Introduction to Biostatistics

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2025-2026
ghassand7@Gmail.com

References

1. Trial M.F. (2010) Elementary Statistics, Eleventh edition. Addison –Wesley.

2. Wayne W. D. (2010) Biostatistics. John Wiley & Sons, INC.

3- خاشع الراوي مقدمة في الاحصاء

Biostatistics

- (a portmanteau word made from biology and statistics)
- The application of statistics to a wide range of topics in biology.

Biostatistics

It is the science which deals with development and application of the most appropriate methods for the:

- Collection of data.
- Presentation of the collected data.
- Analysis and interpretation of the results.
- Making decisions on the basis of such analysis

Some basic concepts

- Data: the row material of Statistics

Source of data

1- Routinely kept records

2- survey

3-Experiment

4-External resources

Types Simple random sample

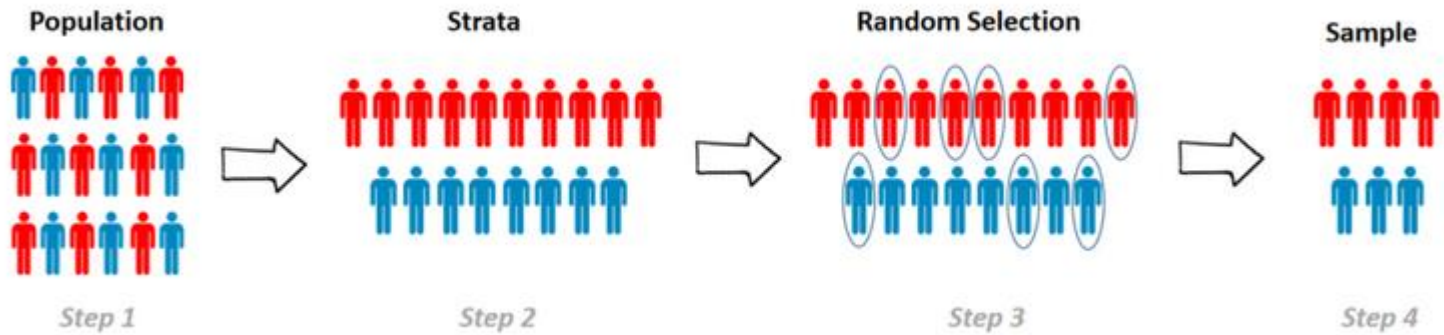
- Simple random sample
  - Systematic random sample
  - Unsystematic random sample
  - Stratified random sample
  - Cluster sample
  - Purposive sample

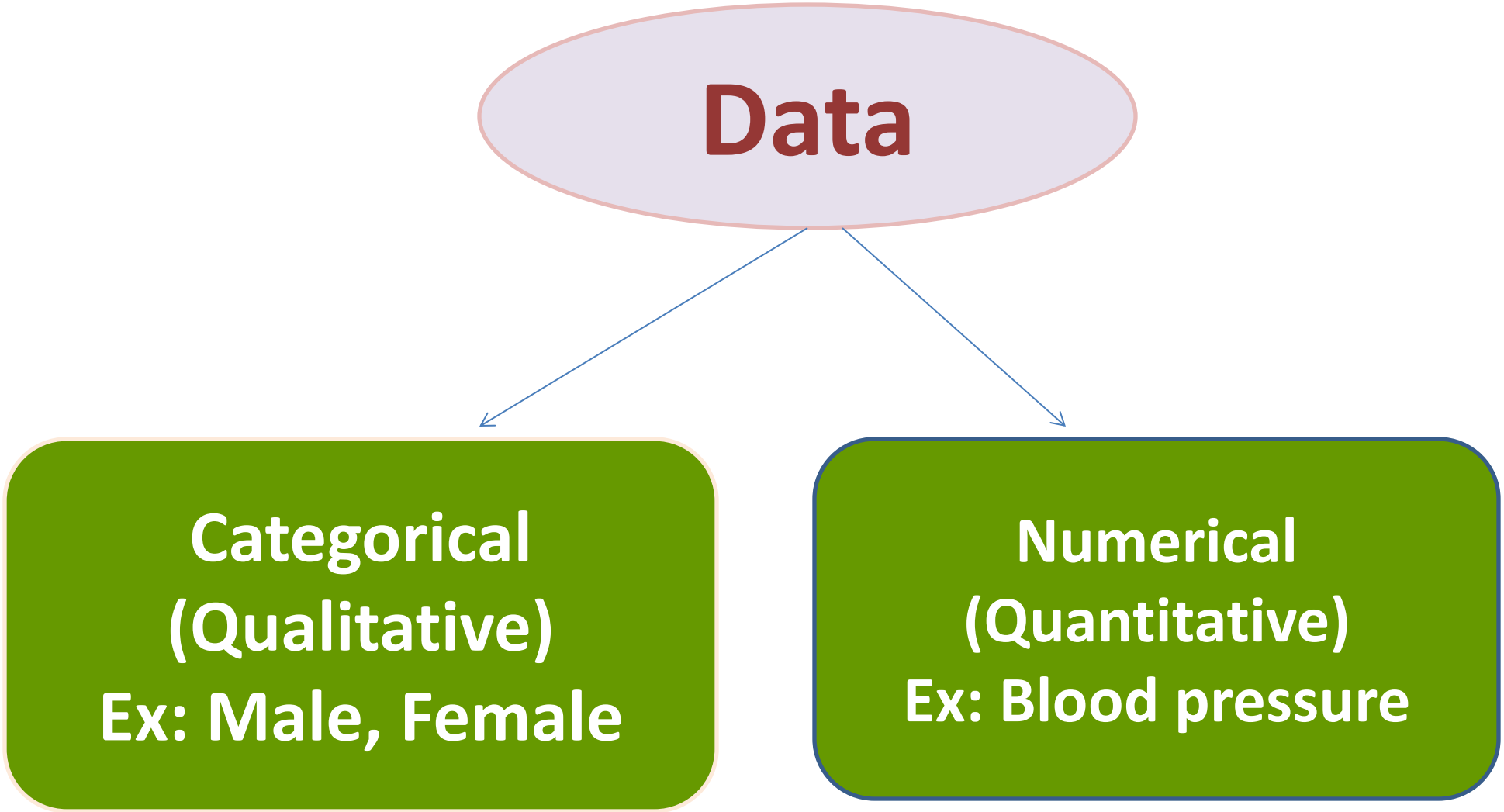
Types Simple random sample

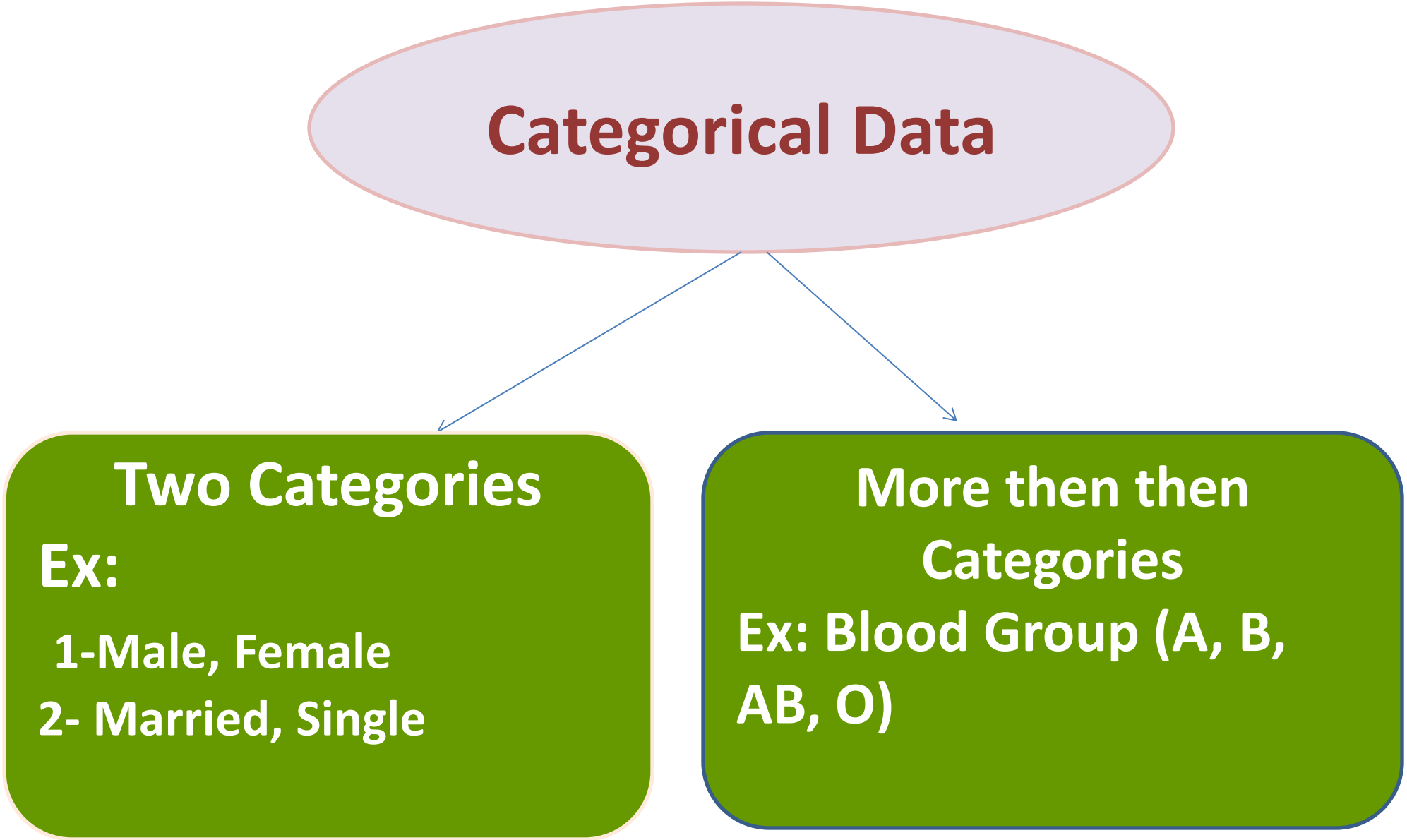
- Simple random sample
  - Systematic random sample
  - A systematic random sample is created by selecting a random starting point from a list and then choosing every nth element thereafter to form the sample. For example, to survey 200 out of 1,000 customers, calculate the sampling interval by dividing the population size by the desired sample size (1,000/200 = 5). Then, select a random starting point from the first five customers (e.g., the 3rd customer), and proceed to select every 5th customer from that point forward (3rd, 8th, 13th, etc.) until the sample of 200 is complete.

Types Simple random sample

- Simple random sample
  - Stratified random sample




Data
Categorical
(Qualitative)
Ex: Male, Female
Numerical
(Quantitative)
Ex: Blood pressure


Categorical Data
Two Categories
Ex:
1-Male, Female
2- Married, Single
More then then Categories
Ex: Blood Group (A, B, AB, O)

Numerical Data

- .1-Discrete data: Observation commonly take certain numerical.

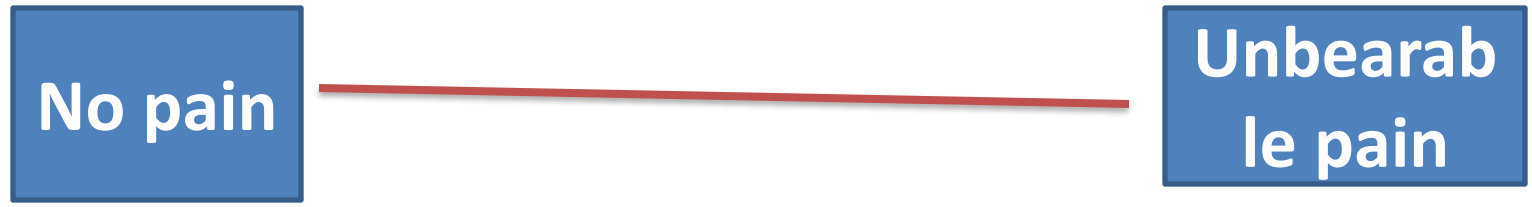
Ex: Number at children in a family.

- Continuous data: data are usually obtained by some format at measurements.

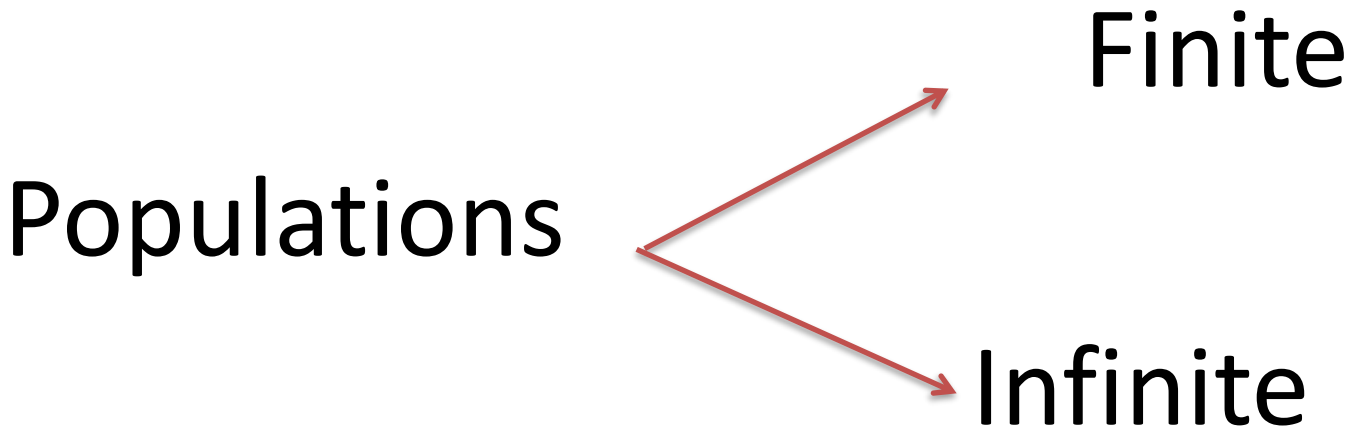
Ex: Height, Weight ,age.

Other type of data

1. Ranks. Ex:drug1, drug2, drug3
2. Percentages. Ex: 85%, 90%
3. Rations. Ex: death rate.
4. Scores. Ex: patients classify a skin rash
5. Scale. Ex:



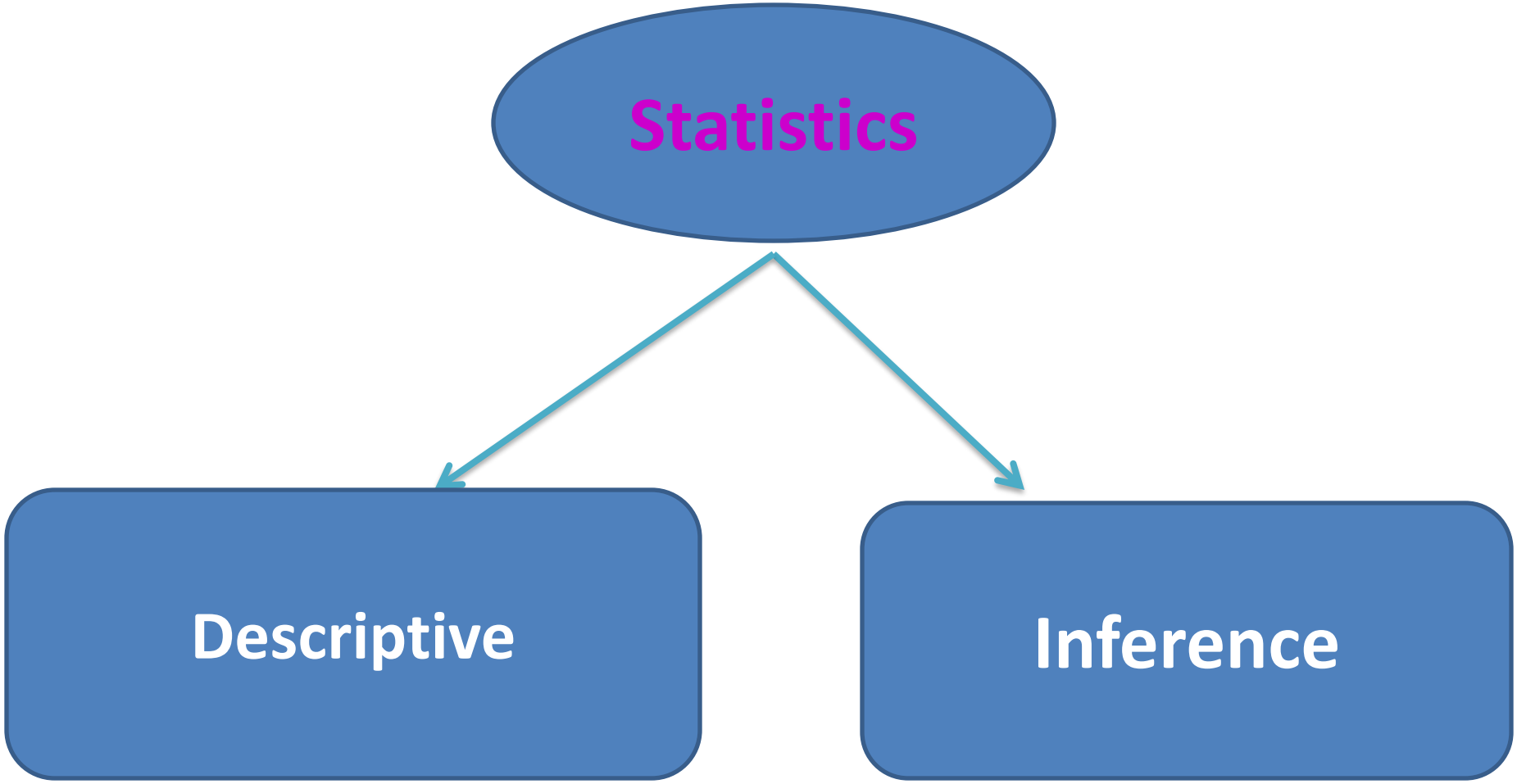
Populations



Population

- **Finite**: It is possible to collect data on everyone in the population
- Example student's height at university of Kufa

- **Infinite**: It is not possible to collect data on everyone in the population
- Example: number of bacteria in a study field.


Statistics
Descriptive
Inference

Thank You

Methods of presentation of data

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2023-2024
ghassand7@Gmail.com
gmohameed@atu.edu.iq

Methods of presentation of data

1. Numerical presentation
2. Graphical presentation
3. Mathematical presentation

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Title

Name of variable (Units of variable)	Frequency	%
- - Categories -		
Total		

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

- Ex: The following data shows the age of 30 patients

23	27	20	22	33	33	43	42	39
27	36	43	24	25	26	31	40	43
30	35	27	29	28	34	21	27	26
32	37	37						

- **From these data Find *A frequency distribution table***

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Sol:

1- simple size (n)=30

2- Range=max value- min value = 43-20 =23

3- Number of class= 1+3.322*log(n)*=1+3.322*log(30)* =1+3.322*(1.477)=6*

4- Leangth of class=R(Range)+1/K(number of class)=23+1/6=4

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Class (5)	Frequency (f_i) (6)
20-23	4
24-27	8
28-31	4
32-35	5
36-39	4
40-43	5

EX: 50 patients *at the surgical department of Alexandria hospital in May 2008* according to their ABO blood groups

no.	ABO blood	no.	ABO blood	no.	ABO blood	no.	ABO blood	no.	ABO blood
1	A	11	AB	21	B	31	AB	41	A
2	O	12	B	22	B	32	B	42	B
3	O	13	AB	23	B	33	B	43	B
4	AB	14	A	24	AB	34	O	44	O
5	B	15	A	25	A	35	O	45	O
6	B	16	B	26	O	36	O	46	A
7	A	17	O	27	A	37	A	47	B
8	A	18	O	28	O	38	A	48	B
9	O	19	O	29	B	39	B	49	B
10	O	20	B	30	A	40	O	50	B

Table (I): Distribution of 50 patients *at the surgical department of Alexandria hospital in May 2008* according to their ABO blood groups

Blood group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30
Total	50	100

EX: The 20 lung cancer patients at the chest department of Alexandria hospital and 40 controls in May 2008 according to smoking

no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer
1	**Smoker**	**Cases**	**13**	**Smoker**	**Cases**	**25**	**Smoker**	**Control**	**37**	**Non Smoker**	**Control**	**49**	**Non Smoker**	**Control**
2	**Smoker**	**Cases**	**14**	**Smoker**	**Cases**	**26**	**Smoker**	**Control**	**38**	**Non Smoker**	**Control**	**50**	**Non Smoker**	**Control**
3	**Smoker**	**Cases**	**15**	**Smoker**	**Cases**	**27**	**Smoker**	**Control**	**39**	**Non Smoker**	**Control**	**51**	**Non Smoker**	**Control**
4	**Smoker**	**Cases**	**16**	**Non Smoker**	**Cases**	**28**	**Smoker**	**Control**	**40**	**Non Smoker**	**Control**	**52**	**Non Smoker**	**Control**
5	**Smoker**	**Cases**	**17**	**Non Smoker**	**Cases**	**29**	**Non Smoker**	**Control**	**41**	**Non Smoker**	**Control**	**53**	**Non Smoker**	**Control**
6	**Smoker**	**Cases**	**18**	**Non Smoker**	**Cases**	**30**	**Non Smoker**	**Control**	**42**	**Non Smoker**	**Control**	**54**	**Non Smoker**	**Control**
7	**Smoker**	**Cases**	**19**	**Non Smoker**	**Cases**	**31**	**Non Smoker**	**Control**	**43**	**Non Smoker**	**Control**	**55**	**Non Smoker**	**Control**
8	**Smoker**	**Cases**	**20**	**Non Smoker**	**Cases**	**32**	**Non Smoker**	**Control**	**44**	**Non Smoker**	**Control**	**56**	**Non Smoker**	**Control**
9	**Smoker**	**Cases**	**21**	**Smoker**	**Control**	**33**	**Non Smoker**	**Control**	**45**	**Non Smoker**	**Control**	**57**	**Non Smoker**	**Control**
10	**Smoker**	**Cases**	**22**	**Smoker**	**Control**	**34**	**Non Smoker**	**Control**	**46**	**Non Smoker**	**Control**	**58**	**Non Smoker**	**Control**
11	**Smoker**	**Cases**	**23**	**Smoker**	**Control**	**35**	**Non Smoker**	**Control**	**47**	**Non Smoker**	**Control**	**59**	**Non Smoker**	**Control**
12	**Smoker**	**Cases**	**24**	**Smoker**	**Control**	**36**	**Non Smoker**	**Control**	**48**	**Non Smoker**	**Control**	**60**	**Non Smoker**	**Control**

Complex frequency distribution Table

Table (III): Distribution of 20 lung cancer patients at the chest department of Alexandria hospital and 40 controls in May 2008 according to smoking

Smoking	Lung cancer				Total	
	Cases		Control			
	No.	%	No.	%	No.	%
Smoker	15	75%	8	20%	23	38.33
Non smoker	5	25%	32	80%	37	61.67
Total	20	100	40	100	60	100

EX: The 60 patients *at the chest department of Alexandria hospital in May 2008* according to smoking & lung cancer

no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer	no.	Smoking	Lung cancer
1	Smoker	positive	13	Smoker	positive	25	Smoker	negative	37	Non Smoker	negative	49	Non Smoker	negative
2	Smoker	positive	14	Smoker	positive	26	Smoker	negative	38	Non Smoker	negative	50	Non Smoker	negative
3	Smoker	positive	15	Smoker	positive	27	Smoker	negative	39	Non Smoker	negative	51	Non Smoker	negative
4	Smoker	positive	16	Non Smoker	positive	28	Smoker	negative	40	Non Smoker	negative	52	Non Smoker	negative
5	Smoker	positive	17	Non Smoker	positive	29	Non Smoker	negative	41	Non Smoker	negative	53	Non Smoker	negative
6	Smoker	positive	18	Non Smoker	positive	30	Non Smoker	negative	42	Non Smoker	negative	54	Non Smoker	negative
7	Smoker	positive	19	Non Smoker	positive	31	Non Smoker	negative	43	Non Smoker	negative	55	Non Smoker	negative
8	Smoker	positive	20	Non Smoker	positive	32	Non Smoker	negative	44	Non Smoker	negative	56	Non Smoker	negative
9	Smoker	positive	21	Smoker	negative	33	Non Smoker	negative	45	Non Smoker	negative	57	Non Smoker	negative
10	Smoker	positive	22	Smoker	negative	34	Non Smoker	negative	46	Non Smoker	negative	58	Non Smoker	negative
11	Smoker	positive	23	Smoker	negative	35	Non Smoker	negative	47	Non Smoker	negative	59	Non Smoker	negative
12	Smoker	positive	24	Smoker	negative	36	Non Smoker	negative	48	Non Smoker	negative	60	Non Smoker	negative

Complex frequency distribution Table

Table (IV): Distribution of 60 patients *at the chest department of Alexandria hospital in May 2008* according to smoking & lung cancer

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

Thank You

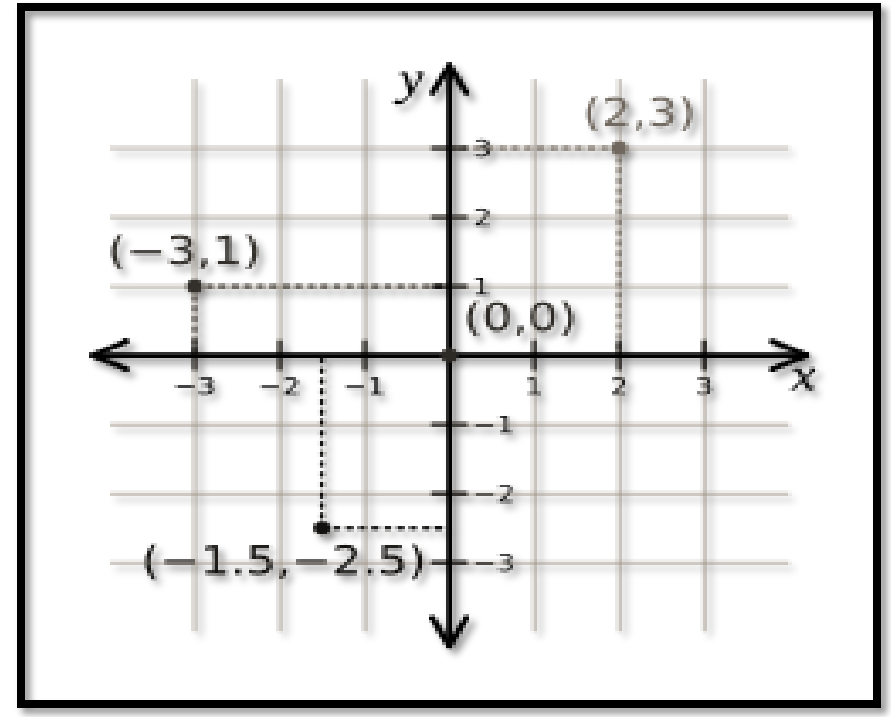
Graphical presentation

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2024-2025
ghassand7@Gmail.com
gmohameed@atu.edu.iq

2- Graphical presentation

1. *Graphs drawn using Cartesian coordinates*

- Line graph
- Frequency polygon
- Frequency curve
- Histogram
- Bar graph
- Scatter plot

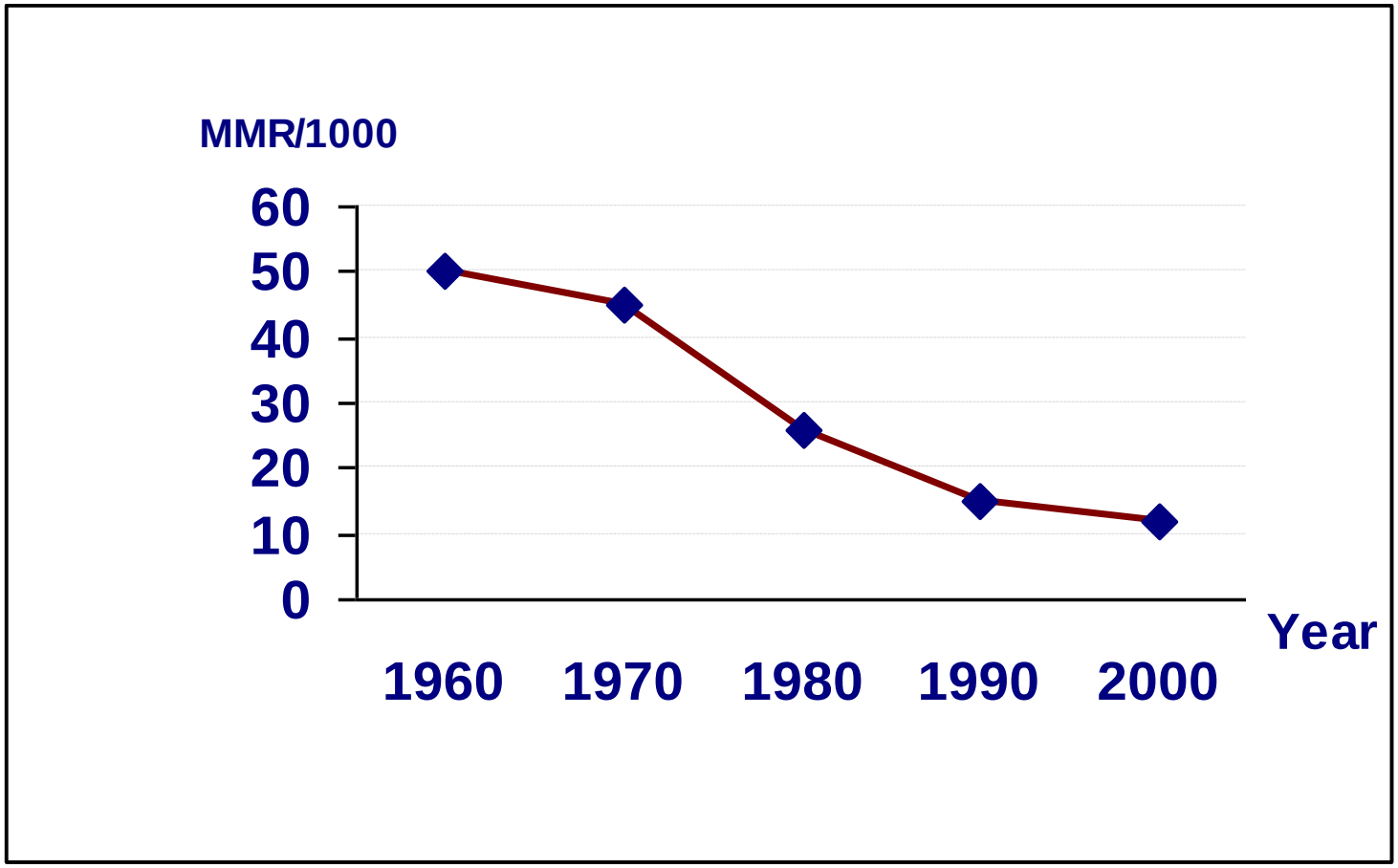


2. *Pie chart*

3. *Statistical maps*

rules

Line Graph



Year	MMR
1960	50
1970	45
1980	26
1990	15
2000	12

Figure (1): Maternal mortality rate of (country), 1960-2000

Frequency polygon

Age (years)	Sex		Mid-point of interval
	Males	Females	
20 -	3 (12%)	2 (10%)	(20+30) / 2 = 25
30 -	9 (36%)	6 (30%)	(30+40) / 2 = 35
40-	7 (32%)	5 (25%)	(40+50) / 2 = 45
50 -	4 (16%)	3 (15%)	(50+60) / 2 = 55
60 - 70	2 (8%)	4 (20%)	(60+70) / 2 = 65
Total	25(100%)	20(100%)	

Age	Sex		M-P
	M	F	
20-	(12%)	(10%)	25
30-	(36%)	(30%)	35
40-	(8%)	(25%)	45
50-	(16%)	(15%)	55
60-70	(8%)	(20%)	65

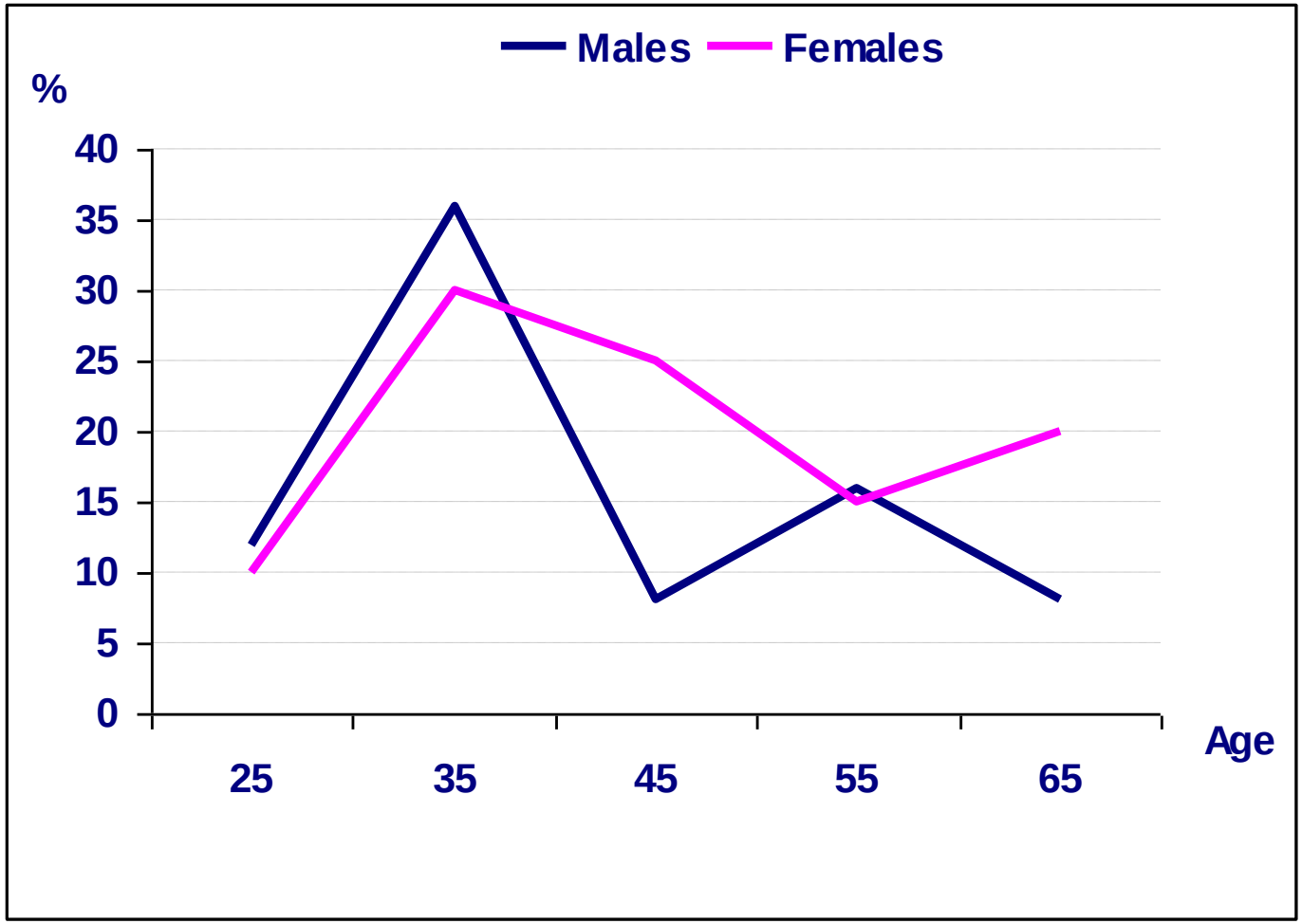


Figure (2): Distribution of 45 patients at (place) , in (time) by age and sex

Frequency curve

Distribution of a group of cholera patients by age

Histogram

Age (years)	Frequency	%
25-	3	14.3
30-	5	23.8
40-	7	33.3
45-	4	19.0
60-65	2	9.5
Total	21	100

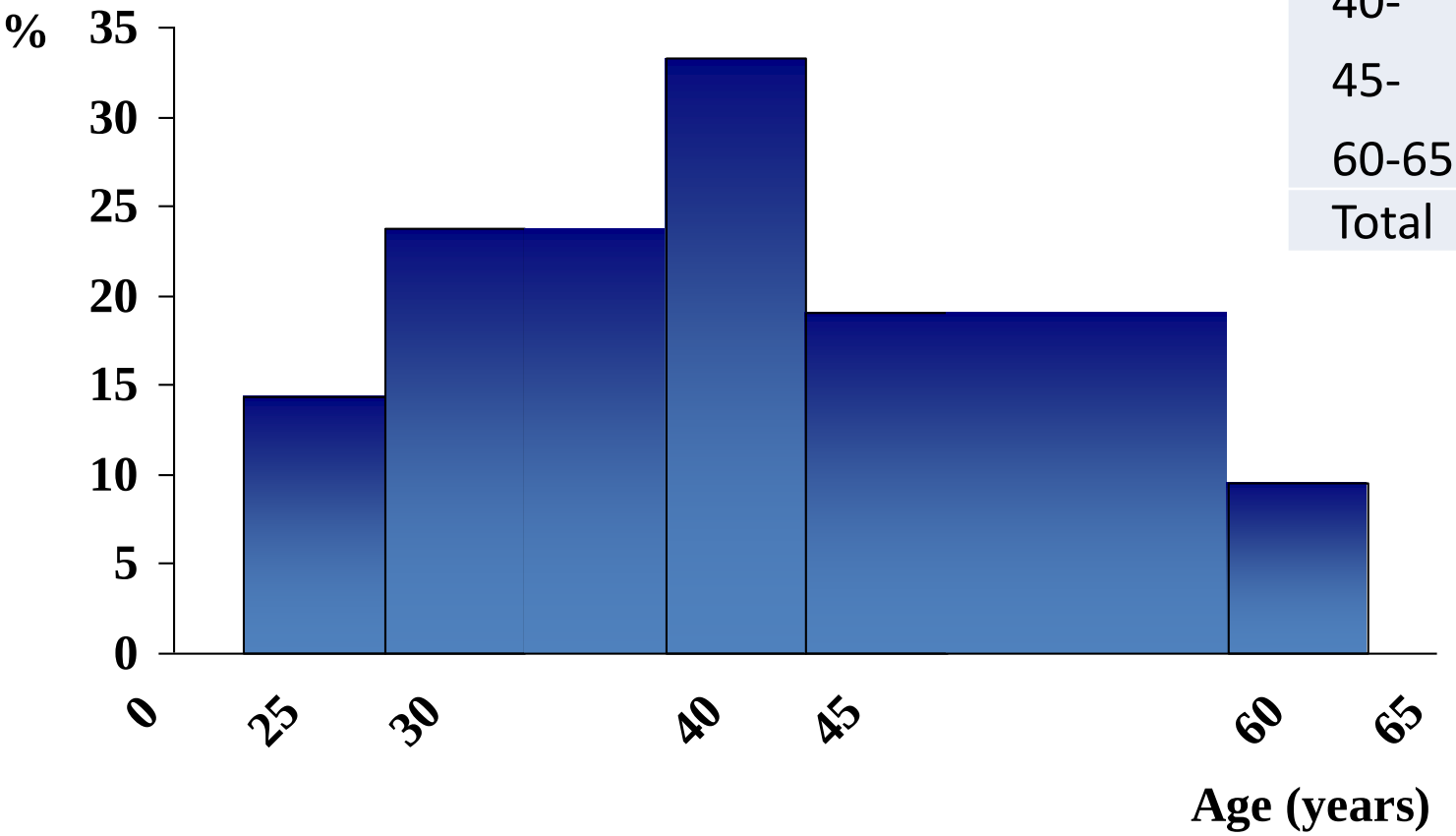
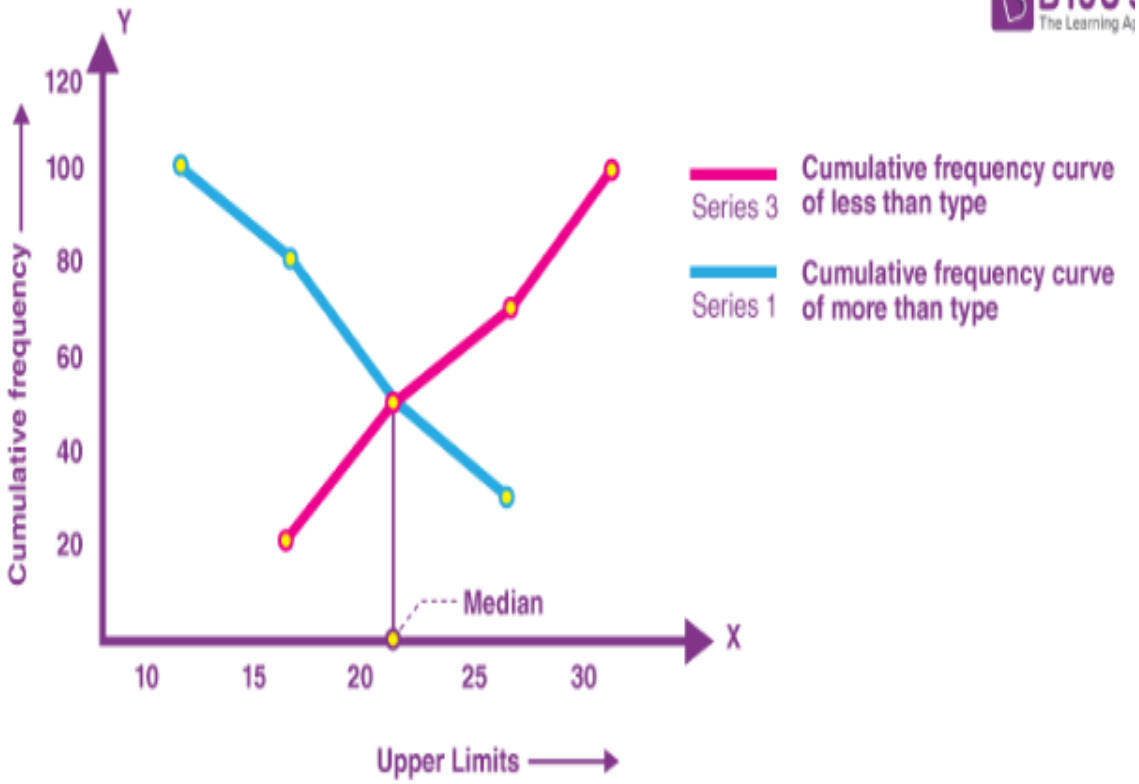


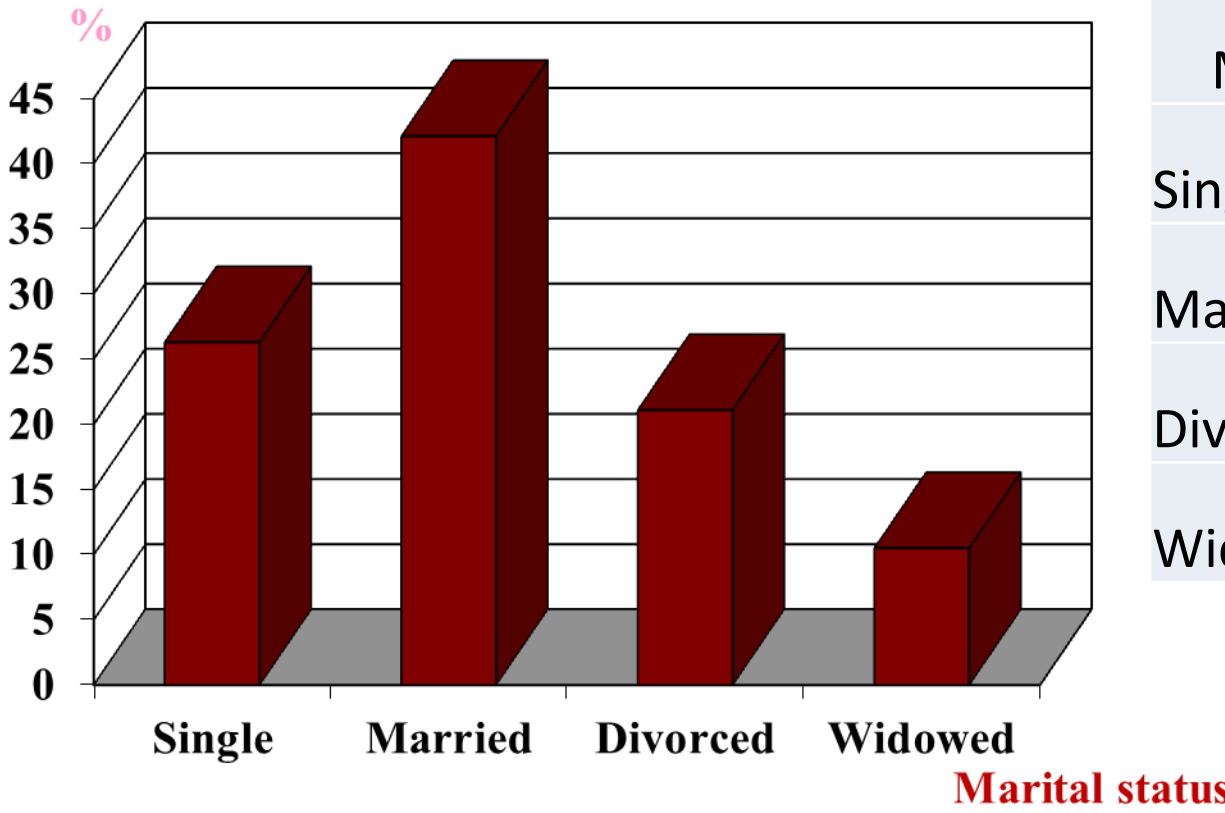
Figure (2): Distribution of 100 cholera patients at (place) , in (time) by age

Upper Cumulative Frequency (U. C. F)
Lower Cumulative Frequency (L. C. F)



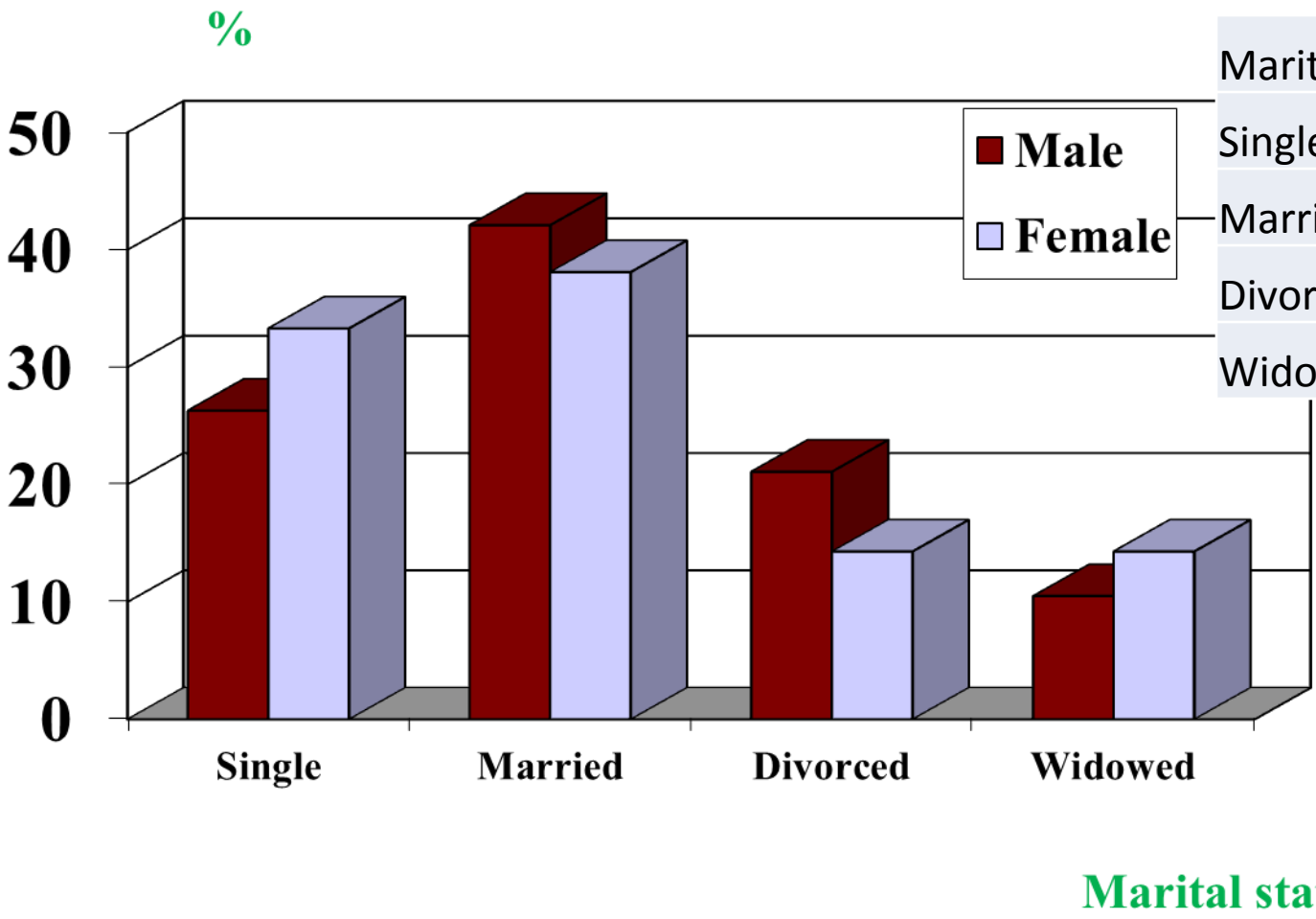
Age Group	F_i	Upper Cumulative Frequency		Lower Cumulative Frequency	
		Age group	Cumulative Frequency	Age group	Cumulative Frequency
10-15	20	-15	20	10-	100
15-20	32	-20	52	15-	80
20-25	18	-25	70	20-	48
25-30	30	-30	100	25-	30

Bar chart



Marital status	%
Single	26.3
Married	42.1
Divorced	21.1
Widowed	10.5

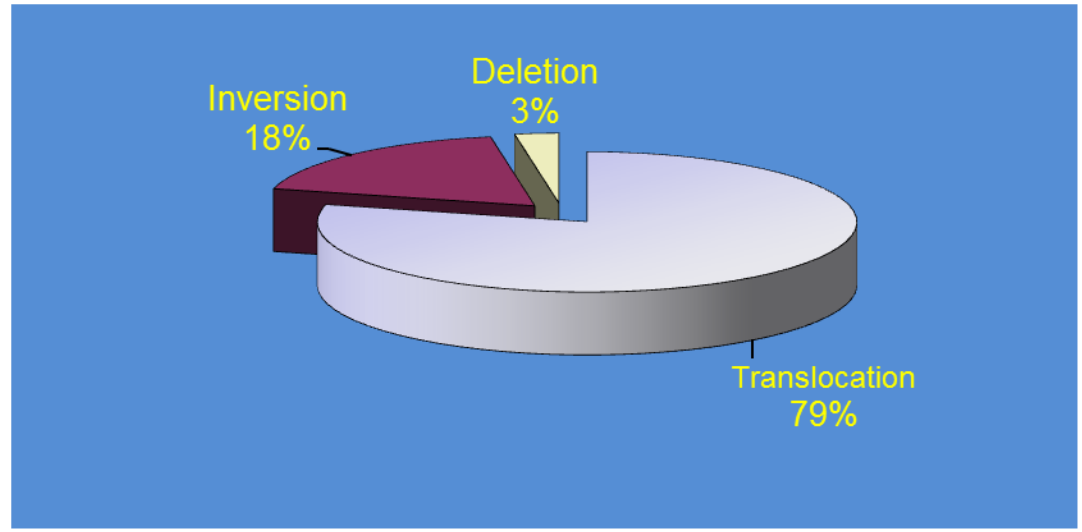
Bar chart



Marital status	Male %	Female %
Single	26.3	33.6
Married	42.1	39.4
Divorced	21.1	13.5
Widowed	10.5	13.5

Pie chart

Translocation	79
Inverse	18
Deletion	3



Thank You

Mathematical presentation

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2024-2025
ghassand7@Gmail.com
gmohameed@atu.edu.iq

3-Mathematical presentation
Univariate Descriptive Statistics

❖Why do we need descriptive statistics

- We use the label *univariate descriptive statistics to refer to a variety of measures of center and variation that are useful for understanding the nature and distribution of a single variable.*
- They can allow us to quickly understand a large amount of information about a single variable.
- *They make data meaningful !*

The Menu of Basic Descriptive Statistics

- **Measures of central tendency**
  - **Mean, Median, Mode, Midrange**
- **Measures of distribution**
  - **Range, Min, Max, Percentiles**
- **Measures of Variation**
  - **Standard Deviation, Variance, Coefficient of Variation**

Measures of Central Tendency - Mean

- **The sample mean is the mathematical average of the data and is the measure of central tendency we use most often.**

- $$\text{sample Mean: } \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of Central Tendency - Mean

Observation #	Age of Volunteer
1	15
2	17
3	17
4	19
5	22
6	26
7	39
	155

Sample Mean:

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

$$\bar{X} = \frac{155}{7}$$

$$\bar{X} = 22.14$$

The sum of all of the observations

n = the number of observations

The sample mean

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

f_i	Class	x_i	f_i
f_1 f_2 . . . f_k	a_1-b_1 a_2-b_2 . . . a_k-b_k	x_1 x_2 . . . x_k	f_1 f_2 . . . f_k
$\sum_{i=1}^{k} f_i$			

Ex: Find the arithmetic mean of the following data

f_i		x_i	$f_i x_i$
7	37-39	38	266
2	40-42	41	82
4	43-45	44	176
2	46-48	47	94
15			618

$$\bar{X} = \frac{618}{15} = 41.13$$

Weighted arithmetic mean

$$\overline{X} = \frac{\sum_{i=1}^{k} X_i W_i}{\sum_{i=1}^{k} W_i}$$

X_i	W_i	$X_i W_i$
X_1	W_1	$X_1 W_1$
X_2	W_2	$X_2 W_2$
.	.	.
.	.	.
.	.	.
X_i	W_i	$X_i W_i$
	$\sum_{i=1}^{k} W_i$	$\sum_{i=1}^{k} X_i W_i$

Ex:Find the weighted arithmetic mean of the following data

X_i	W_i	X_iW_i
90	6	540
70	5	350
60	2	120
	13	1010

$$\bar{X} = \frac{1010}{13} = 77.69$$

Thank You

Mathematical presentation (MEDIAN &MODE)

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2024-2025
ghassand7@Gmail.com
gmohameed@atu.edu.iq

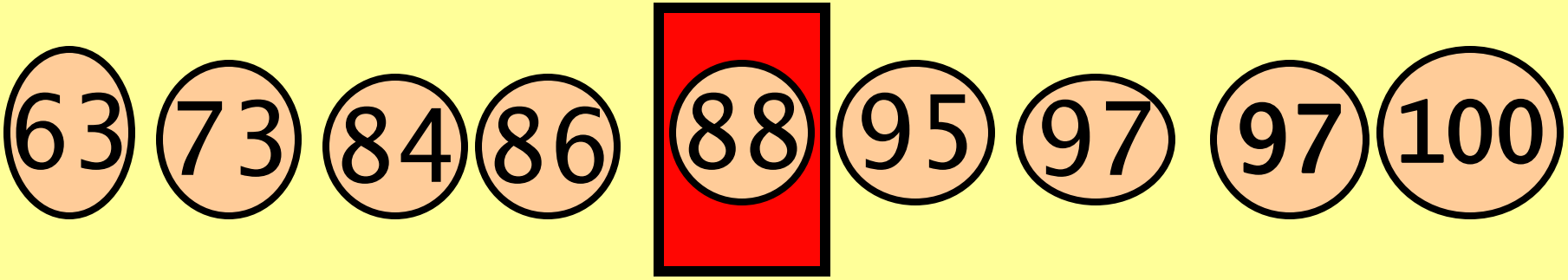
What is the MEDIAN?
How do we find it?

The MEDIAN is the number that is in the middle of a set of data

1. Arrange the numbers in the set in order from least to greatest.

2. Then find the number that is in the middle.

Arrange values from least to greatest.



Find the number that is in the middle.

The median is **88.**

Half the numbers are <u>less</u> than the median.

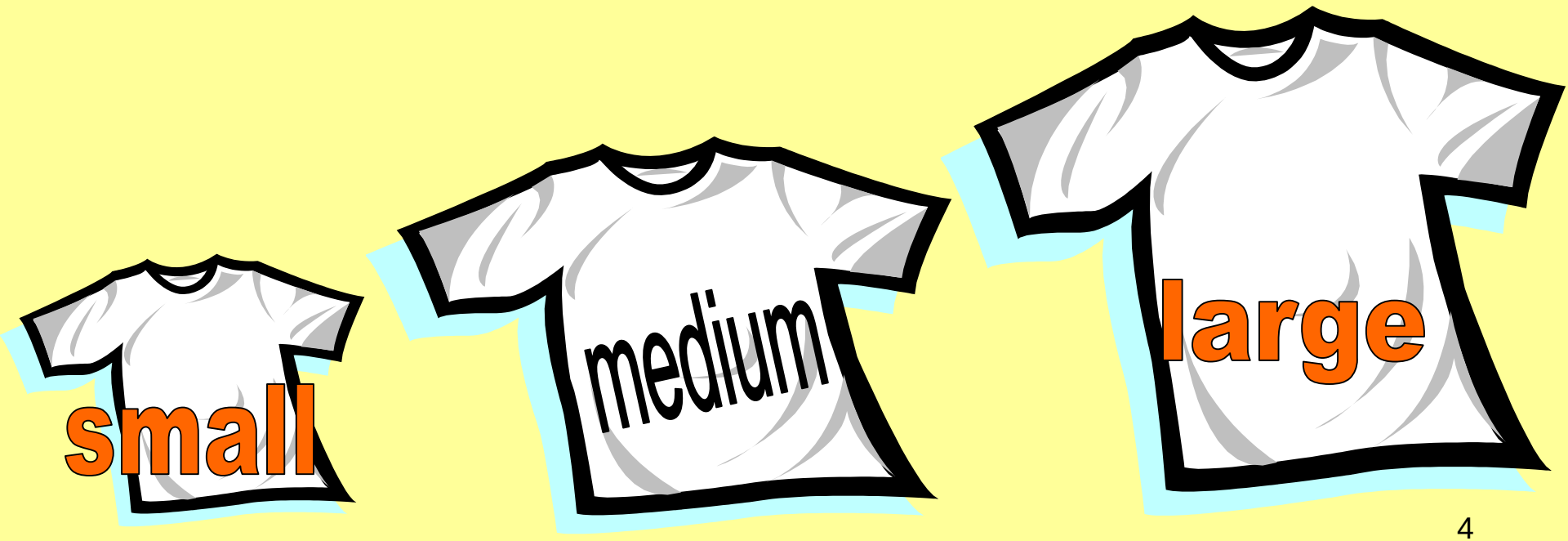
Half the numbers are <u>greater</u> than the median.

Median

Sounds like

MEDIUM

Think **middle** when you hear median.



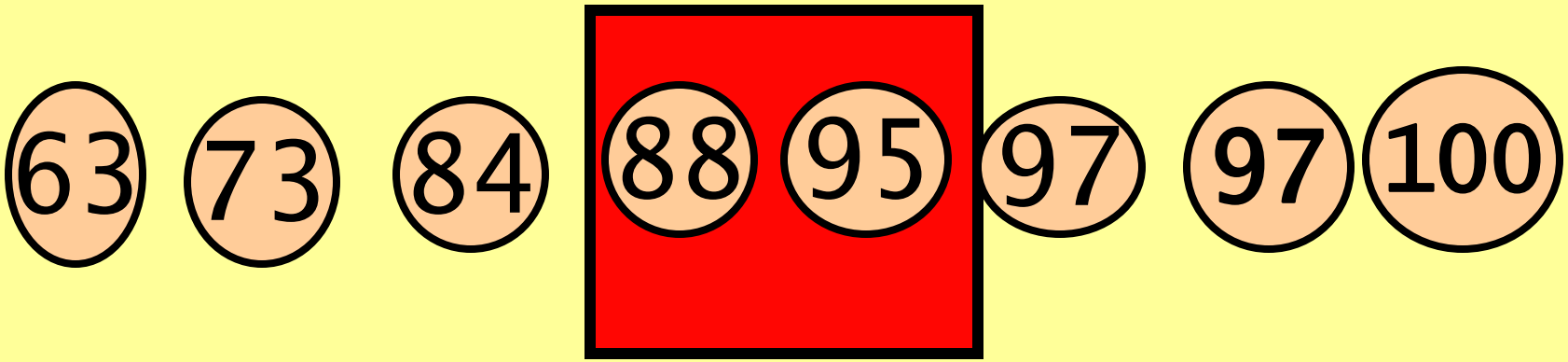
How do we find the MEDIAN

when two numbers are in the middle?

1. **Add the two numbers.**

2. **Then divide by 2.**

Arrange values from least to greatest.



There are two numbers in the middle.

Add the 2 numbers.

88 + 95 = 183

Divide by 2.

183 ÷ 2

The median is **91.5**

Question Size	f
0 - 5	20
5 - 10	24
10 - 15	32
15 - 20	28
20 - 25	20
25 - 30	16
30 - 35	34
35 - 40	10
40 - 45	8

Grouped data

$$m_e = ak + \left(\frac{\frac{\sum_{i=1}^{n} f_i}{2} - c.f}{f_k} \right) * h$$

$ak = lower\ limit\ median\ class$

$c.f = cumulative\ frequency\ of\ class\ prior$

$h = class$ length

$f_k = frequency\ of\ Median\ class$

Answer Step: 1 Class interval	f	C.f
0 - 5	20	20
5 - 10	24	44
10 - 15	32	**76** cf
a_k 15 – 20	**28** f_k	104
20 - 25	20	124
25 - 30	16	140
30 - 35	34	174
35 - 40	10	184
40 - 45	8	1927

$$m_e = ak + \left(\frac{\frac{\sum_{i=1}^{n} f_i}{2} - c.f}{f_k} \right) * h$$

$ak = lower\ limit\ median\ class$

$c.f = cumulative\ frequency\ of\ class\ prior$

$h = class$ length

$f_k = frequency\ of\ Median\ class$

Step: 2

Here,
N= 192, so 192 / 2 = 96
h = 5 cf = 76

Median = N/2 th item which lies in (15 – 20) group.

$$=15 + \frac{96-76}{28} * 5$$

= 15 + 3.58

= 18.58

=> **Median** = 18.58

What is the MODE?

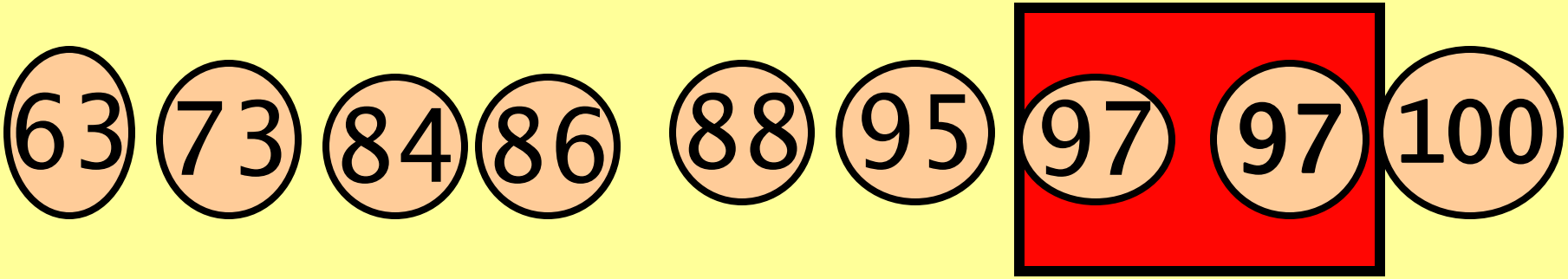
How do we find it?

The MODE is the **piece of data that occurs most frequently in the data set.**

<u>A set of data can have</u>:

- **One mode**
- **More than one mode**
- **No mode**

Arranging values from least to greatest makes it easier to find the mode.



Find the number that appears more or most frequently.

The value 97 appears <u>twice.</u>

All other numbers appear just once.

97 is the MODE

MODE

A Hint for remembering the MODE…

The first two letters give you a hint… MOde

Most **O**ften

MODE

MOST OFTEN

Which set of data has **ONE MODE**?

A 9, 11, 16, 6, 7, 17, 18

B 18, 7, 10, 7, 18

C 9, 11, 16, 8, 16

Which set of data has **NO MODE**?

A 9, 11, 16, 6, 7, 17, 18

B 18, 7, 10, 7, 18

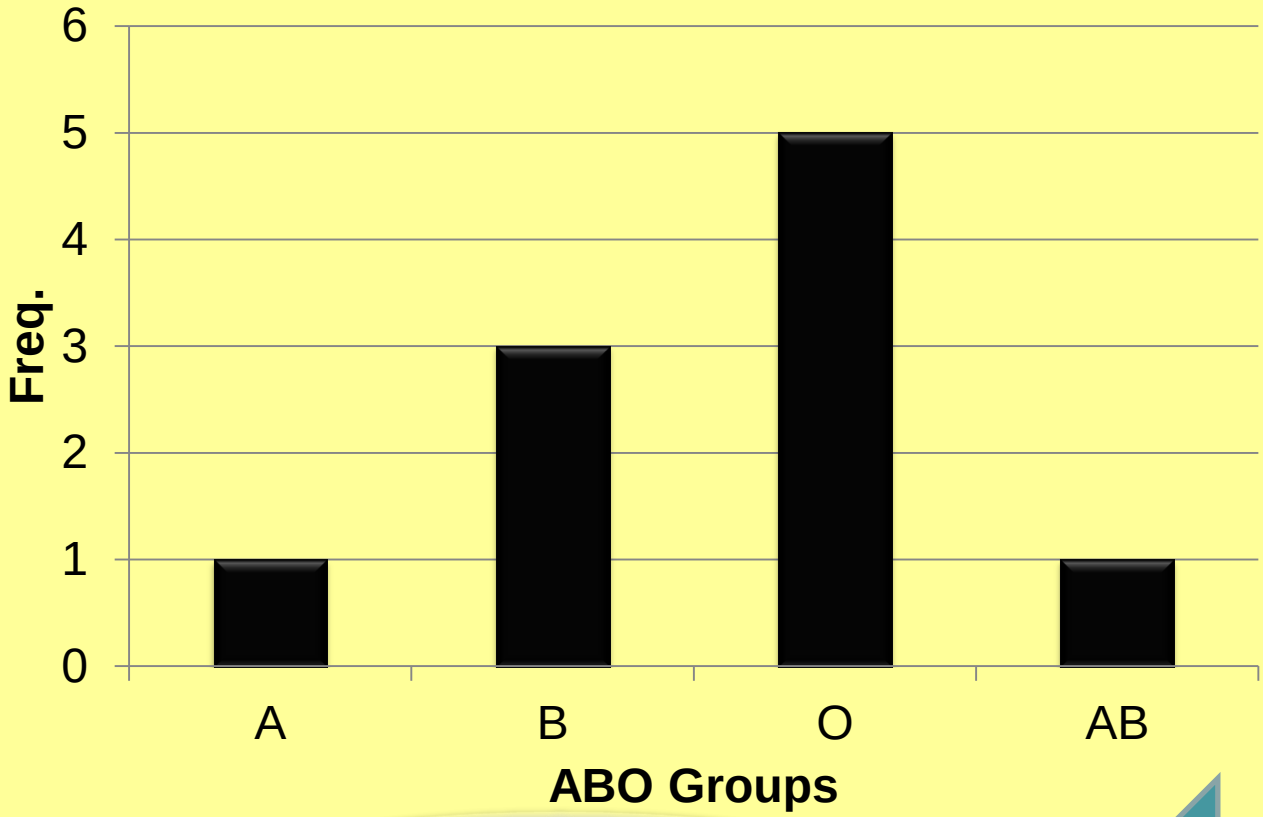
C 13, 12, 12, 11, 12

Which set of data has
MORE THAN ONE MODE?

A 9, 11, 16, 8, 16

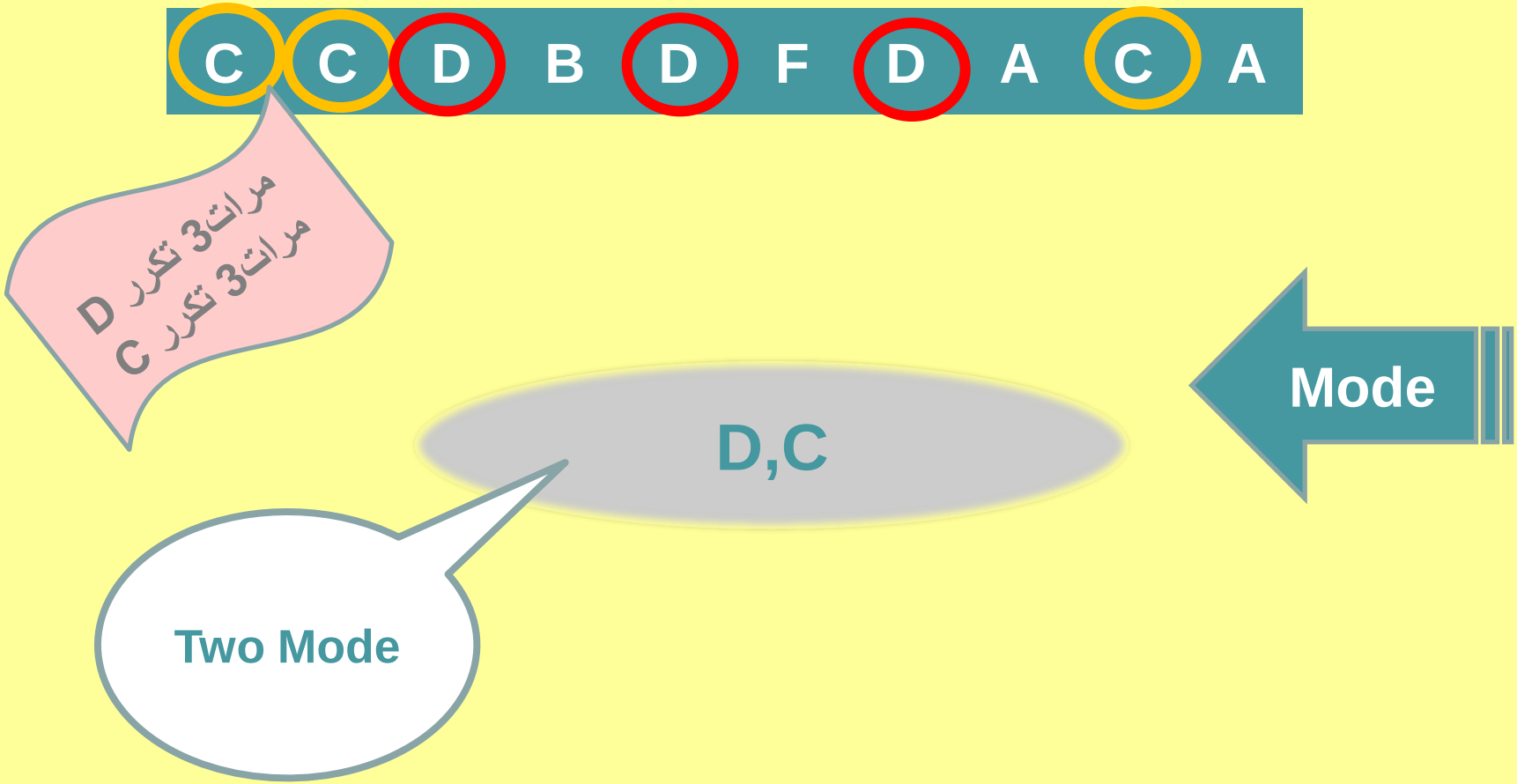
B 9, 11, 16, 6, 7, 17, 18

C 18, 7, 10, 7, 18



O

mode


C C D B D F D A C A
مرات3 تكرر D
مرات3 تكرر C
D,C
Mode
Two Mode

Ex:In ***A frequency distribution table find The*** **MODE**

Age class	F_i
20-30	5
30-40	4
40-50	7
50-60	6
60-70	5
70-80	3
	$\sum f_i = 30$

$$mo = a_k + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) * h = 40 + \left(\frac{3}{3+1}\right) * 10 = 47.5$$

Mean

This one is the requires more work than the others.

Median

Right in the MIDDLE.

Mode

This one is the easiest to find— Just L**OO**K.

Find the….


Mean


Median


Mode

9, 10, 10, 13, 13

Mean

11

Median

10

Mode

10, 13

Find the….


Mean


Median


Mode

8, 8, 9, 10, 10, 12, 12, 13, 17

Mean

11

Median

10

Mode

8, 10, 12

Thank You

Scatter Measures

Measures of distribution

Measures of Variation

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)

2024-2025

ghassan87d@gmail.com
gmohameed@atu.edu.iq

Example

Set 1 : 7 , 8 , 9 , 10 , 11 Mean = 9 Range = 4

Set 2 : 3 , 6 , 9 ,12 ,15 Mean = 9 Range= 12

Set 3 : 1 , 5 , 9 , 13 , 17 Mean = 9 Range =16

Measures of distribution

The Range

The range of a sample is the difference between the highest value and the lowest value.

15 17 17 19 22 26 39

In our example the Range = 39 – 15 or 24; there are 24 years between our youngest and oldest volunteers in the sample.

Measures of Variance

Where measures of central tendency try to give us an idea of where the middle of the data lies, measures of variance (or variation) tell us about how the data is distributed around that center.

Our three primary measures of variance are:

- Mean Deviation
–Standard Deviation,
–Variance and
–Coefficient of Variation

IF $X_1, X_2, \ldots\ldots\ldots, X_n$

Than:

$$M.D = \frac{\sum|X_i - \bar{X}|}{n}$$

where $\bar{X}$ is mean

Mean Deviation

If $X_1, X_2, \ldots\ldots\ldots, X_n$ is mid class

And $f_1, f_2 \ldots\ldots f_n$ are frequency

Than: $M.D = \frac{\sum|X_i - \bar{X}|f_i}{n}$

frequency distribution Table

Example: 6 7 10 8 5 4 9 7

$$\bar{X} = \frac{\sum X_i}{n} = \frac{56}{8} = 7$$

Find the mean

$\lvert X_i - 7 \rvert$	$X_i - 7$	$\lvert X_i - 7 \rvert$
6	-1	1
7	0	0
10	3	3
8	1	1
5	-2	2
4	-3	3
9	2	2
7	0	0
$\sum$		12

Mean Deviation

$$M.D = \frac{12}{8} = 1.5$$

Example: find the Mean Deviation for frequency distribution Table

مثال

class	X_i	fi	$X_i f_i$	$X_i - \bar{X}$	$\lvert X_i - \bar{X} \rvert$	$\lvert X_i - \bar{X} \rvert f_i$
2-4	3	2	6	- 8.32	8.32	16.64
5-9	7	5	35	- 4.32	4.32	21.60
10-12	11	9	99	- 0.32	0.32	2.88
13-17	15	7	105	3.68	3.68	25.76
18-20	19	2	38	7.68	7.68	15.36
	$\sum$	25	283			82.24

$$\bar{X} = \frac{\sum X_i f_i}{n} = \frac{283}{25} = 11.32$$

$$M.D = \frac{\sum |X_i - \bar{X}| f_i}{n} = \frac{82.24}{25} = 3.29$$

Thank You

Scatter Measures

Measures of Variation

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)

2024-2025

ghassan87d@gmail.com
gmohameed@atu.edu.iq

Measures of Variance – Standard Deviation

A:Sample Standard Deviation: $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$

Population Standard Deviation: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{N}}$

The Standard Deviation is a measure of the variation of values around the mean.

B: **Sample Standard Deviation:** $S = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{X})^2}{\sum_{i=1}^{n} f_i - 1}}$

Population Standard Deviation: σ = $\sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{X})^2}{\sum_{i=1}^{n} f_i}}$

Some Key Points for Understanding Standard Deviation

- The standard deviation is always positive.
- The standard deviation of a sample will always be in the ***same units*** as the observations in the sample.
- Extreme values or ***outliers*** can change the value of the standard deviation substantially.
- ***The size of the sample will affect*** the size of the standard deviation; as the sample size increases, the size of the standard deviation decreases.

Measures of Variance - Variance

- The variance of a sample is just the standard deviation of the sample squared.

A: Sample Variance: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{N}$

B: **Sample Variance:** $S^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{X})^2}{\sum_{i=1}^{n} f_i - 1}$

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{X})^2}{\sum_{i=1}^{n} f_i}$

Note: *Also can be calculator* standard deviation *by used the*

formula:

$$s = \sqrt{\frac{n\sum(x_i^2) - (\sum x_i)^2}{n(n-1)}}$$

and variance

$$s^2 = \frac{n\sum(x_i^2) - (\sum x_i)^2}{n(n-1)}$$

Back to our example

- In our sample of volunteer ages, the mean was 22.14 years.

15 17 17 19 22 26 39

- We can calculate the standard deviation to better understand how the values or distributed around that mean.

x_i	$\bar{X}$	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
15	22.14	-7.14	50.9796
17	22.14	-5.14	26.4196
17	22.14	-5.14	26.4196
19	22.14	-3.14	9.8596
22	22.14	-0.14	0.0196
26	22.14	3.86	14.8996
39	22.14	16.86	284.2596
			412.8572

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

$$S = \sqrt{\frac{412.8572}{7-1}}$$

$$S = 8.3$$

$$S^2 = 68.89$$

Ex: Find **Sample Standard Deviation** to table following

$$S = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{X})^2}{\sum_{i=1}^{n} f_i - 1}}$$

f_i		M_i	$\bar{X}$	$(M_i - \bar{X})$	$(M_i - \bar{X})^2$	$f_i(M_i - \bar{X})^2$
1	1-2	1.5	4	-2.5	6.25	6.25
1	3-4	3.5	4	-0.5	0.25	0.25
2	5-6	5.5	4	1.5	2.25	4.50
4						11

$$S = \sqrt{\frac{11}{3}}$$

S= 1.915

S^2 =3.667

Measures of Center – Coefficient of Variation

- The Coefficient of Variation (CV) is a measure of the standard deviation of a sample relative to its mean.
- CV's can be useful when you are comparing the standard deviations of variables that are in two different units.

An example: You are comparing the heights and weights of fourth graders.

Height

 = 52"

S = 4" .

Weight

 = *80 lbs*.

S = 10 lbs.

Which variable has greater variance? How can we compare 4” to 10 lbs?

$$CV = \frac{S}{\bar{X}} * 100\%$$

Height

$\bar{X}$ =52

S = 4” .

$$CV = \frac{4}{52} * 100\%$$

CV=8%

Weight

$\bar{X}$ =80 lbs

S = 10 lbs.

$$CV = \frac{10}{80} * 10$$

CV=12.5%

The standard deviation of height is 8% of the mean of height, where as the standard deviation of weight is 12.5% of the mean of weight, so there is greater variation in the weight of the fourth graders than in the height.

Correlation Coefficients

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)

2024-2025

ghassand7@Gmail.com
gmohameed@atu.edu.iq

Correlation Coefficients

The Meaning of Correlation
Correlation and Data Types
Pearson's r
Spearman rho
Other Coefficients of Note
Coefficient of Determination r^2

The concept of correlation was introduced in Chapters 1 and 5. Our focus since Chapter 16 has been basic statistical procedures that measure *differences between groups* -- one-sample, two-sample, and *k*-sample tests.

Now we turn our attention to basic statistical procedures that measure the *degree of association between variables.*

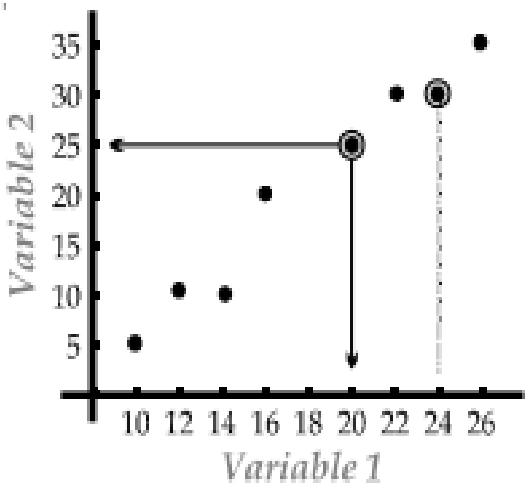
Dr. Wesley Black studied the relationship between rankings of selected learning objectives in a youth discipleship taxonomy between full-time church staff youth ministers and seminary students enrolled in youth education courses at Southwestern Seminary.[1] Questionnaires were returned by 318 students and 184 youth ministers.[2] Ten objectives in each of five categories (Personal Ministry, Christian Theology and Baptist Doctrine, Christian Ethics, Baptist Heritage, and Church Polity and Organization) were ranked by these two groups.

The basic question raised by Black in this study was whether students prioritized discipleship training objectives for youth in the same way as full-time ministers in the field. Using the Spearman rho correlation coefficient, Black found the correlations of rankings generated by students and ministers of the ten items for each of five categories were as follows: Personal Ministry, 0.915; Christian Theology and Baptist Doctrine, 0.867; Christian Ethics, 0.939; Baptist Heritage, 0.939; and Church Polity and Organization, 0.927.

Each of these are strong positive correlations[3] between the rankings of objectives by students and ministers.

The Meaning of Correlation

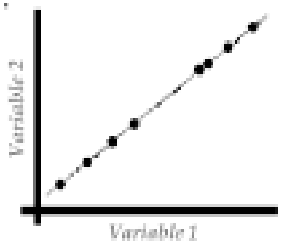
When we discussed the frequency distribution (chapter 15), we plotted *values* of X on the x-axis, and the *frequency* of the X-values on the y-axis. In this plot, there was one score per subject.



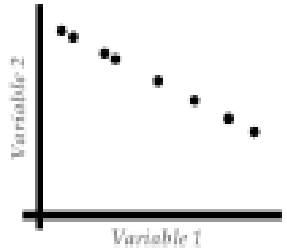
In graphing a correlation between two variables, there are *two scores* **per subject** — an X-score and a Y-score. We plot the X-scores on the x-axis and Y-scores on the y-axis. A single dot represents the intersection between each X-Y pair. Notice the diagram to the left. The single point in the shaded circle represents two scores from a single subject in a study: a ***20*** on variable 1 and a **25** on variable 2.

Notice how the dots form a pattern in two-dimensional space. The tighter the pattern, the higher the correlation. The looser the pattern, the lower the correlation.

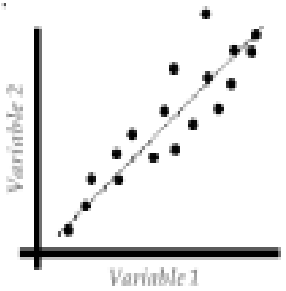
These patterns are called ***scatterplots***. The scatterplots at left illustrate various kinds of associations. The first scatterplot shows a *perfect positive correlation*. The correlation is positive because **variable 2 increases as variable 1 increases**. It is a *perfect* correlation because all the points fall on a straight line. (The line has been included in this diagram, but is not part of the scatterplot.)



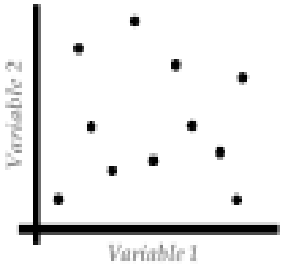
The second scatterplot shows a *perfect negative correlation*. The correlation is negative because **variable 2 *decreases* as variable 1 increases**. It is *perfect* because all the points fall on a straight line (not shown in this diagram).



The third scatterplot shows a *moderately positive correlation*. Notice how most of the data points do not fall on the line. It is a moderate correlation, however, because the points fall in a tight pattern around the line. Notice the pattern is ***linear*** -- that is, a pattern suggesting a line.



The fourth scatterplot shows *no correlation*. Scores on one variable have no systematic association with scores on the other. The scatterplot presents ***no linear pattern*** among the points.



Beyond the graphical representation of association, we can mathematically compute the *degree of association* between two variables. The numerical result of such a computation is called a ***correlation coefficient***. The value of these coefficients usually range from -1.00 to +1.00.

A ***positive*** coefficient indicates that two variables systematically vary in the same direction: *as one variable increases, the other variable tends to increase.* The closer the coefficient is to +1.00, the stronger the positive association.

A ***negative*** coefficient indicates that two variables systematically vary in opposite directions: *as one variable increases, the other variable tends to decrease.* The closer the coefficient is to -1.00, the stronger the negative association.

A coefficient ***close to zero*** indicates that no systematic co-varying exists between the variables. There are several important correlation procedures. They differ according to the data types of the variables.

Correlation and Data Types

Since chapter 16, we have focused on interval or ratio data types. In this chapter, we broaden our focus. There are correlational procedures for all four data types (nominal, ordinal, interval, ratio).

The *Pearson's Product Moment Correlation Coefficient* (r_{xy}) computes the correlation between two interval or ratio variables. *Spearman's rho* (r_s) computes the correlation between two ordinal, or ranked, variables. The *Contingency Coefficient (C) and Cramer's Phi* (ϕ_c) compute the strength of relationship when testing nominal data analyzed by a χI procedure (Chapter 23). The *Phi Coefficient* (r_ϕ) computes the correlation between two ***dichotomous variables*** (two and only two categories (Yes or No, True or False).

Additionally, a study may require the computation of a correlation coefficient between mixed data types. *Point biserial* is used when one variable is interval/ratio and the second is dichotomous. *Rank biserial* is used when one variable is ordinal and the second is dichotomous. We can summarize these various coefficients like this:

Variable 1

	Variable 1			
Variable 2	Interval/Ratio	Ordinal	Nominal	Dichotomous
Interval/Ratio	r_{xy}	r_s*		Point Biserial
Ordinal	r_s*	r_s		Rank Biserial
Nominal			C, ϕ_C**	
Dichotomous	Point Biserial	Rank Biserial		r_ϕ

*requires interval/ratio data to be ranked

**Requires χ^2 value

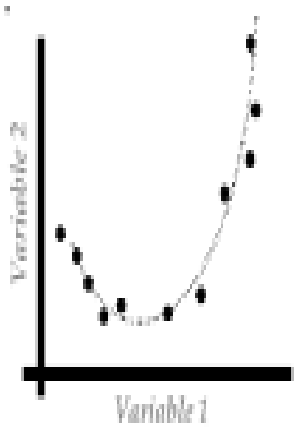
Finally, *Kendall's Coefficient of Concordance (W)* computes the correlation of three or more rankings of of items. Now we'll look at how each of these correlation coefficients are computed.

Pearson's Product Moment Correlation Coefficient (r_{xy})

The most popular correlation coefficient is the Product Moment correlation coefficient, better known as *Pearson's r*. Pearson's r is used to determine the correlation between two variables under three conditions.

First, both variables must be interval or ratio measures (i.e. attitude scales, test scores).

Second, the relationship between the two variables must be linear – the data points must generally fall along a straight line. A non-linear relationship between variables, shown at right, produces a Pearson's r near zero, even though is it clear from the example that there is a strong ("quadratic," of the type $y=x^2$) relationship between the two variables.



The third condition is that both variables are normally distributed. A skewed

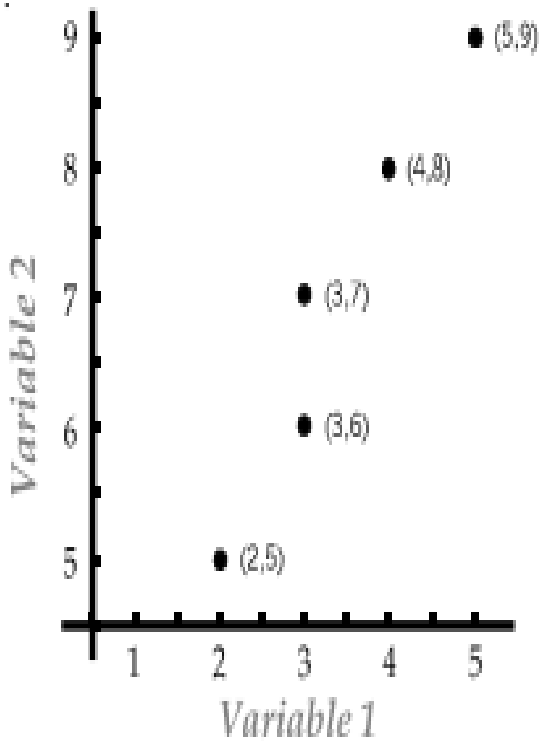
distribution produces a smaller r than a normal distribution.

Use a large scale for variables in correlational analysis, since the larger the variability, the stronger the coefficient will be. A common mistake in research design is to use age categories rather than actual ages, or salary categories rather than actual dollar values. The range of categories will always be much smaller than the range of actual data, reducing the value of r.

Pearson's r is computed with the following formula:

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2][n(\Sigma Y^2) - (\Sigma Y)^2]}}$$

where n equals the number of score-pairs, and X and Y equal paired scores. While this formula is somewhat foreboding, it consists of the following simple components:



ΣXY	Multiply X by Y and sum
ΣX	Sum all the X scores
ΣY	Sum all the Y scores
ΣX^2	Square all X's and sum
ΣY^2	Square all Y's and sum
$(\Sigma X)^2$	Square the sum of X
$(\Sigma Y)^2$	Square the sum of Y

Let's say we have a set of 5 paired scores: (3,6), (5,9), (2,5), (3,7), and (4,8). A scatter-plot of this data is shown at left. From what you can see in this graph, do you predict a strong or weak correlation coefficient?

We've put the paired X-Y values in the chart below to facilitate computing the various elements of the Pearson's r formula. The letters in the chart (A-G) refer to the step below (A-G).

X^2	X	XY	Y	Y^2
9	3	18	6	36
25	5	45	9	81
4	2	10	5	25
9	3	21	7	49
16	4	32	8	64
63	17	126	35	255
ΣX^2	ΣX	ΣXY	ΣY	ΣY^2
F	A	E	C	G

289 $(\Sigma X)^2$ B

1225 $(\Sigma Y)^2$ D

A. Add up the Xs. This is ΣX, and equals 17

B. $(\Sigma X)^2 = 17 \times 17 = 289$
C. Add up the Y's. This is ΣY, and equals 35
D. $(\Sigma Y)^2 = 35 \times 35 = 1225$
E. Multiply each XY pair together and add. This is (ΣXY), and equals 126
F. Square each X and add up the squared values. $\Sigma X^2 = 63$
G. Square each Y and add up the squared values. $\Sigma Y^2 = 255$

Now substituting into the raw score equation we have: *Before going on, be sure to identify each term in the equation above with the chart above and the equation on the previous page.*

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n(\Sigma X^2)-(\Sigma X)^2][n(\Sigma Y^2)-(\Sigma Y)^2]}} = \frac{5(126) - (17)(35)}{\sqrt{[5(63)-(289)][5(255)-(1225)]}}$$

$$= \frac{630 - 595}{\sqrt{[315 - 289][1275 - 1225]}} = \frac{35}{\sqrt{[26][50]}} = \frac{35}{36.056} = 0.971$$

The Pearson r value of +0.971 indicates a very strong -- nearly perfect -- positive correlation between these two variables.

Spearman's rho Correlation Coefficient (r_s)

Spearman's rho yields a correlation coefficient between two ordinal, or ranked, variables.[4] The formula is:

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

where D is the difference between paired ranks. The number "6" is a constant.

Suppose a pastor asked two staff members to rank ten church objectives according to how well they were being accomplished by the church. Here are the rankings of the ministers.

Objective	*Min Ed Rank*	*Min Youth Rank*
1	1	2
2	2	1
3	3	5
4	4	3
5	5	7
6	6	6
7	7	4
8	8	10
9	9	9
10	10	8

Question: Do these two staff members agree in their evaluation of the objectives?

What is the strength of their agreement?

First we compute the differences (D) between ranks, then square the differences (D^2), sum the squares (D^2), and substitute into the formula. The table below summarizes the process:

Objective	Min Ed Rank		Min Youth Rank		D	D^2
1	1	-	2	=	-1	1
2	2		1		+1	1
3	3		5		-2	4
4	4		3		+1	1
5	5		7		-2	4
6	6		6		0	0
7	7		4		+3	9
8	8		10		-2	4
9	9		9		0	0
n= 10	10		8		+2	4
						$\Sigma D^2=28$

Objective 1 was ranked highest (1) by the minister of education and second (2) by the minister of youth. Subtracting 2 from 1 yields a difference (D) of **-1**. Squaring D yields a D^2 of **1**. Notice that the sum of differences (D) equals 0.

Summing the D^2 values, we get **28**.

Substituting the value of $\mathbf{D^2}$ and **n** into the Spearman formula, we have

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6(28)}{10(99)} = 1 - 0.17 = 0.83$$

The coefficient of +0.83 indicates a strong agreement between the two staff ministers with respect to the rankings of church objectives.

Other Important Correlation Coefficients

Several other correlation coefficients will be mentioned. These will be described but not illustrated by example in the interest of space.

Point Biserial Coefficient

The point biserial correlation coefficient is computed between one interval or ratio variable and one dichotomous variable. The term "biserial" refers to the fact that there are two groups of persons (X= 0,1) being observed on the continuous variable (Y).

Use this procedure to test correlations of attitude scores or test scores of subjects between "haves" and "have-nots": ministers who graduated from seminary and those who did not, preschoolers who have had a specified early education program and those who haven't, staff members who have a specified evaluation procedure and those who do not, and so forth.

Rank Biserial Coefficient

The rank biserial correlation coefficient is much like the point biserial just discussed, except that it uses an *ordinal variable* in place of an interval/ratio variable. The coefficient measures degree of relationship between a dichotomous condition (1,0) and a ranking.

Phi Coefficient (r_ϕ)

The **Phi Coefficient** measures the strength of relationship between two dichotomous variables. A study of marital status and attrition rate in college might arbitrarily assign a "1" to married and "0" to not married; a "1" to dropped out and a "0" to remaining in school. Any type of variable that can be classified "1" and "0" can use the phi coefficient.

A positive correlation indicates those who score "1" on one variable tend to score "1" on the other. Using the example above, a positive correlation would mean that married students (1) tend to drop out of school (1) more than unmarried students.

Thank You

Introduction to Regression

Msc. Ghassan Dhahir Al-Thabhawee
(M.sc. in University of Kufa)
2024-2025
ghassand7@Gmail.com
gmohameed@atu.edu.iq

Objectives

In this chapter, you learn:

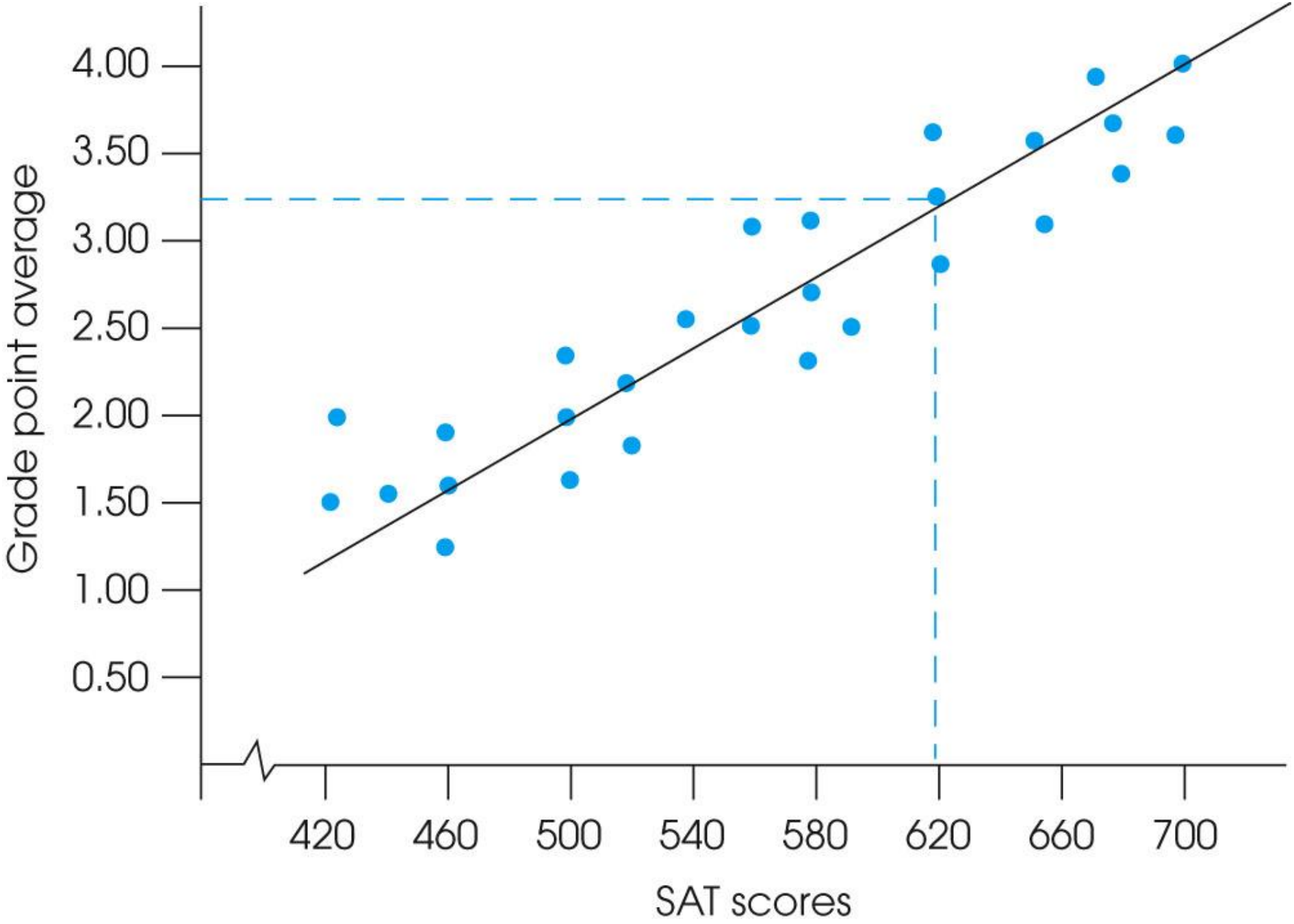
- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable
- To understand the meaning of the regression coefficients a and b
- To evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values

Introduction to Linear Regression

- The Pearson correlation measures the degree to which a set of data points form a straight line relationship.
- **Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.

Introduction to Linear Regression (cont.)

- Any straight line can be represented by an equation of the form $Y = a + bX$, where b and a are constants.
- The value of b is called the slope constant and determines the direction and degree to which the line is tilted.
- The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.


Grade point average
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
420
460
500
540
580
620
660
700
SAT scores

Introduction to Linear Regression

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable we wish to predict or explain
- Independent variable: the variable used to predict or explain the dependent variable

Introduction to Linear Regression

The equation for the regression line is

$$\hat{y} = a + bx$$

whare

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n}$$ and $$a = \bar{y} - b\bar{x}.$$

Simple Regression Example

The following data are diastolic blood pressure (DBP) measurements taken at different times after an intervention for n = 5 persons. For each person, the data available include the time of the measurement and the DBP level. Of interest is the relationship between these two variables.

	Time		DPB		
Patient	x	x^2	y	y^2	xy
1	0	0	72	5,184	0
2	5	25	66	4,356	330
3	10	100	70	4,900	700
4	15	225	64	4,096	960
5	20	400	66	4,356	1,320
Sum	50	750	338	22,892	3,310
Mean	10		67.6		
n	5		5		

For the blood pressure data,

$$\bar{x} = 50/5 = 10,$$

$$\bar{y} = 338/5 = 67.6,$$

the slope is

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n}$$

$$b = \frac{3,310 - (50)(338)/5}{750 - (50)^2/5} = -0.28$$

and the intercept is

$$a = \bar{y} - b\bar{x},$$

$$a = 67.6 - (-0.28)10 = 70.4$$

$$\hat{y} = a + bx = 70.4 - 0.28x$$

The best line is

Patient	Time x	DBP y
1	0	72
2	5	66
3	10	70
4	15	64
5	20	66