TOPIC 2: Correlation and scatter graphs
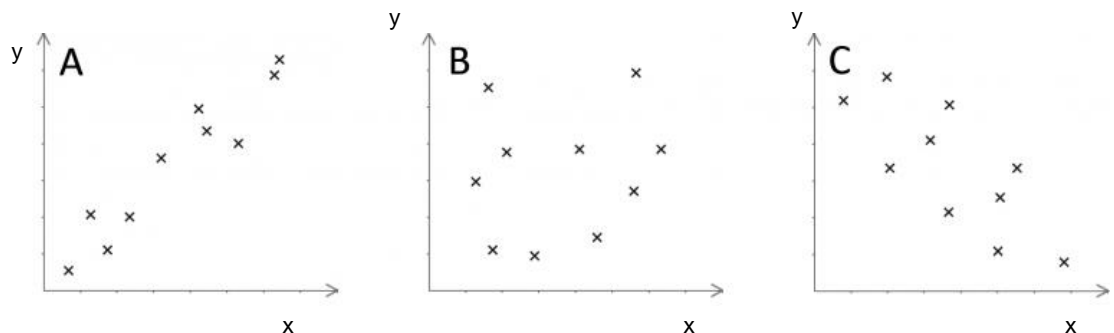
Correlation is the relationship between two variables. The measure of the degree of the relationship between two variables the is correlation coefficient

Scatter graphs:

Is a graphical representation of two variables for a given set of data.

Types of correlation

a) Positive/direct correlation (A): If one variable, x, increases as the other, y, increases, or, as x decreases, y also decreases then the relationship between the two variables is said to positive.

b) Negative / inverse correlation(B): If one variable, x, increases as y reduces, or, x reduces as y increases then the relationship between two variables is said to be negative

c) Zero/no correlation (C): If an increase or decrease in one variable does not affect the other, then the variables are said to have a zero correlation.

Line of best fit

This refers to a line through a scatter plot of data points that best expresses the relationship between these points. To draw this line the following steps are followed:

i)   Find the mean point $M(\bar{X}, \bar{Y})$, where $\bar{X} = \dfrac{\sum X}{n}$ and $\bar{Y} = \dfrac{\sum Y}{n}$

         Where n is the number of pairs of X and Y

ii) Draw the line through the mean point ensuring that there are equal points left on either side of the line

Rank correlation coefficient:

It is the measure of the degree of the relationship between two variables. It is determined on the basis of ranks. There many correlation coefficients. Spearman's correlation coefficient is more preferred.

It is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Where n=number of pairs of the variables

d= difference between the ranks

Interpretation of magnitude of correlation coefficient:

a) **Using magnitude and direction**

Correlation coefficient	comment
0 – 0.19	Very low
0.2 – 0.39	Low
0.40 – 0.59	Moderate
0.60 – 0.79	high
0.80 – 1.0	Very high

When giving a comment it should have magnitude and direction. For example if $\rho = +0.7$, the comment should be high (magnitude), positive (direction) correlation.

b) **Using level of significance:**

After calculating the correlation coefficient, ρ_C, the comment is made basing on the standard value of significance levels for correlation coefficient:(use uneb tables)

i) Determine the corresponding standard value of Spearman's correlation coefficient at either 5% or 1% level of significance from the tables

ii) Compare the two values. If the magnitude of the calculated value, $|\rho_c|$, exceeds the standard value then correlation is significant otherwise not.

Example 1:

The height (cm) and ages (years) of a random sample of ten farmers are given in the table below.

Heights, x(cm)	156	151	152	160	146	157	149	142	158	140
Ages, y years	47	38	44	55	46	49	45	30	45	30

a) (i) Calculate the rank correlation coefficient
   (ii) Comment on your result at 5% level of significance
b) Plot a scatter diagram for the data and comment on the relationship of the data. Draw a line of best fit
c) Use your graph to find Y when X = 147
   Solution:
   a)

R_X	R_Y	d	d^2
4	3	1	1
6	8	-2	4
5	7	-2	4
1	1	0	0
8	4	4	16

3	2	1	1
7	5.5	1.5	2.25
9	9.5	-0.5	0.25
2	5.5	-3.5	12.25
10	9.5	0.5	0.25
			$\sum d^2 = 41$

i) $\rho = 1 - \dfrac{6\sum d^2}{n(n^2-1)} = 1 - \dfrac{6 \times 41}{10(10^2-1)} = 0.7515$

ii) Since $|\rho_c| = 0.7515 > |\rho_T| = 0.65$, correlation is significant at 5%

Or There is a high positive Correlation between heights and ages.

b) $\bar{X} = \dfrac{1511}{10} = 151.1$, $\bar{Y} = \dfrac{429}{10} = 42.9$. Mean point $(\bar{X}, \bar{Y}) = (151.1, 42.9)$
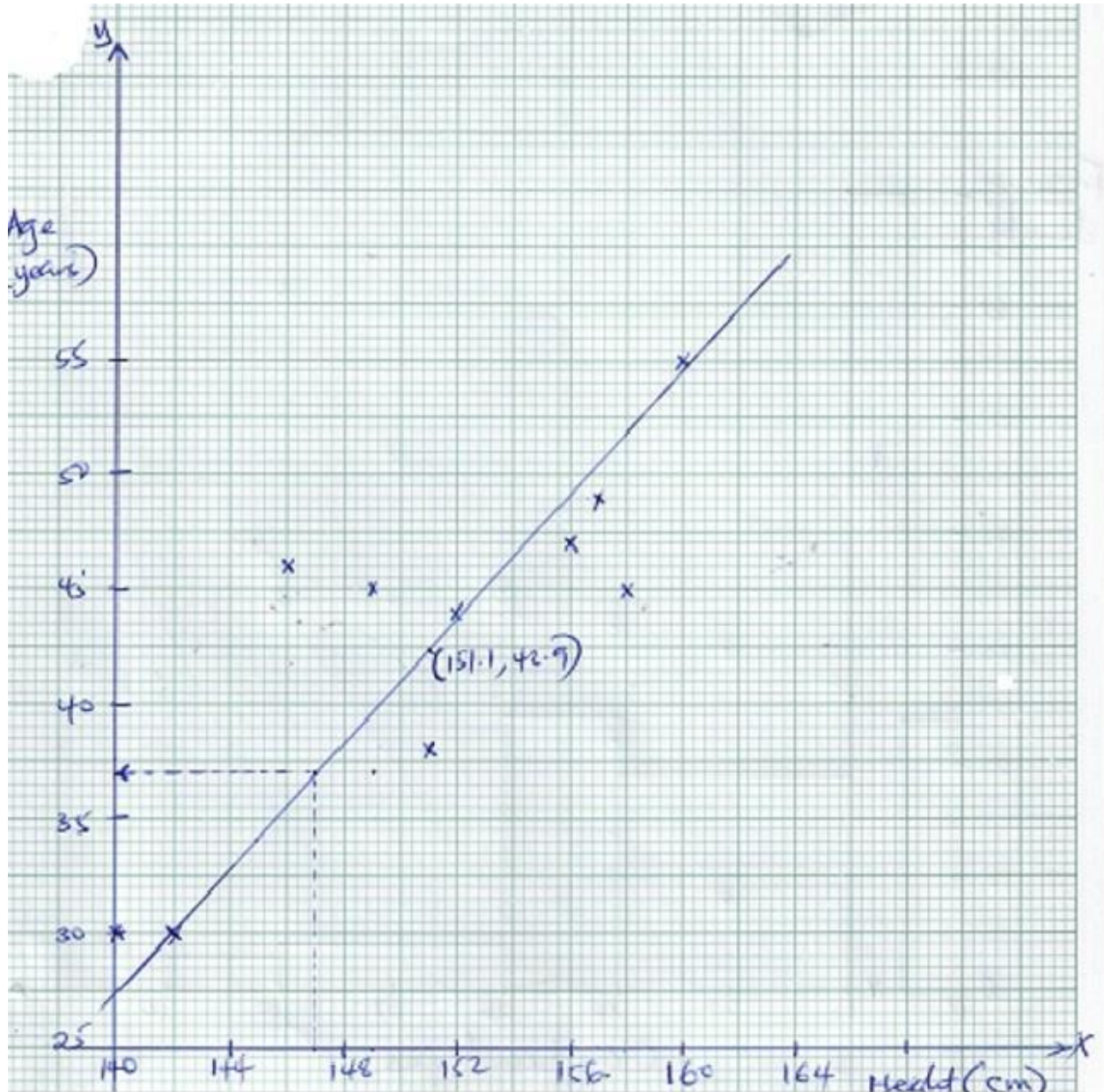
Comment: There is a positive linear correlation between X and Y

c) Y= 37, X=147

Assignment 2.1.2

1. The table below shows scores by candidates in ICT and Mathematics.

ICT	F_9	P_8	C_6	C_4	D_1	C_5	C_4	C_4
Maths	F	E	O	D	A	E	B	C

a.) Calculate the rank correlation for the data.

b.) Test whether the correlation is significant at 5% level of significance.

2. Three examiners X, Y, and Z each marked the scripts of ten candidates who sat Mathematics examination. The table below shows the examiner's ranking of the candidates.

EXAMINER	CANDIDATE									
	A	B	C	D	E	F	G	H	I	J
X	8	5	9	2	10	1	7	6	3	4
Y	5	3	6	1	4	7	2	10	8	9
Z	6	3	7	2	5	4	1	10	9	8

Calculate the coefficient of rank correlation of the rankings between

(i) X and Y
(ii) Y and Z.
iii) State with reason whether there is significant difference between rankings of the three examiners.

3. The course work grades ranging from A to G and examinations marks of 8 candidates are given below.

Course work grade	A	B	C	D	E	F	G	H
Exam mark	92	75	63	54	48	45	34	18

(a) Calculate the spearman's rank correlation coefficient for the two sets of data

(b) Test whether it is significant at 5% level.