بسم الله الرحمن الرحيم

Correlation
and
linear regression

Pearson’s Correlation Coefficient (r)

- $$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n \sum Y^2 - (\sum Y)^2]}}$$

Correlation is defined
as the degree or strength of relationship between two characteristics in a population

- **The aim is**
- to **investigate the linear association between two**
- **continuous quantitative variables**.

Correlation therefore measures the closeness of the association.

For the correlation to be obtained we need the followings;

a. **One population**

b. **two characteristics**

c. **both should be continuous type (quantitative data)**

d. **both should be changing (variables) (not constant)**

e. **There must be some sort of relationship between two**

- ✓ **in order to obtain the strength of this relationship**

- ❑ **After that we need to determine**
- ❖ **which of the two** variables **is X** and
- ❖ which one **is Y according to the following;**

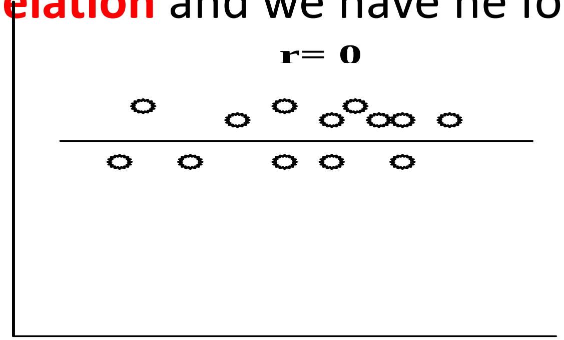
After that we need to determine which of the two variables is X and **which one is Y** according to the following;

X Independent:	Y Dependent:
The change in X is independent on the change in Y	**The change in Y is dependent on the change in X**
Less changing in a short period of time (more constant)	**More changing in a short period of time (more changing)**
As the cause	**As the effect**

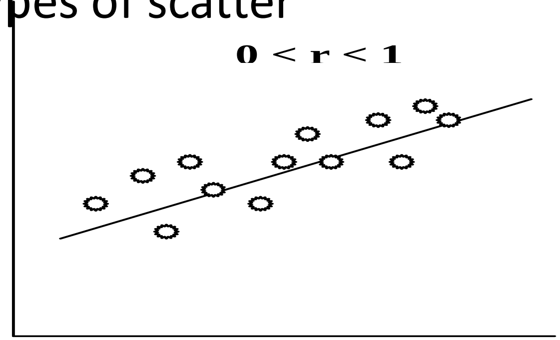
- **After that we need to draw a scatter diagram in** order to **ascertain** the **presence of correlation** and

we have the **following types of scatter**

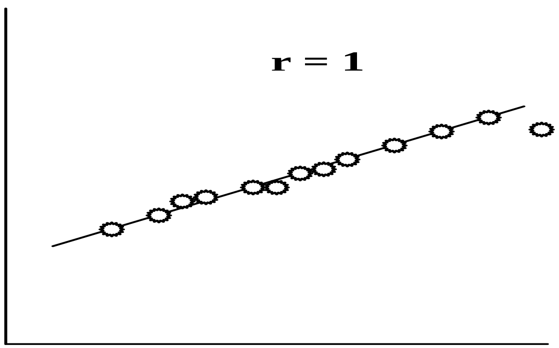
After that we need to **draw a scatter diagram** **in** order to **ascertain** the **presence of correlation** and we have he following types of scatter



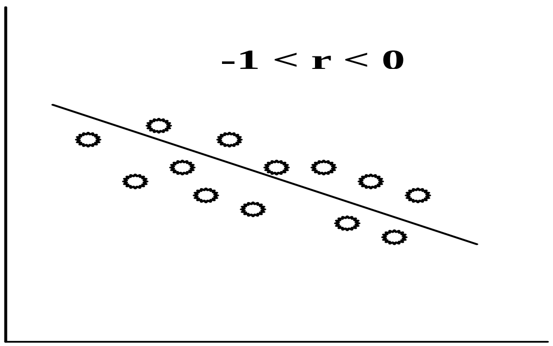
a) No correlation



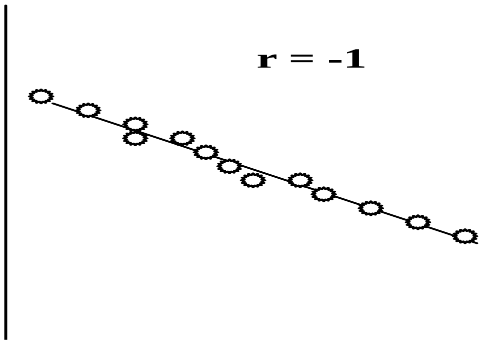
b) Direct positive correlation



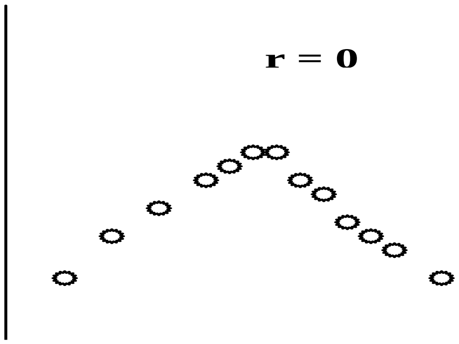
c) Perfect direct positive correlation



d) Inverse negative correlation



e) Perfect inverse negative correlation



f) Strong but non-linear relationship exists

Pearson's Correlation Coefficient (r)

- $$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n \sum Y^2 - (\sum Y)^2]}}$$

r only **measures the linear relationship so we have to draw a scatter diagram** first **to identify non-linear relationship**

Linear regression;

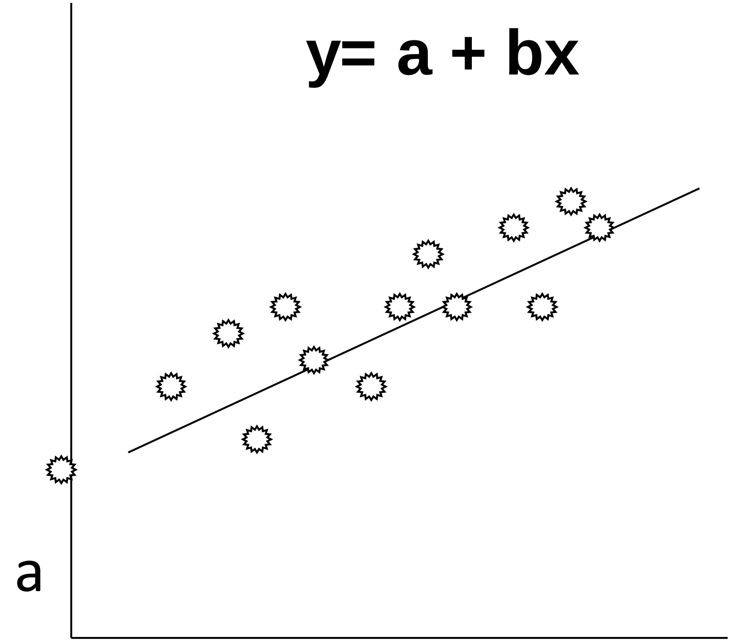
Gives the equation of straight line that best describes it and enables the prediction of one variable from the other.

a= constant, **called y-intercept**, **it is the place where the regression line intercept with y axis**

b=regression coefficient

x= any value of X variable

y= any value of Y variable



Interpretation of r:

- r **is always a number between -1 and +1**
- r is **positive if x and y** tend to **be high or low together, and the larger its value, the closer the association**
- r is **negative if high value of y tend to go with low values of x and vice versa**
- r **only measures the linear relationship so we have to draw a scatter diagram first to identify non-linear relationship.**

Pearson's Correlation Coefficient (r)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation

-
- **The r^2** (The coefficient of determination), i.e. when value of **$r=0.58$, then $r^2=0.34$**,

this means that 34% of the variation in the values of y may be accounted for by **knowing values of x or** vice versa

Pearson’s Correlation Coefficient (r)

- **The value of (r) indicates the strength of the relationship**
- **<0.2 : very weak**
- **0.2- <0.4 : weak**
- **0.4- <0.7 : moderate**
- **0.7- <0.9 : strong**
- **≥0.9 : very strong**

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation e.g; The body weight (Kg) and plasma volume (Liter) of 8 healthy men are presented in this table;

No	**Body weight (Kg)**		**Plasma volume (Liter)**		
	X	**X^2**	**Y**	**Y^2**	**X x Y**
1	**58**	**3364**	**2.75**	**7.56**	**159.50**
2	**70**	**4900**	**2.86**	**8.18**	**200.20**
3	**74**	**5476**	**3.37**	**11.36**	**249.38**
4	**63.5**	**4032.25**	**2.76**	**7.62**	**175.26**
5	**62**	**3844**	**2.62**	**6.86**	**162.44**
6	**70.5**	**4970.25**	**3.49**	**12.18**	**246.05**
7	**71**	**5041**	**3.05**	**9.30**	**216.55**
8	**66**	**4356**	**3.12**	**9.73**	**205.92**
	$\sum x=535$	$\sum x^2=35983.5$	$\sum y=24.02$	$\sum y^2=72.798$	$\sum x.y=1615.292$

Pearson's Correlation Coefficient (r)

- $$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2 \, . \, \sum (Y - \bar{Y})^2}}$$

$$r = \frac{SP\,xy}{\sqrt{(SSx \; . \; SSy)}} = \frac{SP\,xy}{\sqrt{(SQx \; . \; SQy)}}$$

SP = Sum of products of X and Y

SS=SQ= Sum of squares of X or of Y

$SPxy = \sum x.y - (\sum x).(\sum y) / n \Rightarrow 1615.292 - (535x24.02)/ 8 \Rightarrow 8.9545$

$SQx = \sum x^2 - (\sum x)^2 / n \quad ==> 35983.5 - (535)^2 / 8 \quad ==> 0.675$

$SQy = \sum y^2 - (\sum y)^2 / n \quad ==> 72.798 - (24.02)^2 / 8 \quad ==> 205.375$

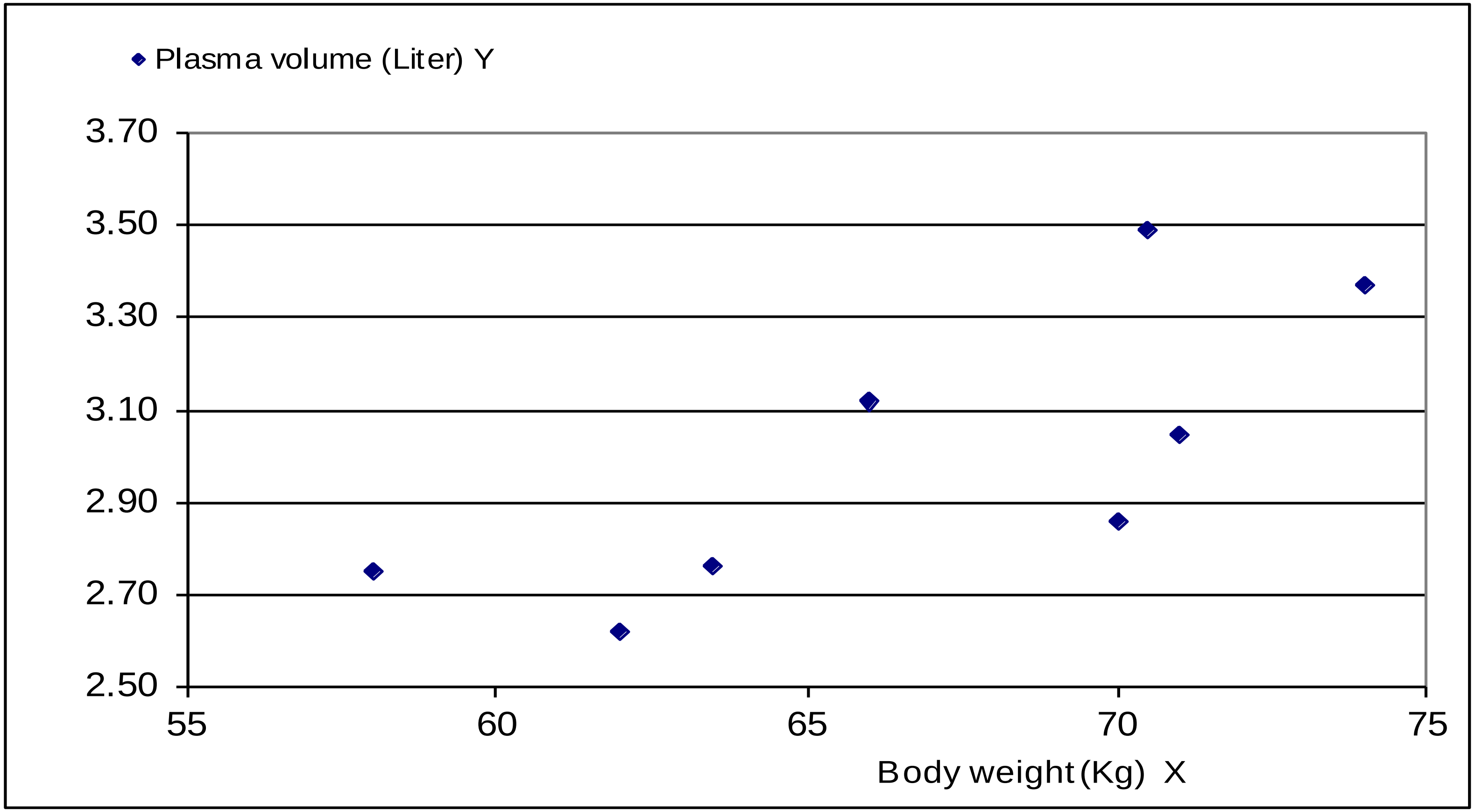
SP = Sum of products of X and Y

SS=SQ= Sum of squares of X or of Y

$$r = \frac{\sum x.y - (\sum x).(\sum y) / n}{\sqrt{[\sum x^2 - (\sum x)^2/ n] [\sum y^2 - (\sum y)^2/ n]}}$$

$r = 8.9545 / \sqrt{(0.675 \times 205.375)} \implies + 0.759$

There is strong direct relationship between weight (Kg) and plasma volume (Liter)



Scatter diagram of plasma volume and body weight showing linear regression line

Regression:

Regression:

Regression: it is the best fit for the relationship between two characteristics in a population.

Regression Line: The best line that fits the relationship between two characteristics in a population. Usually determined by the equation of straight line of the first degree which is→ y=a+bx

Y → any value on y axis, the dependent variable

a→constant, the intercept, the value of y when x equals to zero, it is the distance between the X axis and the point at which the regression line or its extension cuts the Y axis.

.b=regression coefficient= SPxy/SSX → SPxy/SQx

X → any value on X axis, the independant variable

We can get the equation according to the following

~~¥~~=a + b~~X~~

~~¥~~= ∑y/n = 24.02/8 → 3.0025

~~X~~=∑x/n= 535/8 → 66.875

b=SPxy/SQx = 8.9545/205.375→0.0436

~~¥~~=a + b~~X~~

3.0025=a + 0.0436 x 66.875

3.0025 = a + 2.916

a = 3.0025-2.916 → 0.0865

Y = 0.0865 + 0.0436 X

Regression coefficient “b” means that; for each one unit change in x axis there is about (b) unit change in y axis

For each one Kg increase in weight, there is about 0.0867 Liter increase in plasma volume

$Y = a + bX$

$3.0025 = a + 0.0436 \times 66.875$

$3.0025 = a + 2.916$

$a = 3.0025 - 2.916$ ⍰ 0.0865

$$Y = 0.0865 + 0.0436\ X$$

Thank you for attention

year 1ST
medical students