Lecture
Correlation and linear regression

Correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. Correlation is the strength of relationship between 2 characteristics in a population, to be obtained needs the following:

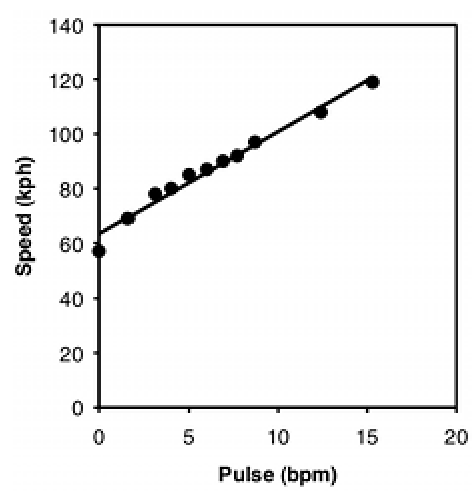
*** One population. Two characteristics**

*** Both should be changing (variables are not constant)**

*** There must be some so sort of relationship between 2 in order to obtain the strength of this relationship.**

Need to determine which of 2 variables is X & Y according to following introduction

X	**Y**
Independent: the changing 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**



Graph of pulse rate versus speed on an elliptical exercise machine

(perfect direct positive correlation) r = 1

Correlation analysis is used to quantify the association between two continuous variables (between an independent and a dependent variable or between two independent variables. Regression analysis is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable or **dependent variable** and the risk factors and confounders are the **predictors**, or **independent variables**. In regression analysis, the dependent variable is denoted "*y*" and the independent variables are denoted by "*x*".

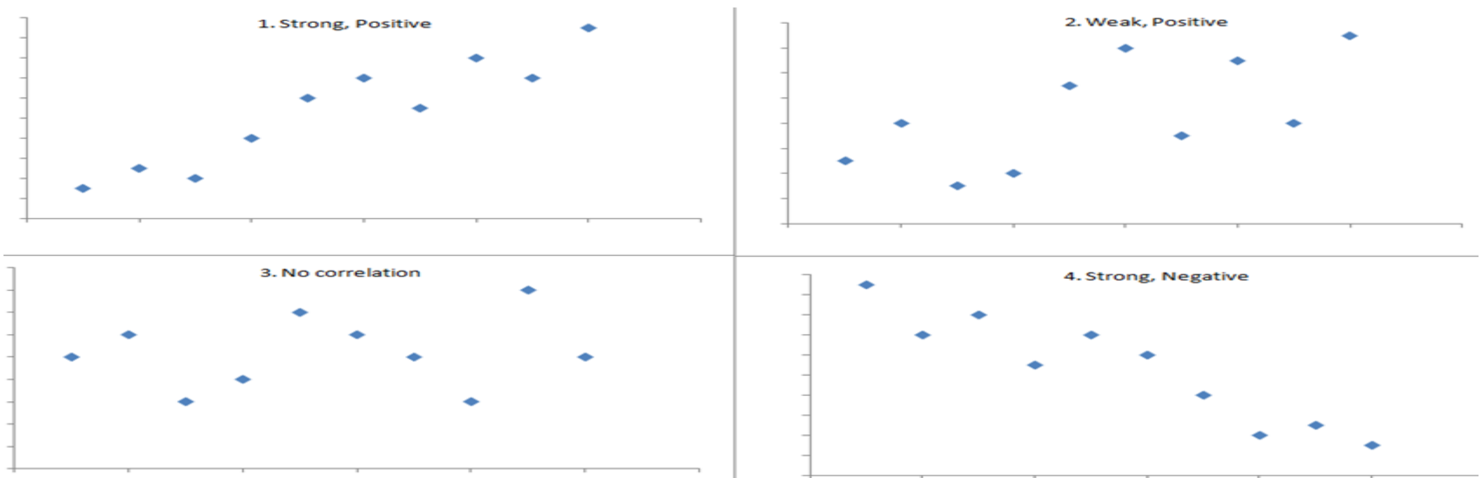
In correlation analysis, we estimate a **correlation coefficient**, more specifically the **Pearson Product Moment correlation coefficient**. The correlation coefficient, denoted **r** Ranges between **-1 and +1** and **quantifies the direction and strength of the linear association** between the two variables. The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other or negative (i.e., higher levels of one variable are associated with lower levels of the other. The sign of the correlation coefficient indicates the direction of the association. **The magnitude of the correlation coefficient indicates the strength of the association.**

For example, a correlation of r = **0.9** suggests a strong positive association between two variables.

Where as a correlation of **r = - 0.2** suggest s a weak, negative association.

A correlation **close to zero suggests no linear association** between two continuous variables.

If r (< 0.3) no----- correlation,
If r (0.3----< 0.5) ---- weak correlation,
If r (0.5 0.7) ---- moderate correlation,
If r (0.7 - 1) strong ve+ correlation,
If r = - 1 strong ve- correlation



a strong positive association (r=0.9), similar to what we might see for the correlation between infant birth weight and birth length.

Types

There are several different measures for the degree of correlation in data, depending on the kind of data, principally whether the data is a measurement, ordinal, or categorical.

Pearson:

The Pearson product-moment correlation coefficient, also known as r, R, or Pearson's r, is a measure of the strength and direction of the linear relationship between two variables .

Intra-class:

A descriptive statistic that can be used to measure the reliability, when quantitative measurements are made on units that are organized into groups; it describes how strongly units in the same group resemble each other, unlike most other correlation measures it operates on data structured as groups. strong correlation is ≥ 0.7.

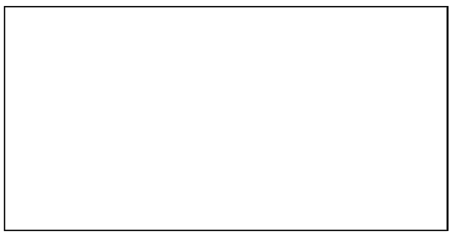
Rank:

Rank correlation is a measure of the relationship between the rankings of two variables, or two rankings of the same variable:

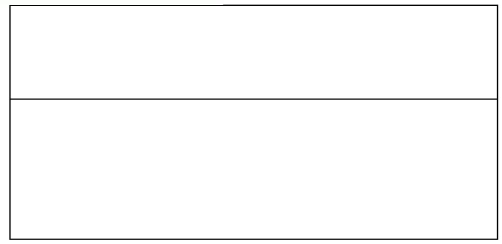
Spearman's rank correlation coefficient is a measure of how well the relationship between two variables can be described by a monotonic function, no need for linear relationship between two variables .

Goodman and Kruskal's gamma is a measure of the strength of association of the cross tabulated data when both variables are measured at the ordinal level.

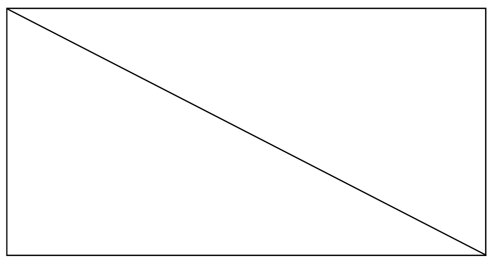
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$



Strong positive correlation



No correlation



Strong negative correlation

Example: the body weight and plasma volume of 8 healthy men are presented in this table: in general high plasma volume tends to be associated with high wt this relationship is measured by Pearson correlation:

No	Body wt 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	2.76	7.62	175.26
5	62	3844	2.62	6.86	162.44
6	70.5	497.25	3.49	12.18	246.05
7	71	541	3.05	9.30	216.55
8	66	4356	3.12	9.73	205.92
	$\Sigma x = 535$	$\Sigma x^2 = 35983.5$	$\Sigma y = 23.95$	$\Sigma y^2 = 72.79$	$\Sigma x.y$ 1615.292

R=8(1615.292)-(535)(23.95)/√ (8X35983.5-286.22)-(8X72.79-573.60

= +0.759, there is a strong relationship between body weight & plasma volume

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Example question: Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	9	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
	247	486	20485	11409	40022

Use the following correlation coefficient formula.

$r=n(\Sigma xy)- \Sigma(x)(y)/ \sqrt{[n\Sigma x^2]- (\Sigma x)^2 [n\Sigma y^2 -(\Sigma y)^2]}$

From our table: $\Sigma x = 247$, $\Sigma y = 486$
$\Sigma xy = 20485$, $\Sigma x^2 = 11409$, $\Sigma y^2 = 40022$
$n = 6$
The correlation coefficient = $6((20485) – (247 \times 486))/\sqrt{[[6(11409) – (247^2)] \times [6(40022) – 486^2]]}$
= 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.00005298, which means the variables have a NO correlation.