# Biostatistics Lecture 11

## Correlation and linear regression:

The aim is to investigate the linear relationship between two continuous variables. Correlation therefore measures the closeness of the relationship.

Correlation is defined as the degree or strength of relationship between two characteristics in a population. For the correlation to be obtained we need the followings;

- One population
- two characteristics
- both should be continuous type (quantitative data)
- both should be changing (variables) (not constant)
- There must be some sort of relationship between two in order to obtain the strength of this relationship

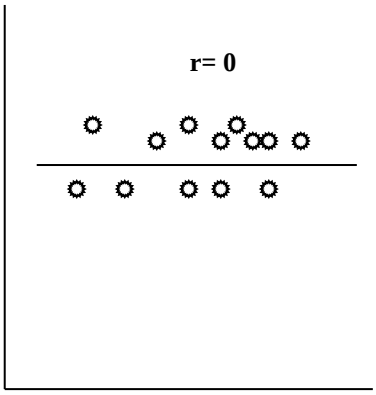
## Uses of correlation;

1- **Description**; to describe the general level of one variable that is related with each level of other variable.
2- **Adjustment**; to provide a mean of adjusting two set of variables when one variable tends to have different values with another variable.
3- **Forecasting** or **prediction**; to aid in forecasting or prediction a level of one variable for a new value of another variable.
4- **Interpreting**; to understand and interpret the mechanism by which one variable is related to another.
5- **Outlier detection**; to detect abnormal values or outliers that may merit detailed individual study.

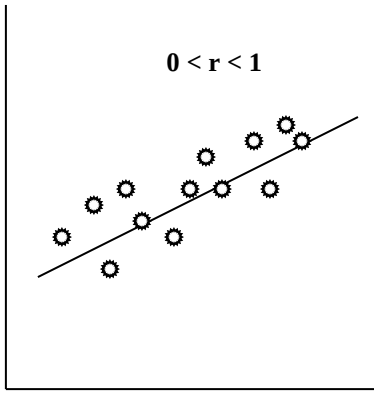
After that we need to determine which of the two variables is X and which one is Y according to the following;

| X | Y |
|---|---|
| Independent: The change in X is independent on the change in Y | Dependent: 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 |
| As explanatory | As out come |

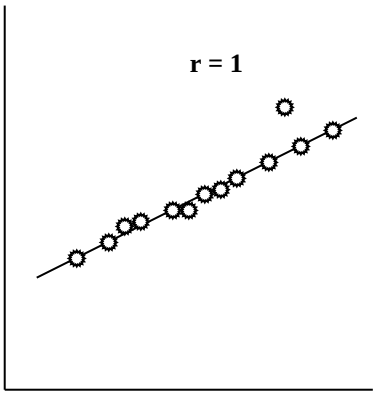
After that we need to draw a scatter diagram in order to ascertain the presence of correlation and we have the following types of scatter;



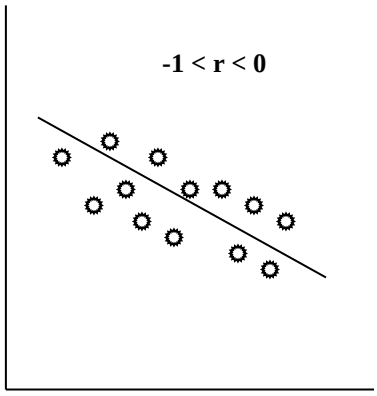

a) No correlation



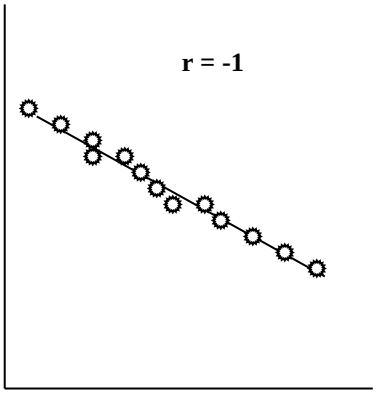

b) Direct positive correlation



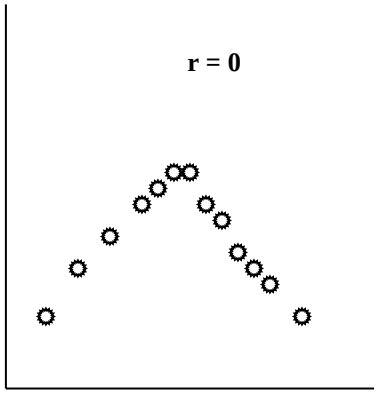

c) Perfect direct positive correlation




d) Inverse negative correlation




e) Perfect inverse negative correlation

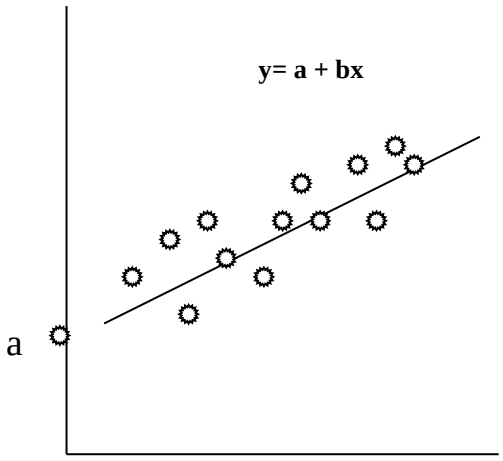



f) Strong but non-linear relationship exists

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 intersept 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 relationship
- r is negative if high value of y tends 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.
- if we calculate r without examining the data the we will miss a strong but non-linear relationship
- if r < 0.3 → No correlation,

  r 0.3-<0.5 → Weak correlation,

  r 0.5-0.7 → Moderate correlation,

  r 0.7- <1 (+ or -) → Strong 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

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 . 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 |
| | $\sum x=535$ | $\sum x^2=35983.5$ | $\sum y=24.02$ | $\sum y^2=72.798$ | $\sum x.y=1615.292$ |

In general, high plasma volume tends to be associated with high weight, this relationship is measured by the Pearson correlation,

$$r = \frac{\sum (X - X/)(Y - Y/)}{\sqrt{[\sum (X - X/)^2 \cdot \sum (Y - Y/)^2}}$$

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

SP = Sum of products of X and Y

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

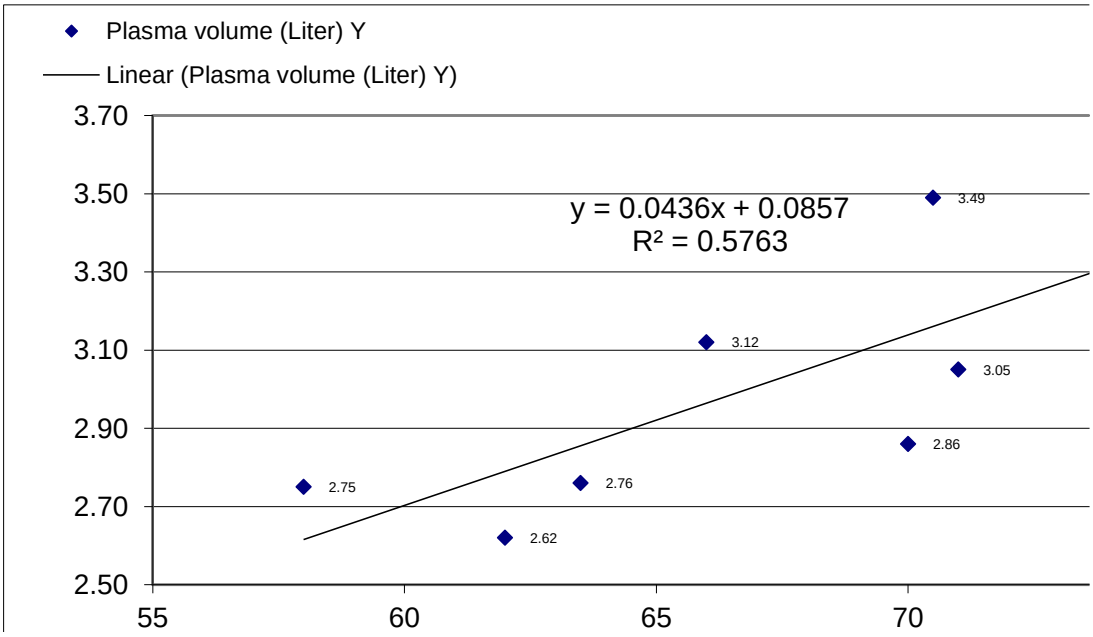
$SPxy = \sum x.y - (\sum x).(\sum y) / n$ ==> 1615.292 - (535x24.02)/ 8 ==> 8.9545

$SQx = \sum x^2 - (\sum x)^2 / n$ ==> $35983.5 - (535)^2 / 8$ ==> 205.375

$SQy = \sum y^2 - (\sum y)^2 / n$ ==> $72.798 - (24.02)^2 / 8$ ==> 0.678

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

$r = 8.9545 / \sqrt{(0.678 \times 205.375)}$ ==> +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

A weak correlation may therefore be statistically significant if based on a large number of observations, while a strong correlation may fail to achieve significance if there are only a few observations.