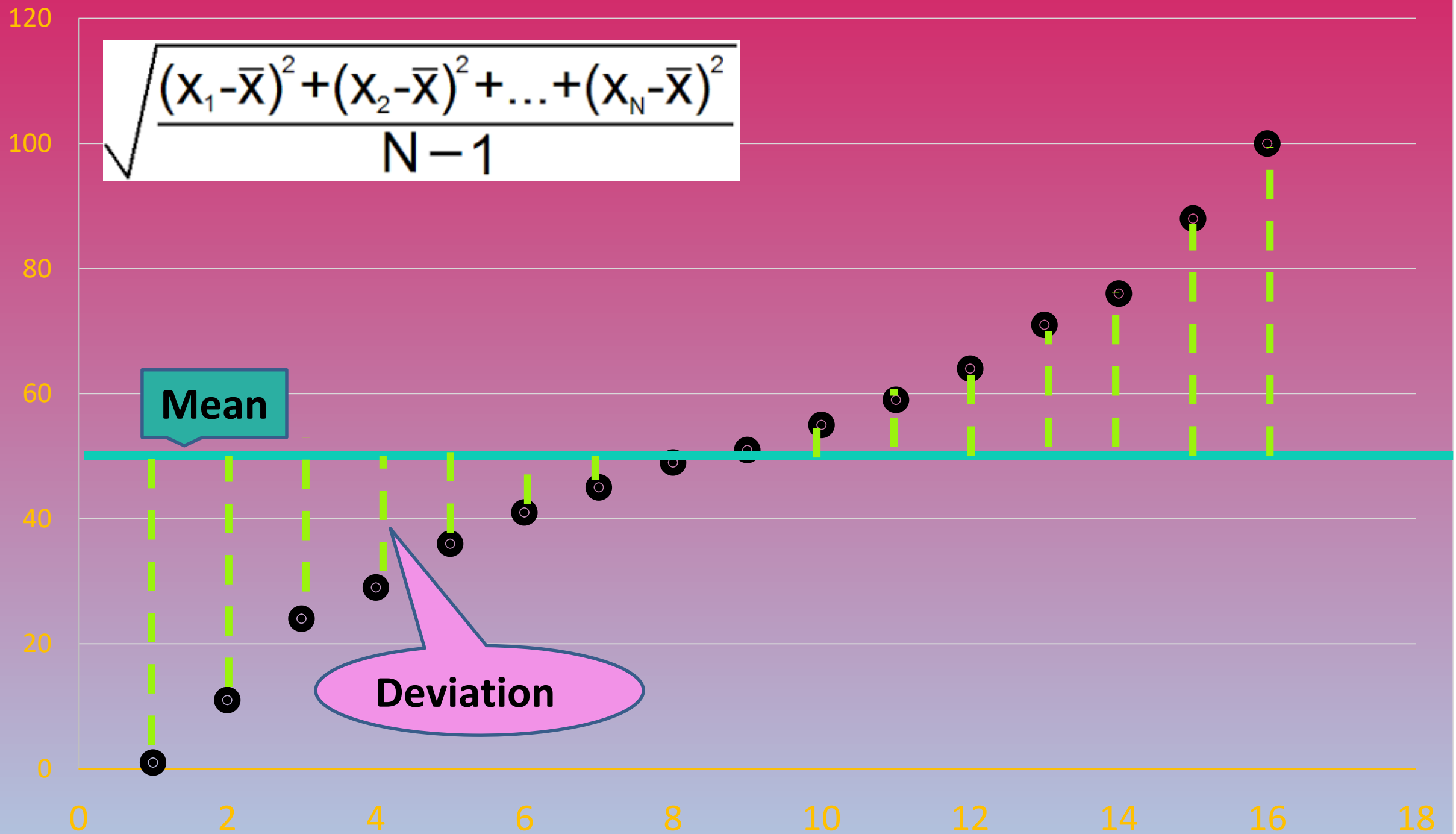
$\bar{X}$ - arithmetic mean

$X_1, \dots, X_N$ - observation

The data







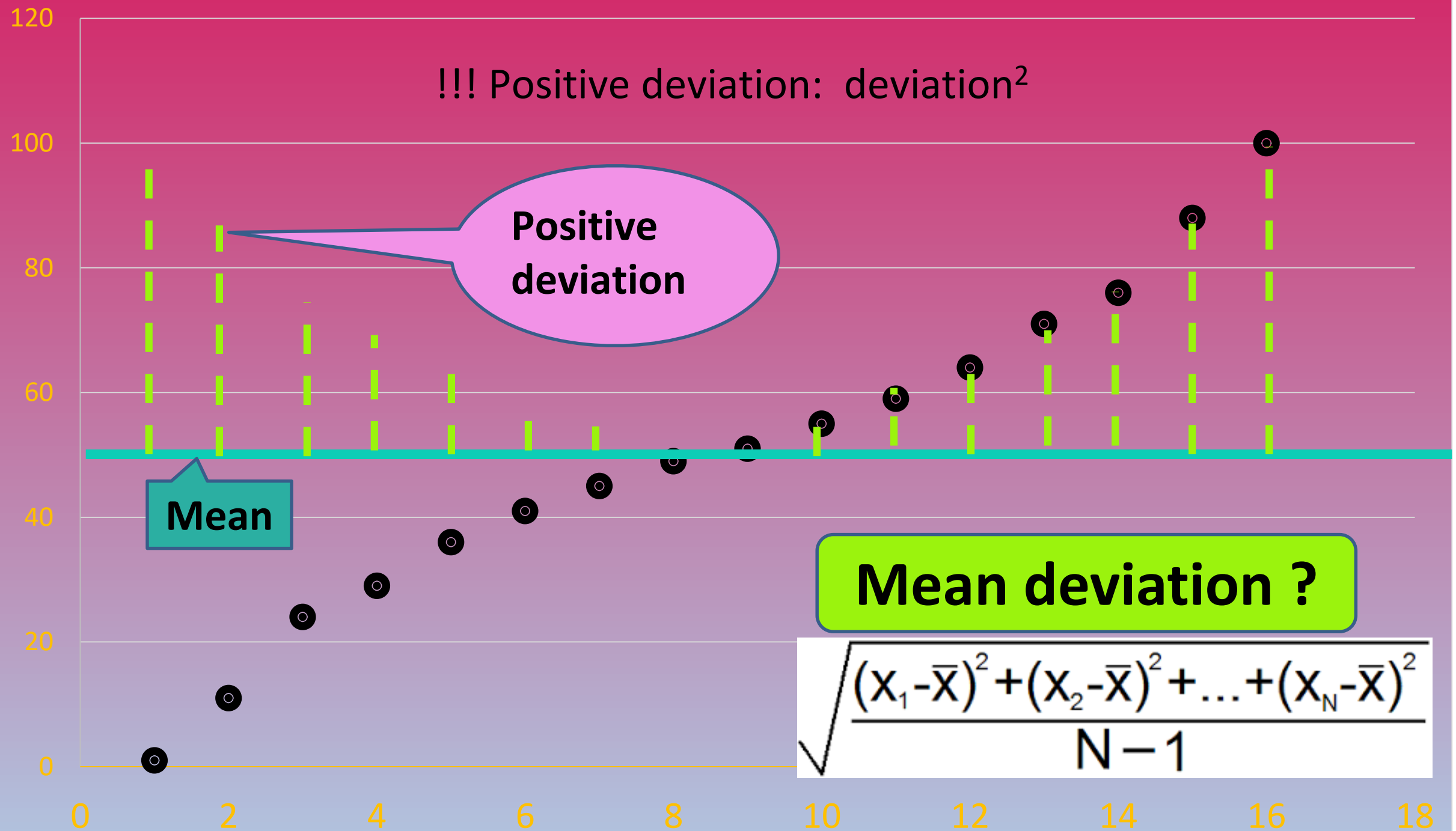
!!! Positive deviation: deviation²

Positive
deviation

Mean

Mean deviation ?

$$\sqrt{\frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_N - \bar{X})^2}{N - 1}}$$





A diagram illustrating the concept of mean deviation. It features a horizontal black line representing the mean. Above and below this line are several vertical red dashed lines of varying heights, representing individual data points. A light blue rounded rectangle with a black border is centered on the horizontal line, containing the text "Mean deviation".

Mean deviation

$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N - 1}}$$



Example

- No. of tooth for 6 patients: 20, 24, 26, 25, 25, 30
- Arithmetic mean $\bar{X} = (20+24+26+25+25+30)/6=25$
- Standard deviation:

$$\begin{aligned} S &= \sqrt{\frac{(20-25)^2 + (24-25)^2 + (26-25)^2 + (25-25)^2 + (25-25)^2 + (30-25)^2}{6-1}} = \\ &= \sqrt{\frac{(-5)^2 + (-1)^2 + 1^2 + 0^2 + 0^2 + 5^2}{5}} = \sqrt{\frac{25+1+1+25}{5}} = \sqrt{\frac{52}{5}} = \sqrt{10.4} = 3.22 \end{aligned}$$

10.4 is the variation

Mean deviation from the mean :-P

Standard Deviation - example

Ex. In a hepatitis A epidemic, a sample of 6 persons had clinical symptoms after 24, 25, 29, 29, 30, 31 days.

$X = 24, 25, 29, 29, 30, 31$ - the individual data, $\bar{X} = 28$

1. $X - \bar{X} = -4, -3, 1, 1, 2, 3$ (each data minus arithmetic mean)
2. $(X - \bar{X})^2 = 16, 9, 1, 1, 4, 9$ (square)
3. $\Sigma(X - \bar{X})^2 = 16+9+1+1+4+9 = 40$ (sum)
4. $\Sigma(X - \bar{X})^2 / (n-1) = 40/5 = 8$ (divide by $6-1=5$) = variance
5. $\sqrt{\Sigma(X - \bar{X})^2 / (n-1)} = 2.83$ This value is the standard deviation

Standard Deviation - example

Cholesterol values are given for 10 patients: 200, 180, 140, 160, 180, 150, 170, 110, 230, 170

	Cholesterol (X)	$X - \bar{X}$	$(X - \bar{X})^2$
	200	31	961
	180	11	121
	140	-29	841
	160	-9	81
	180	11	121
	150	-19	361
	170	1	1
	110	-59	3481
	230	61	3721
	170	1	1
Sum	1690	0	9690
$\bar{X}$	169	$(X - \bar{X})^2 / (n-1)$	1076.6
		St.Dev.	32.8

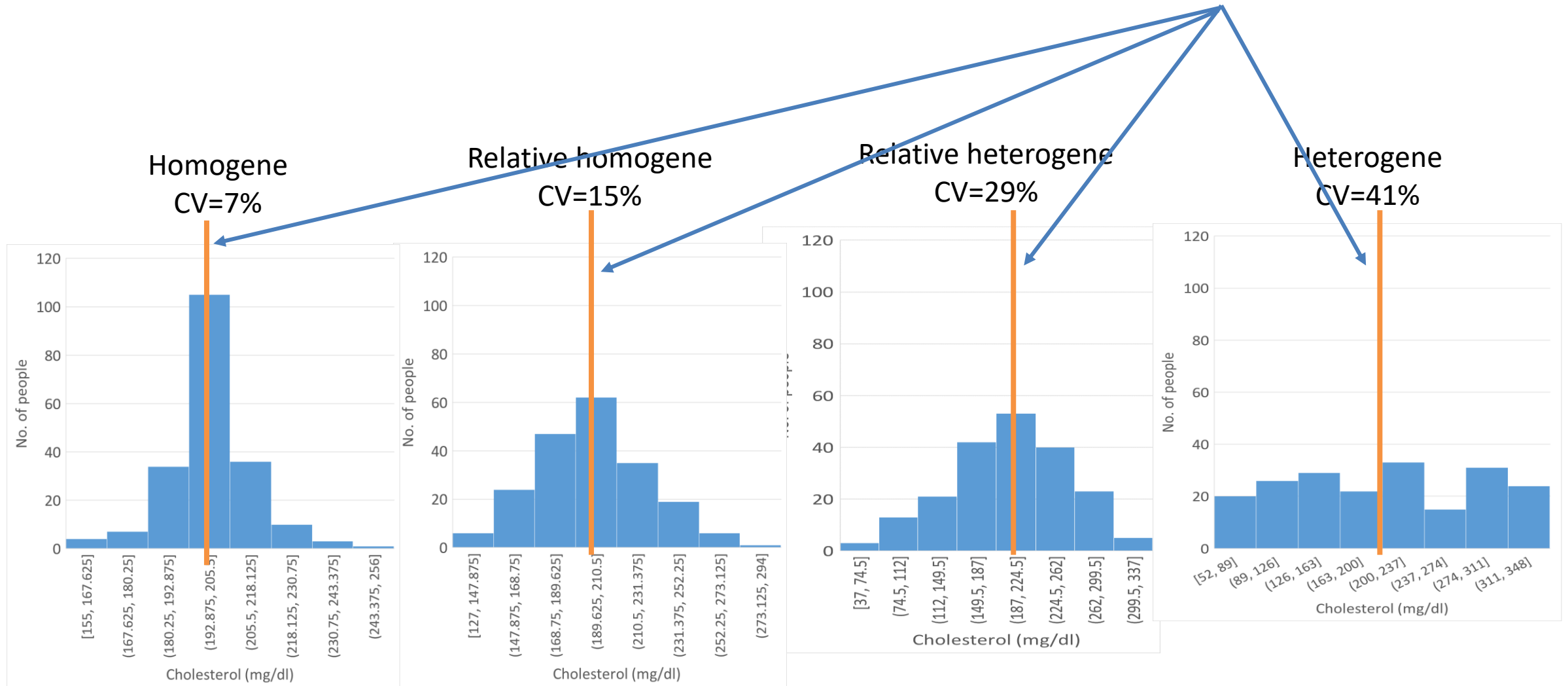
Deviation
from mean

Positive
deviation
from mean

Other measures

- **The coefficient of variation ($CV = s/\bar{X} * 100$)**
 - interpretation
 - $0\% \leq CV < 10\%$ data are homogeneous
 - $10\% \leq CV < 20\%$ data are relatively homogeneous
 - $20\% \leq CV < 30\%$ data are relatively heterogeneous
 - $30\% \leq CV$ data are heterogeneous
- where s =standard deviation, $\bar{X}$ =arithmetic mean

homogene = values are in vicinity of the arithmetic mean
heterogene = values are not in vicinity of the arithmetic mean

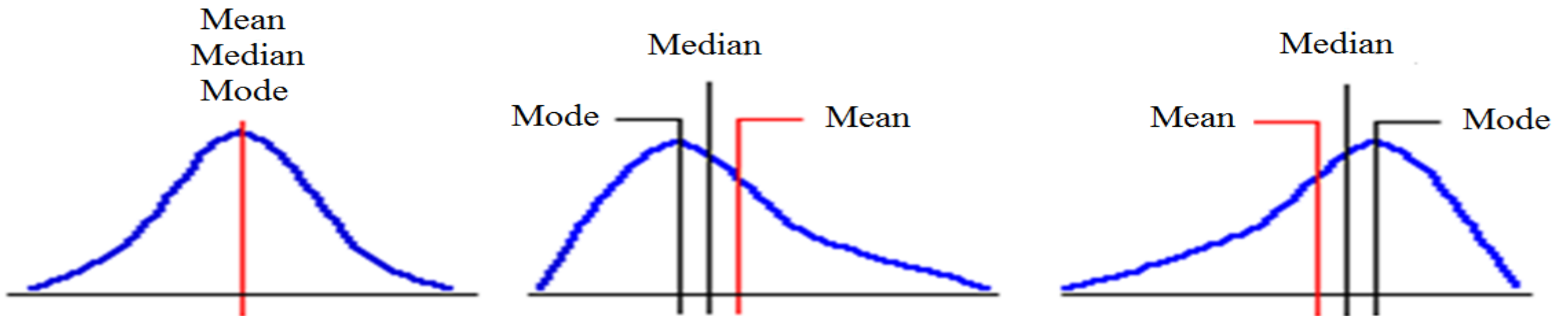


100 patients

100 patients

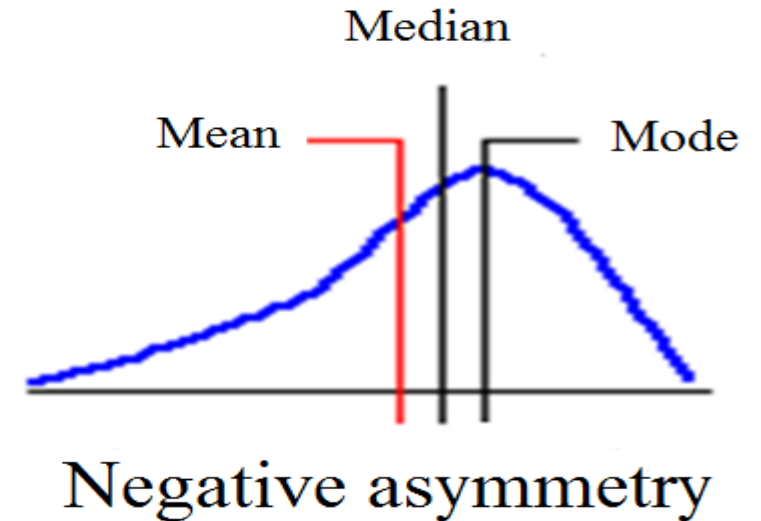
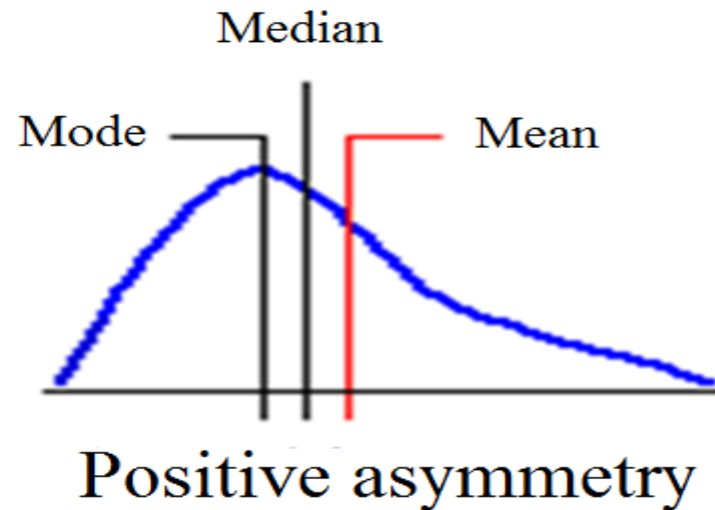
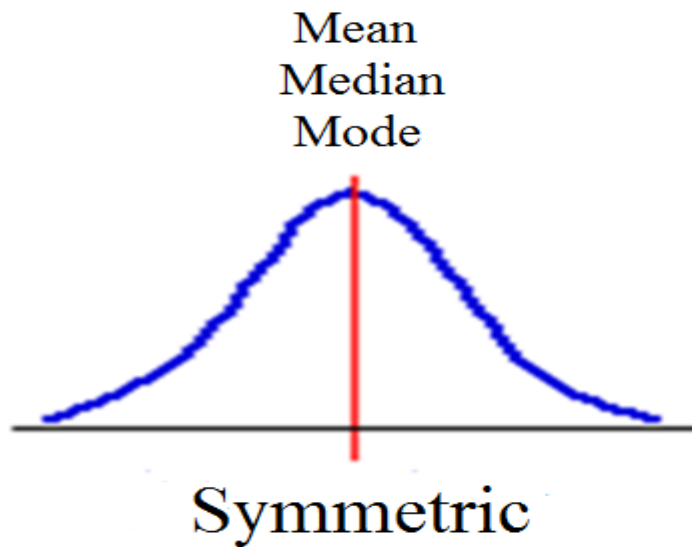
Measure of centrality - applicability

How to interpret the mean, median and mode:



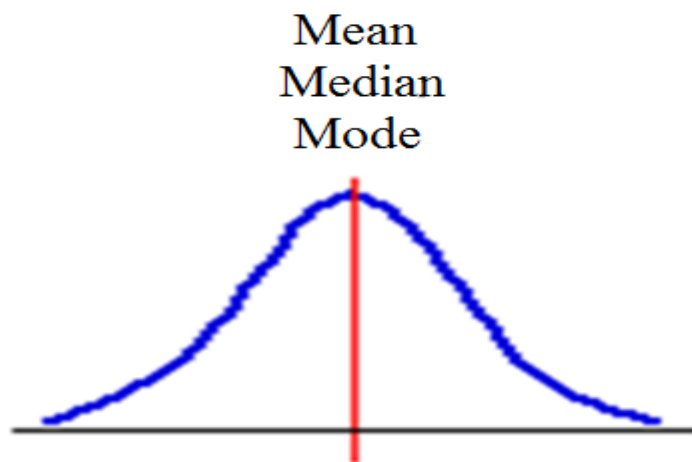
Measure of centrality - applicability

How to interpret the mean, median and mode:

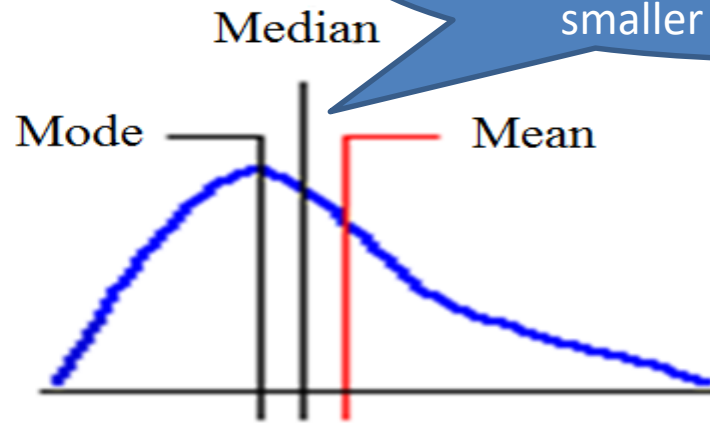


Measure of centrality - applicability

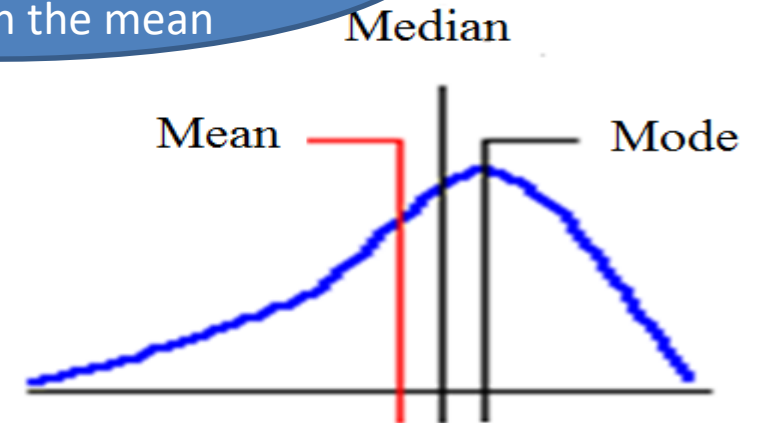
How to interpret the mean, median and mode:



Symmetric



Positive asymmetry



Negative asymmetry

Mean higher than median and mode, outliers are high values, but most of the values are smaller than the mean

Mean lower than median and mode, outliers are smaller values, most of the values are higher than the mean



Skewness

- Skewness – Asymmetry measure

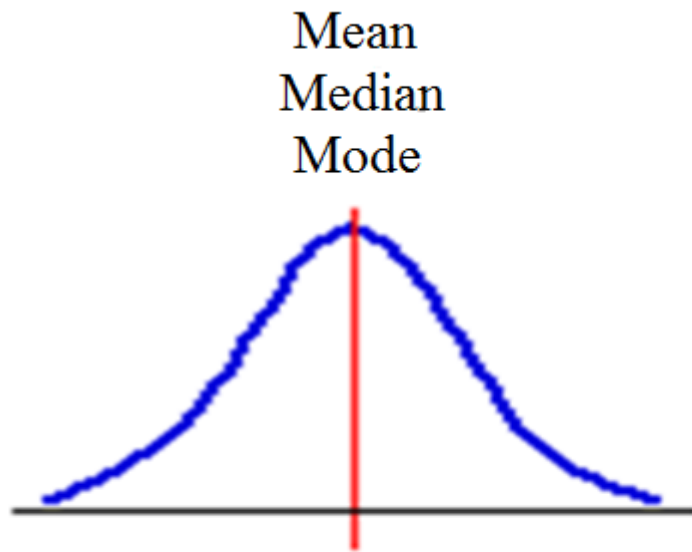
$$S_K \approx \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}$$

where:

s = sample standard deviation , $\bar{X}$ = arithmetic mean

Skewness – Asymmetry measure

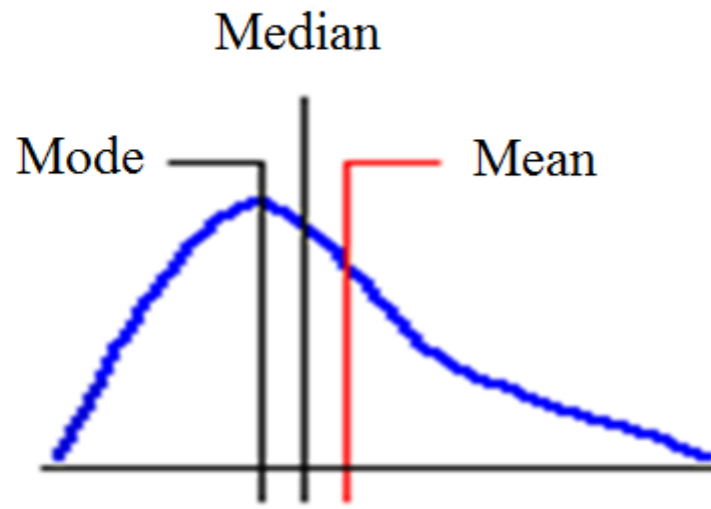
- **Skewness** is a measure of the asymmetry of the distribution of a quantitative variable.



Symmetric

Skewness=0

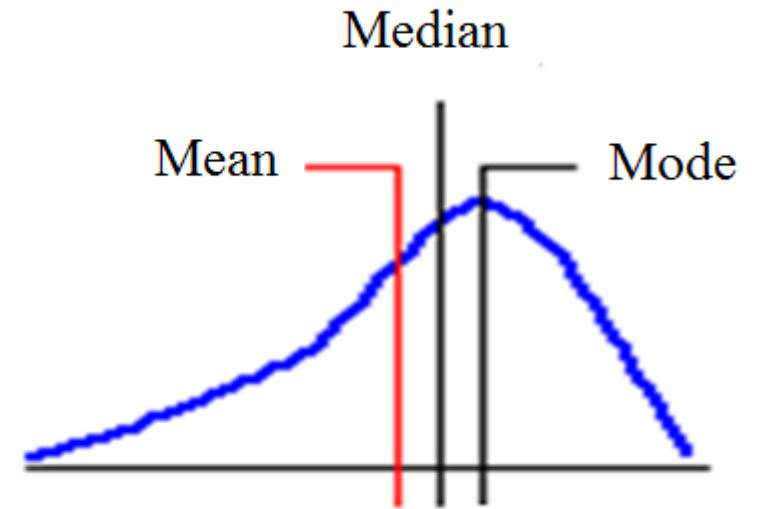
symmetrical distribution



Positive asymmetry

Skewness>0

asymmetry to the right



Negative asymmetry

Skewness<0

asymmetry to the left



Median = 2900 €

Arithmetic mean = 5500 €

2400€ 2600 €



3200 €



19000 €

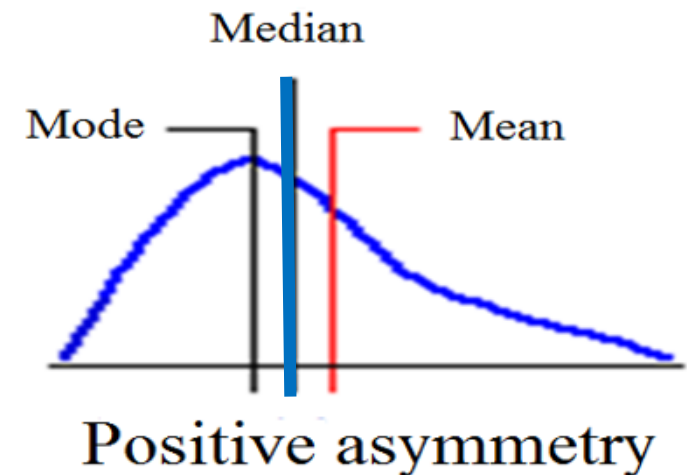


1 value

2900 €

2900 €

5 values



Kurtosis

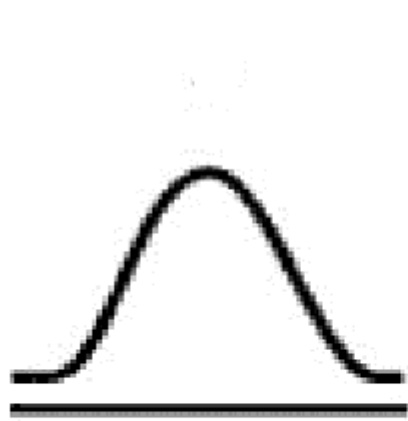
- Kurtosis – Peakedness, flatness, tailedness measure

$$K_E \approx \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{s^4} - 3$$

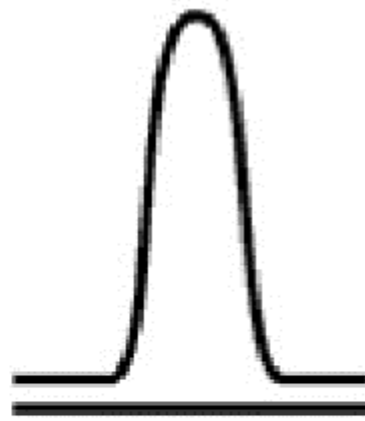
s = sample standard deviation , $\bar{X}$ = arithmetic mean

Kurtosis - Peakedness, flatness, tailedness measure

- Kurtosis is a measure of the "tailedness" of the distribution of a variable



Mesokurtic
 $\text{kurtosis}=0$



Leptokurtic
 $\text{kurtosis}>0$



Platykurtic
 $\text{kurtosis}<0$

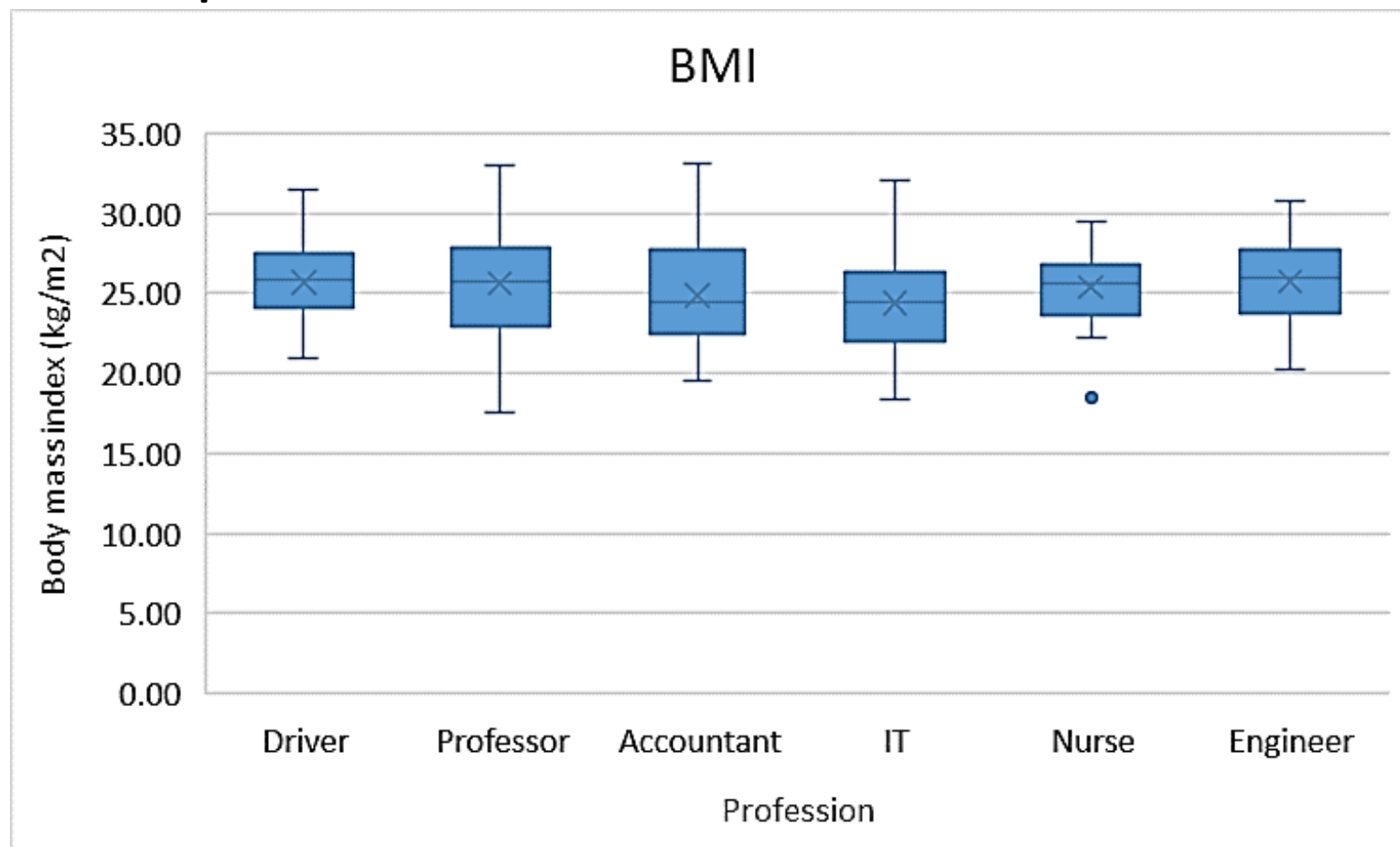


The relationship between two variables (from which one is quantitative or both are quantitative)

	A	B
1	Profession	Body mass index
2	Driver	25.66
3	Driver	27.45
4	Driver	22.33
5	Driver	26.27
6	Driver	25.96
7	Driver	25.24
8	Driver	25.05
9	Driver	28.19
10	Driver	31.52
11	Driver	28.07
12	Driver	23.36
13	Driver	21.99
14	Driver	25.80
15	Driver	25.81
16	Driver	24.39
17	Driver	20.90
18	Driver	26.76
19	Driver	27.68
20	Professor	32.98
21	Professor	31.10
22	Professor	23.72
23	Professor	28.09
24	Professor	22.92
25	Professor	25.76
26	Professor	21.38
27	Professor	19.13
28	Professor	25.28
29	Professor	27.47
30	Professor	24.89

The relationship between two variables (one quantitative, one qualitative)

- Box-whisker plot



Statistics - two variables (one quantitative, one qualitative)

Professions

BMI (kg/m ²)	IT	Driver	Professor	Accountant	Nurse	Engineer
Mean	29.82	32.79	25.77	24.83	25.38	25.72
Standard Error	0.50	0.60	0.77	0.75	0.46	0.49
Median	29.78	32.54	25.73	24.50	25.67	26.00
Mode	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Standard Deviation	2.55	2.53	3.95	3.53	2.44	2.63
Sample Variance	6.53	6.38	15.58	12.44	5.95	6.90
Kurtosis	-0.57	0.38	1.12	0.05	1.03	-0.52
Skewness	-0.06	0.66	0.35	0.54	-0.51	-0.25
Range	10.51	9.16	18.56	13.63	11.01	10.51
Minimum	24.43	29.28	17.51	19.51	18.46	20.27
Maximum	34.93	38.44	36.07	33.13	29.47	30.78
Sum	775	590	670	546	710	745
Count	26	18	26	22	28	29

there is not one value with the highest frequency

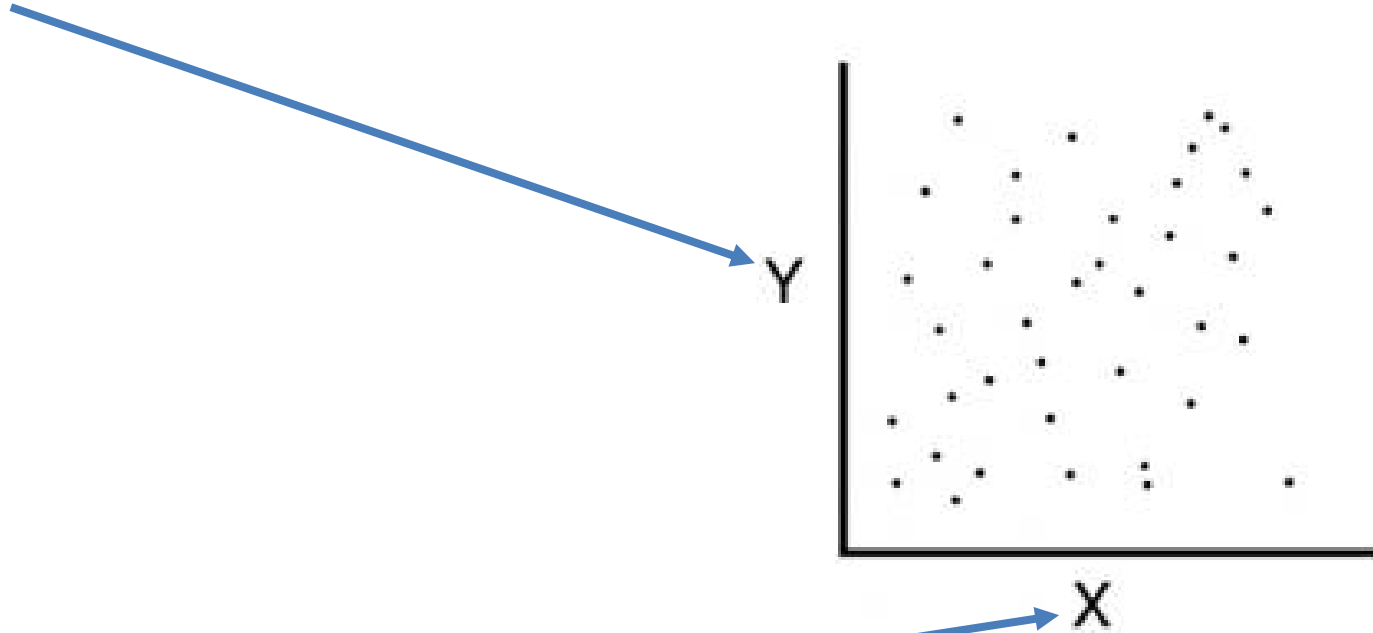
no. of people in each group

XY Scatter chart

- The relationship between two quantitative or ordinal (scores) characteristics
- How is the relationship?

XY scatter chart

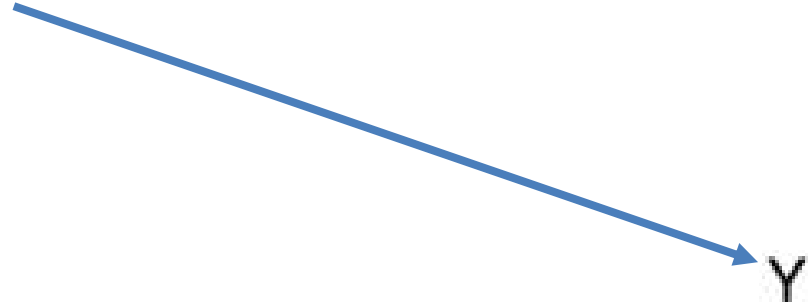
Y – quantitative variable or ordinal (score)



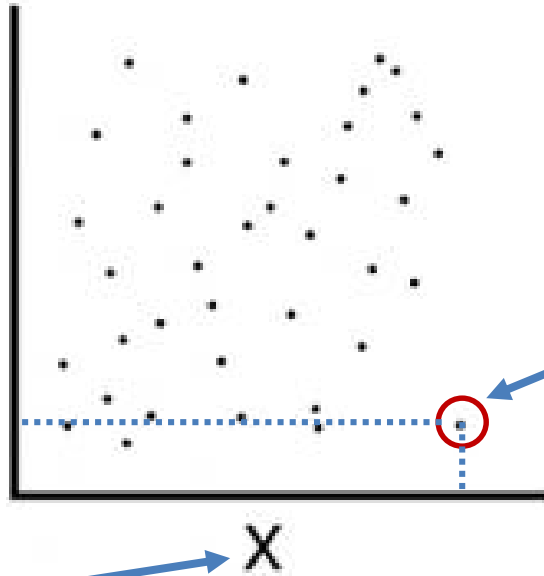
X – quantitative variable or ordinal (score)

XY scatter chart

Y – quantitative variable or ordinal (score)



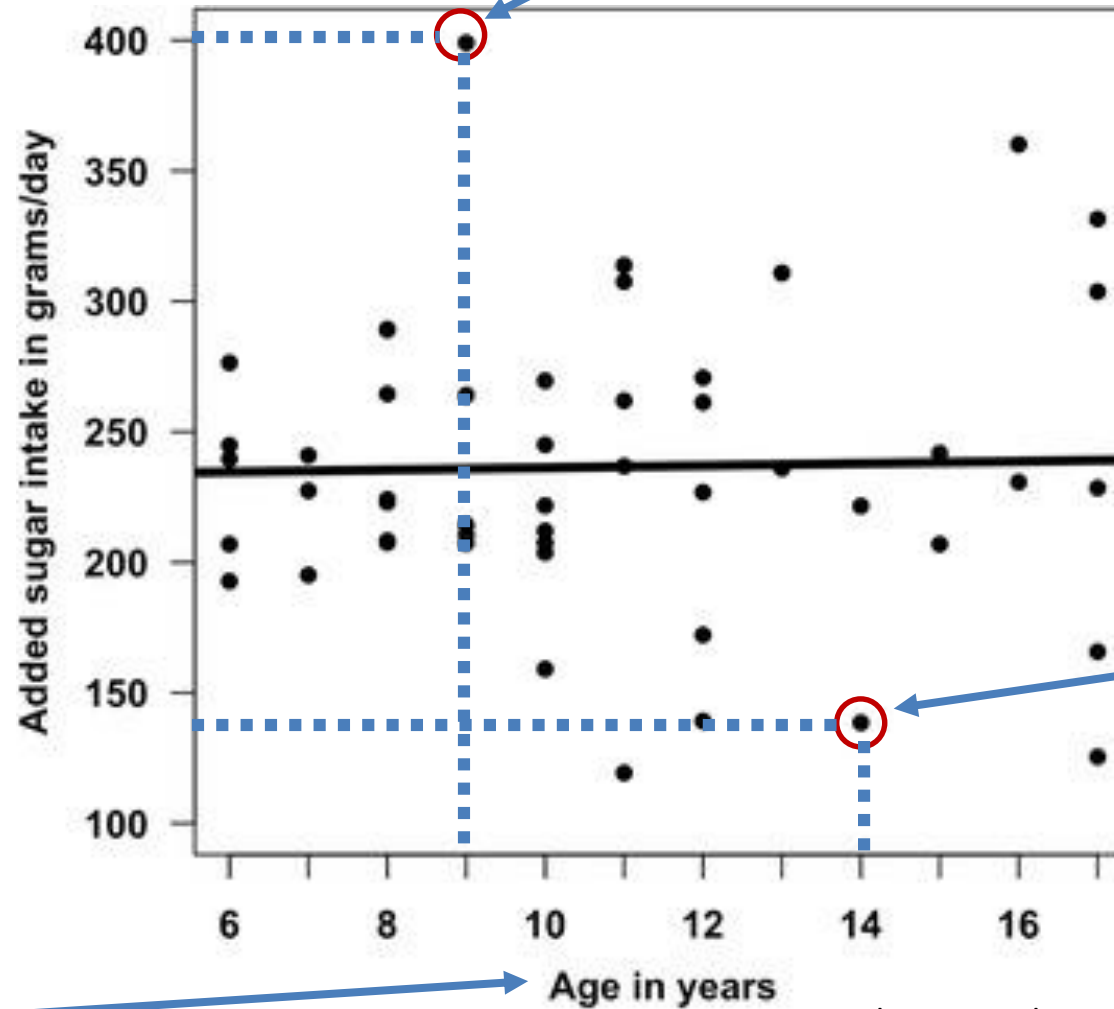
A subject



X – quantitative variable or ordinal (score)

Y – quantitative variable - **Sugar intake**

A subject with age=9 years and 400 grams/days sugar intake



A subject with age=14 years and 140 grams/days sugar intake

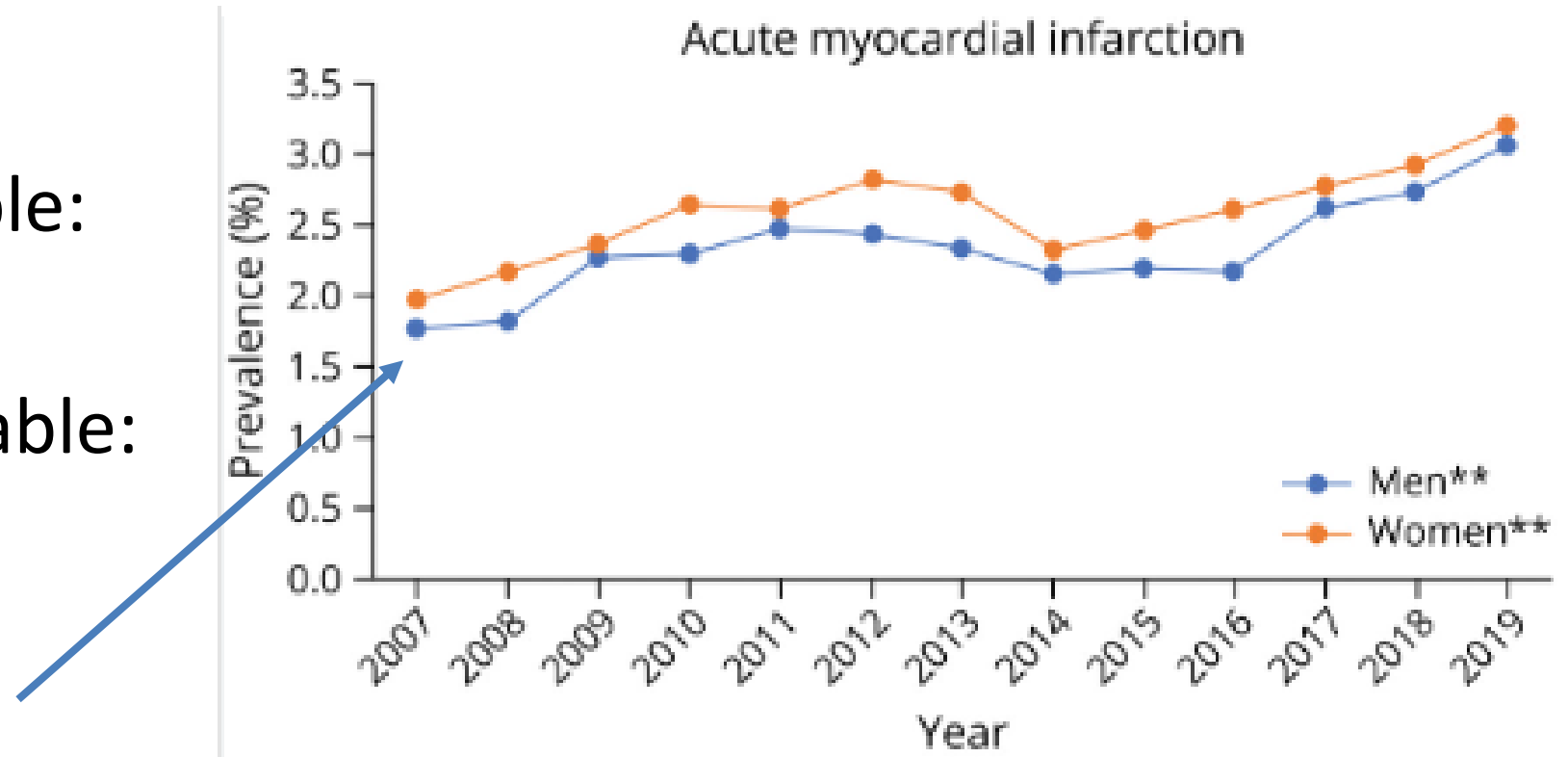
X – quantitative variable - **Age**

Chi DL, Hopkins S, O'Brien D, Mancl L, Orr E, Lenaker D. Association between added sugar intake and dental caries in Yup'ik children using a novel hair biomarker. BMC Oral Health. 2015 Oct 9;15(1):121.

Other charts

Line charts

- qualitative variable:
 - hypertension
- quantitative variable:
 - time

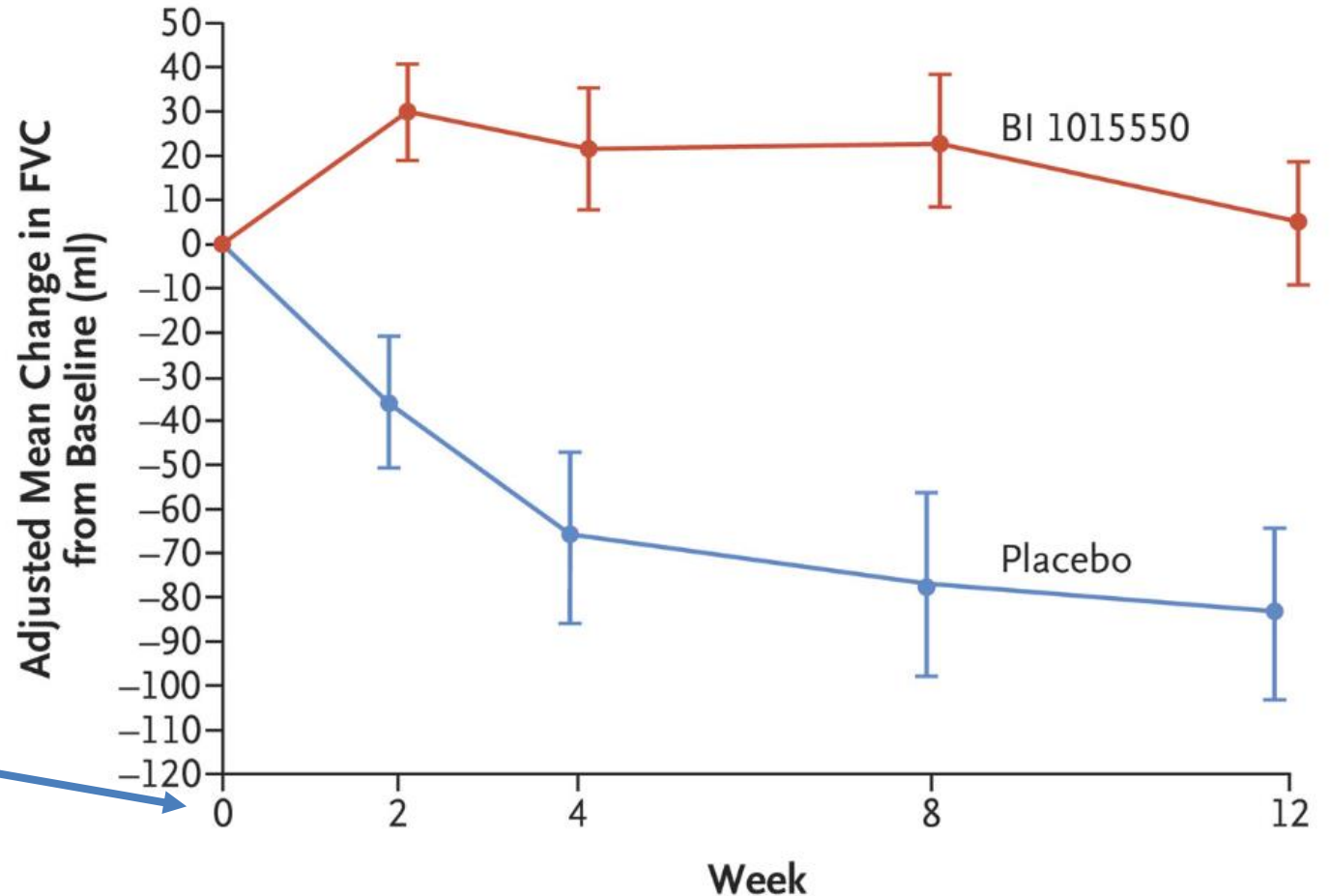


Trends in the frequency of nonInfectious complications after acute ischemic stroke admission in the united states according to sex

Other charts

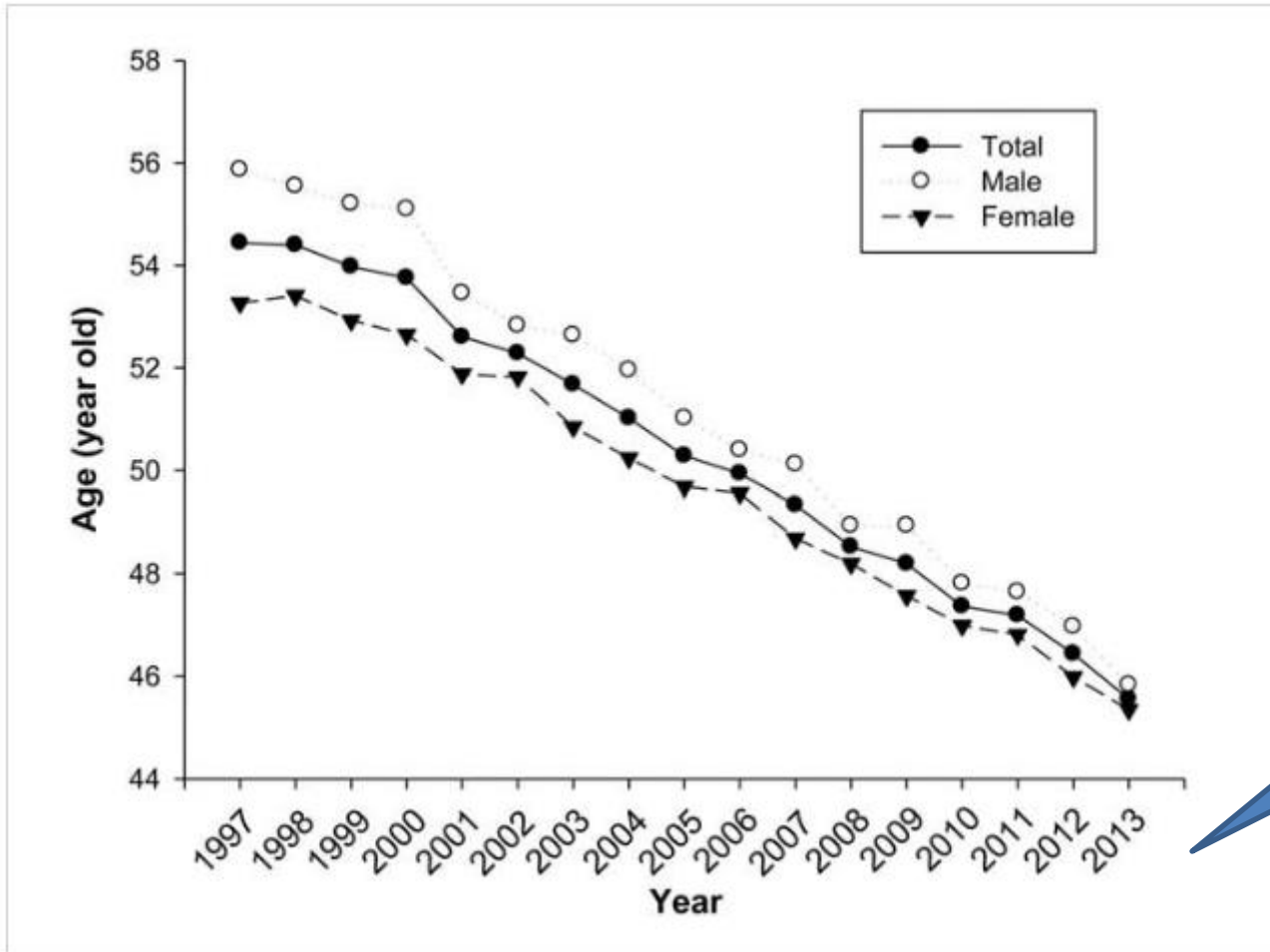
Line charts

- quantitative variable:
 - FVC (forced vital capacity)
 - mean, standard error
- quantitative variable:
 - time



Change in FVC over time for subjects with BI 1015550 versus placebo

Line charts with arithmetic means



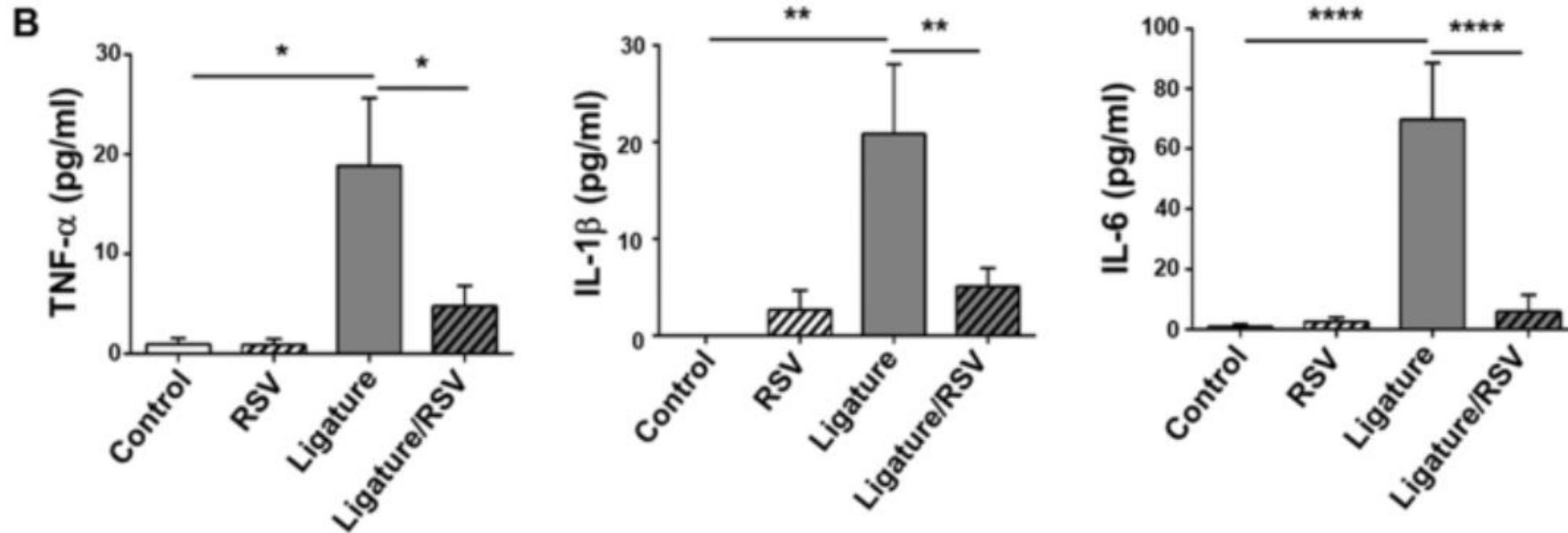
Mean age of patients with periodontitis in Taiwan. The mean age for periodontitis was shown a decrease pattern from 1997 to 2013.

Time

Other measures

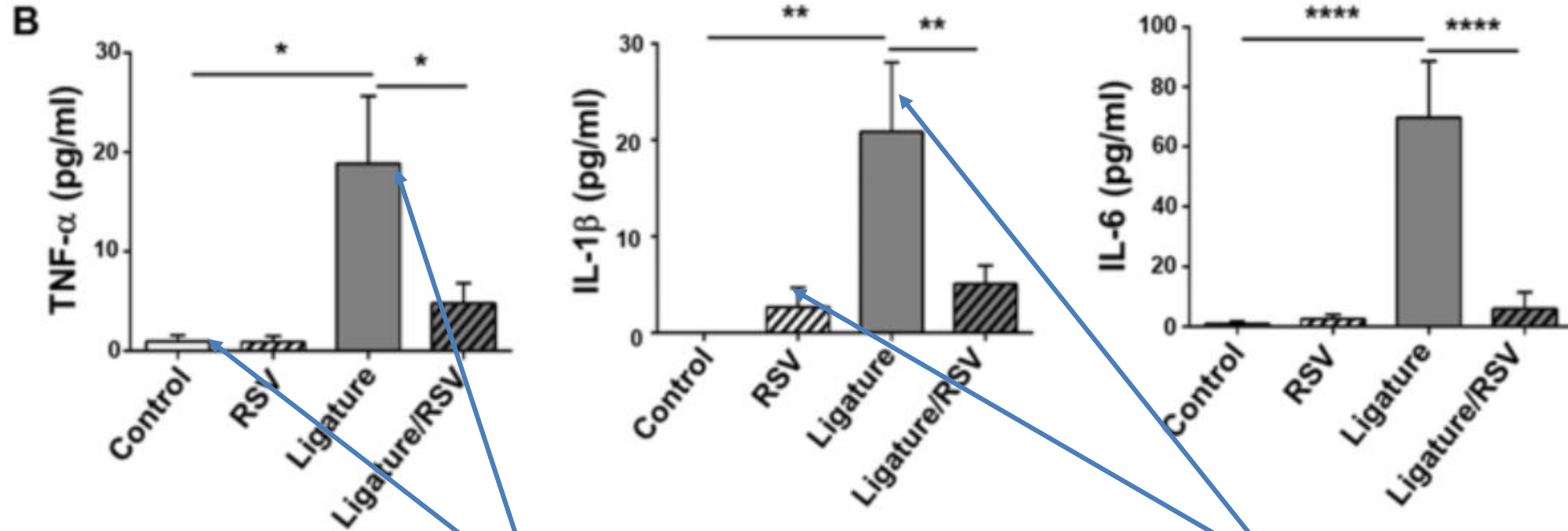
- The **standard error** ($E_s = s/\sqrt{n}$)
- where s =standard deviation, n =number of data

Mean error plot



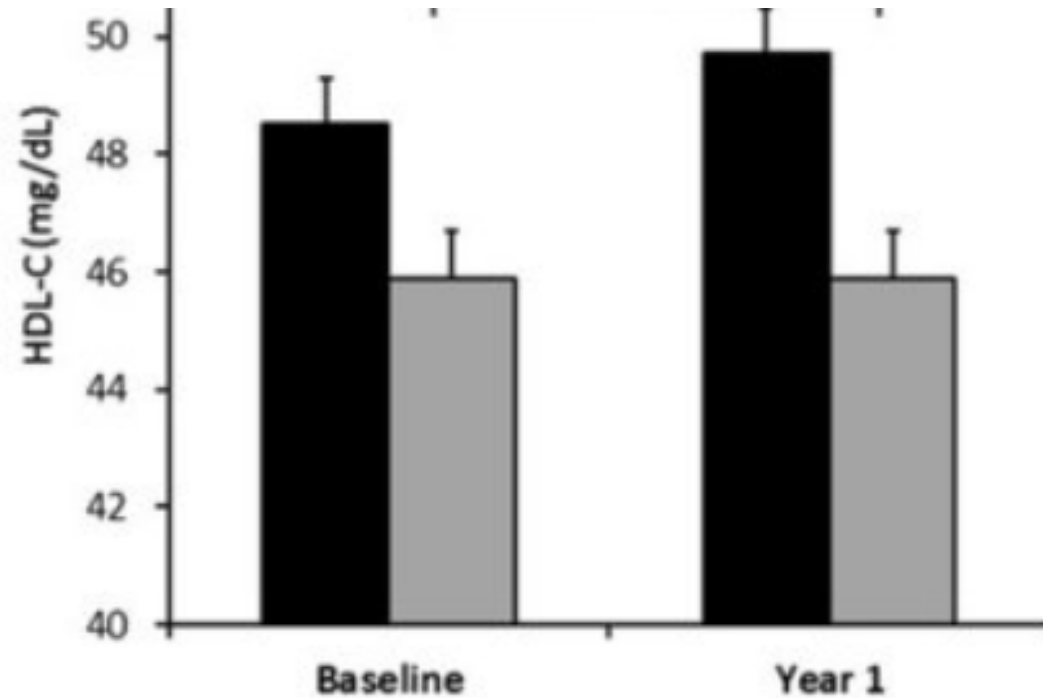
- Ligature placement (paradontosis disease induced in mice) notably increased the serum levels of pro-inflammatory cytokines, and rosovastatin (RSV) completely inverted such increases.
- **(B)** Levels of TNF- α , IL-1 β and IL-6 from the mice
- means + SE

Mean error plot



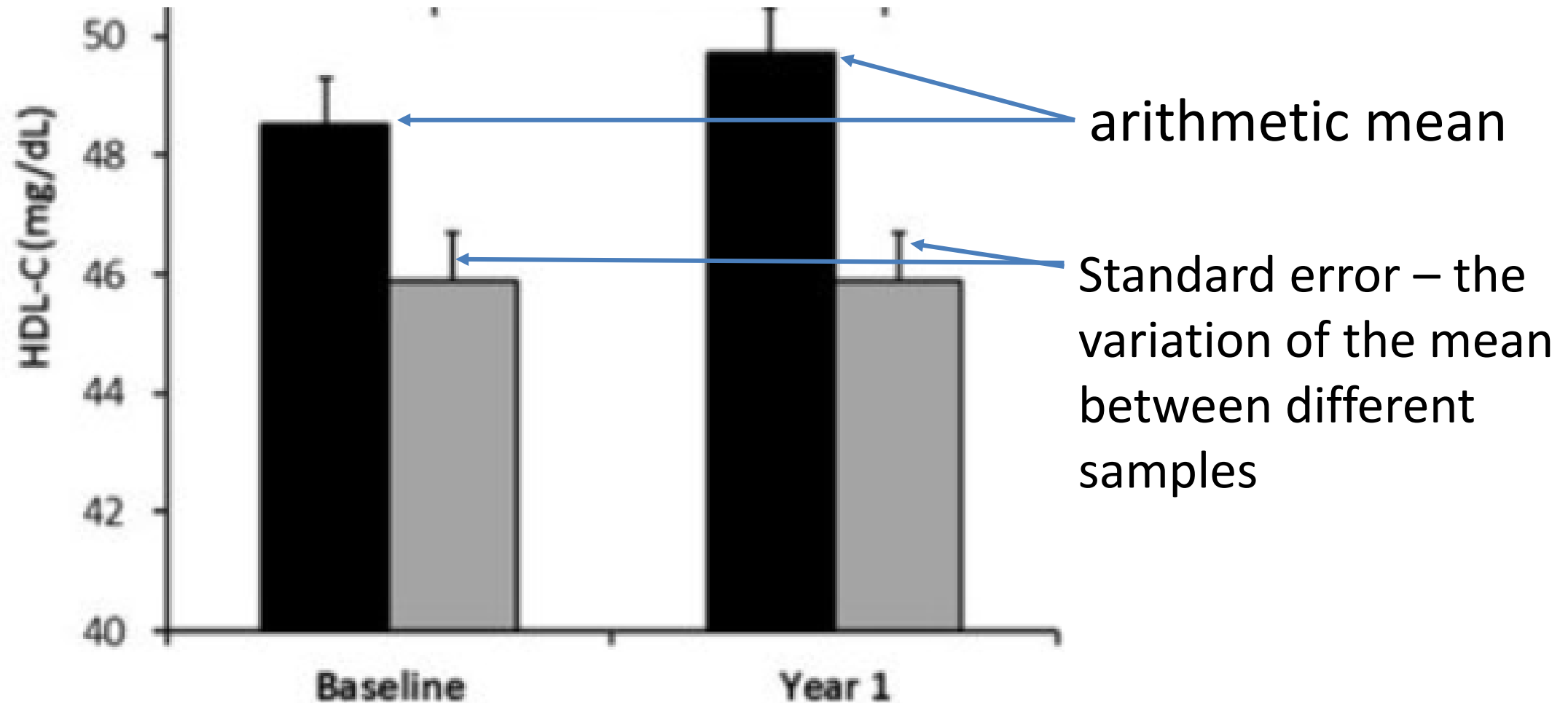
arithmetic mean

Standard error – the variation of the mean between different samples

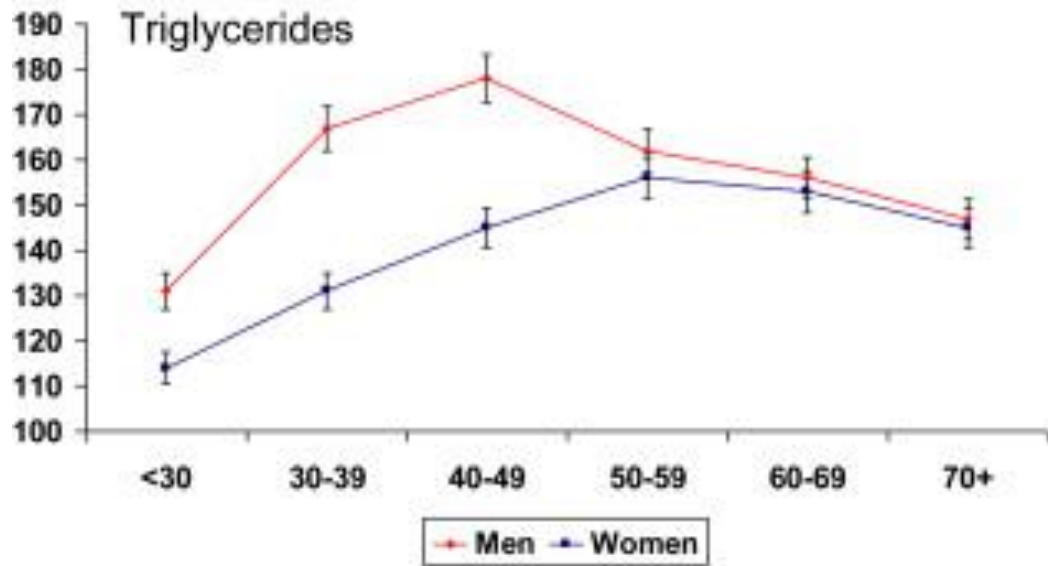
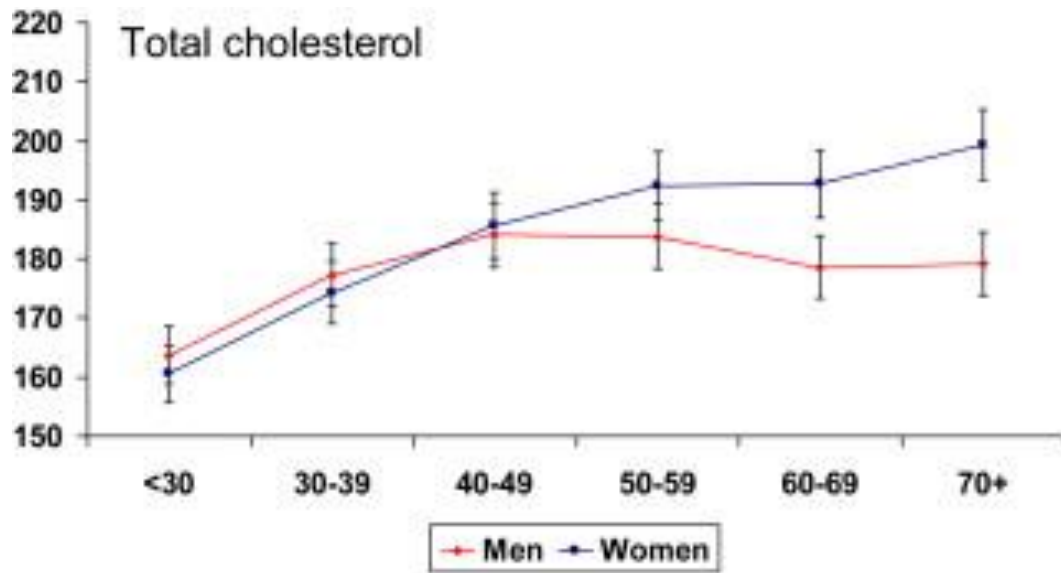


- Error-plot
- HDL-cholesterol in subjects treated with Metformin at baseline and after one year
- 221 African Americans

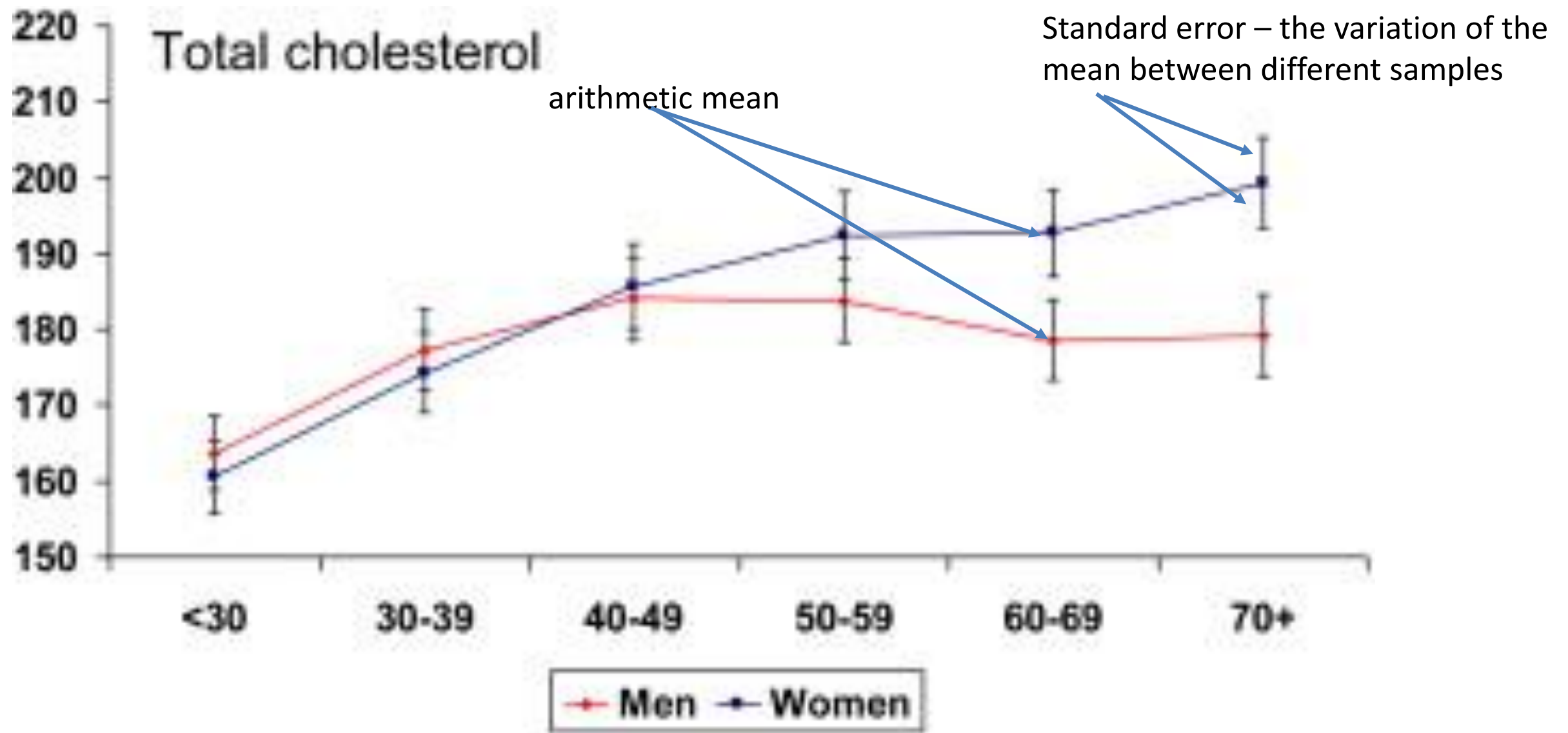
- Error-plot



Zhang C, Gao F, Luo H, Zhang CT, Zhang R. Differential response in levels of high-density lipoprotein cholesterol to one-year metformin treatment in prediabetic patients by race/ethnicity. *Cardiovasc Diabetol*. 2015 Jun 12;14:79.



- 6123 subjects (men 3388)
- Age-group specific levels of total cholesterol, triglycerides (mg/dl) in women and men.



Guptha S, Gupta R, Deedwania P, Bhansali A, Maheshwari A, Gupta A, Gupta B, Saboo B, Singh J, Achari V, Sharma KK. Cholesterol lipoproteins and prevalence of dyslipidemias in urban Asian Indians: a cross sectional study. Indian Heart J. 2014 May-Jun;66(3):280-8.

Summarizing Numerical Data (Continuous, Discrete)



One quantitative variable

- Measures of location
 - Quartiles
 - Percentiles
 - Measure of centrality
 - Arithmetic mean
 - Median
 - Mode
- Measures of asymmetry
 - Skewness
 - Kurtosis
- Measure of spread
 - Variances
 - Standard Deviation
 - Coefficient of variation
 - Standard error
- Charts
 - Histograms

Two quantitative variables

- Charts
 - XY Scatter
 - line charts

One qualitative, one quantitative variable

- Charts
 - Box-whiskers
 - Error plot
 - Line

Thank you!