Correlation Analysis

Dr Kumar Chinta
Professor
Department Of Community Medicine
Narayana Medical College
Nellore

Correlation Analysis

Contents

- Introduction
- Quantitative variable
- Correlation
- Dependent and independent variable
- Methods of correlation analysis
- Types of correlation
- Working on Excel
- Summary

Correlation Analysis

Introduction

- Correlation Analysis Is a method of bivariate statistical analysis
- Used to study the relationship between two quantitative variables
- Studies the direction and magnitude of correlation

Correlation Analysis

Quantitative Variable?

- Variables are measured on quantitative scale or qualitative scale

S .No	Sex	Weight in kgs
1	M	50
2	F	55
3	F	60
4	M	50

Correlation Analysis

What is correlation?

Distribution of children by age and weight

Age	Weight
Birth	2.6
3 mts	5.3
6 mts	6.7
9 mts	7.4
1yr	8.4
2 yrs	10.1
3 yrs	11.8
4 yrs	13.5
S yrs	14.8
6 yrs	16.3
7 yrs	18
8 yrs	19.7
9 yrs	21.5
10 yrs	23.5

Correlation Analysis

What is correlation?

Distribution of children by age and weight

Age	Weight
Birth	2.6
3 mts	5.3
6 mts	6.7
9 mts	7.4
1yr	8.4
2 yrs	10.1
3 yrs	11.8
4 yrs	13.5
5 yrs	14.8
6 yrs	16.3
7 yrs	18
8 yrs	19.7
9 yrs	21.5
10 yrs	23.5

This phenomenon of change in magnitude of one variable with a change in the magnitude of the other associated variable is called **correlation**

Correlation Analysis

Dependent & **independent** variable ?

Distribution of children by age and weight

Age	Weight
Birth	2.6
3 mts	5.3
6 mts	6.7
9 mts	7.4
1yr	8.4
2 yrs	10.1
3 yrs	11.8
4 yrs	13.5
5 yrs	14.8
6 yrs	16.3
7 yrs	18
8 yrs	19.7
9 yrs	21.5
10 yrs	23.5

In correlation out of two associated variables one is the **dependent** variable while the other is **independent** variable

Correlation Analysis

In medical research and epidemiology the independent variable is also called **predictor variable /exposure variable**, while the dependent variable is known as **outcome/effect variable**

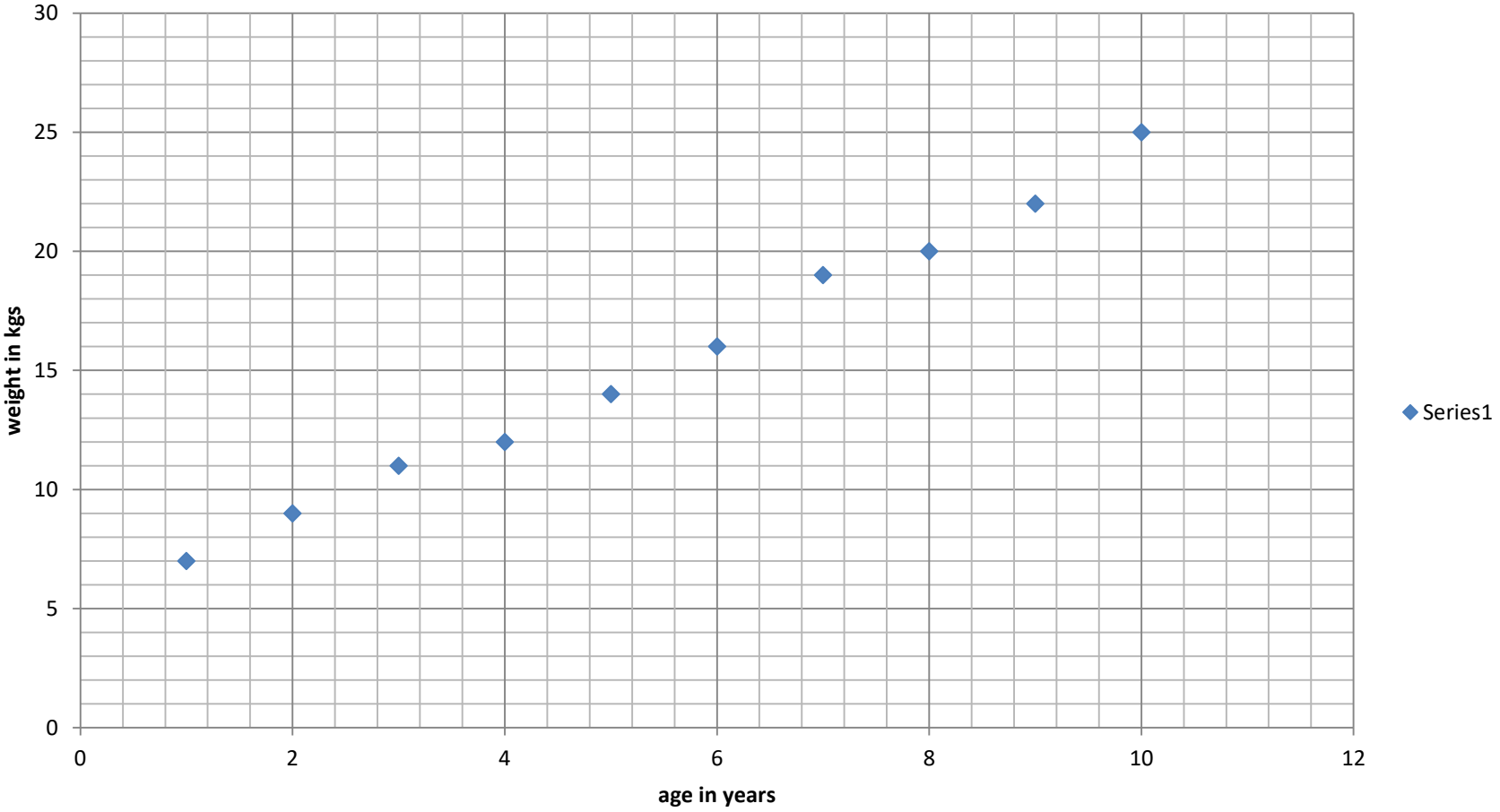
Correlation Analysis

Methods

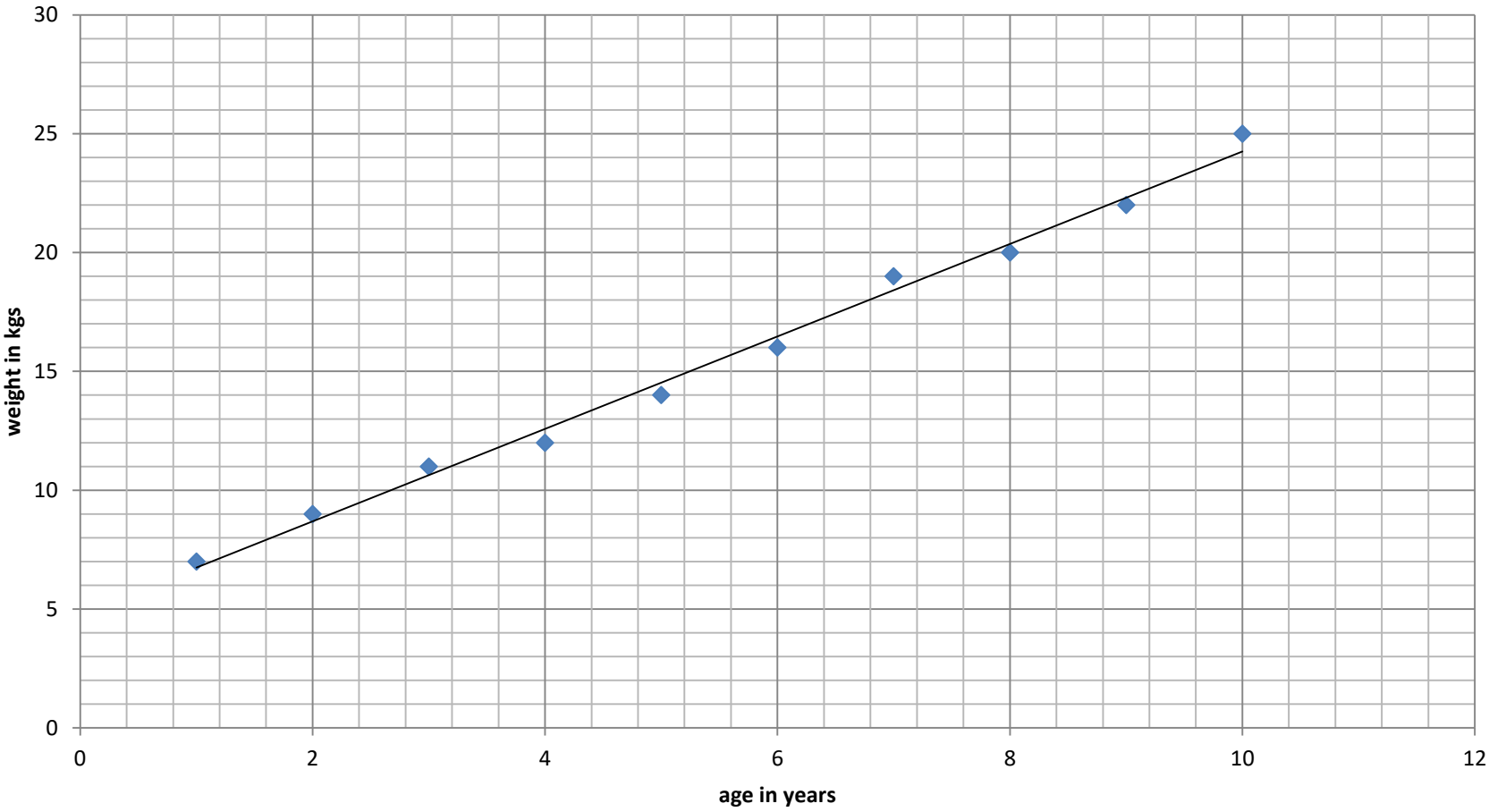
1) Scatter Plot

2) Correlation Coefficient

Correlation Analysis
Scatter plot



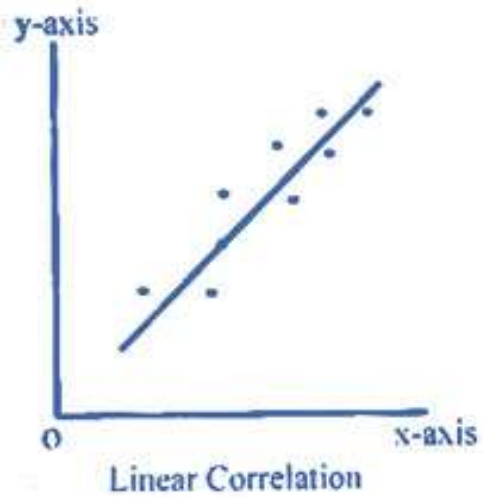
Correlation Analysis



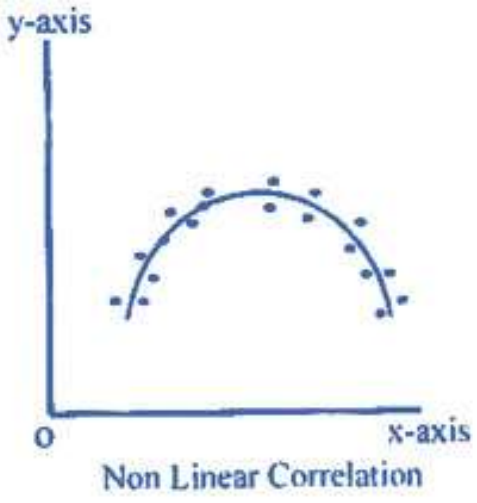
Correlation Analysis

Types of correlation

- Linear correlation
- Non-linear correlation



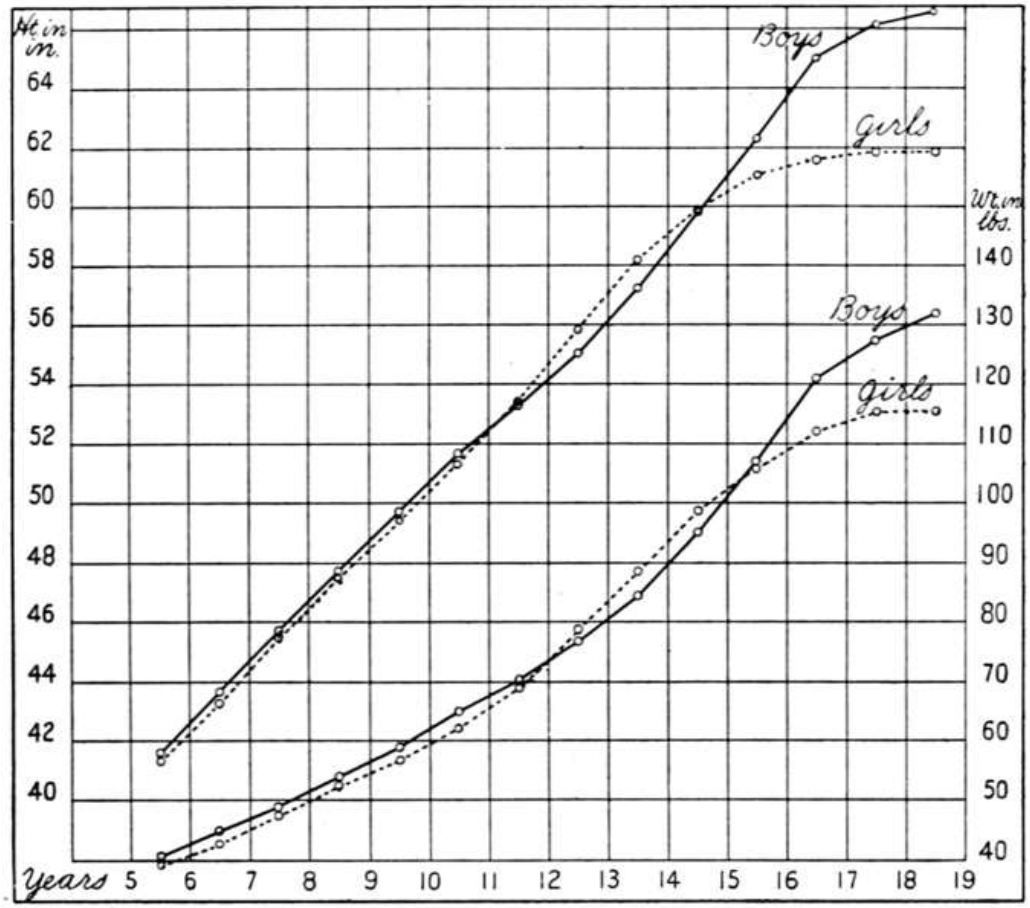
Linear Correlation



Non Linear Correlation

Correlation Analysis
Types of correlation

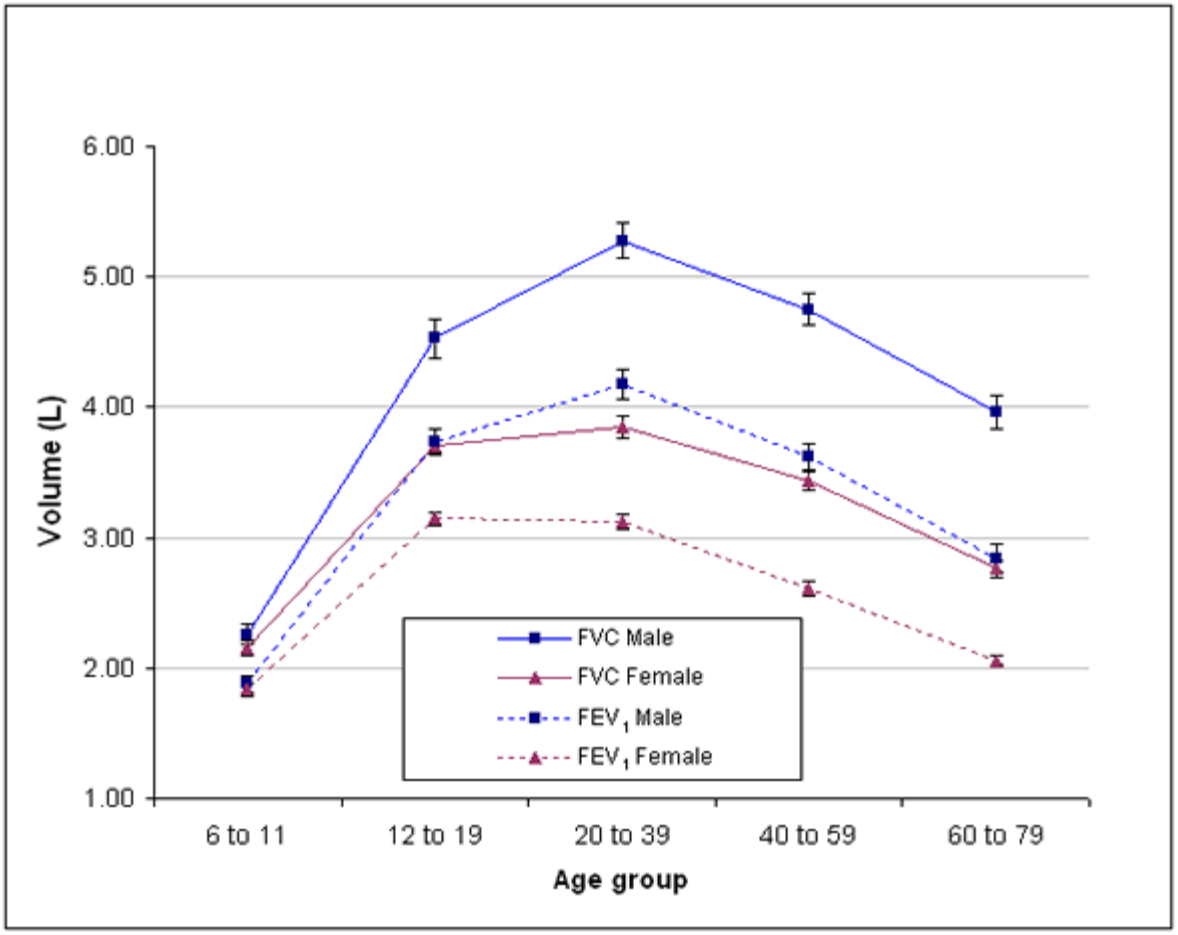
- Graph of the growth of Boston schoolchildren in height and weight



Correlation Analysis

Types of correlation

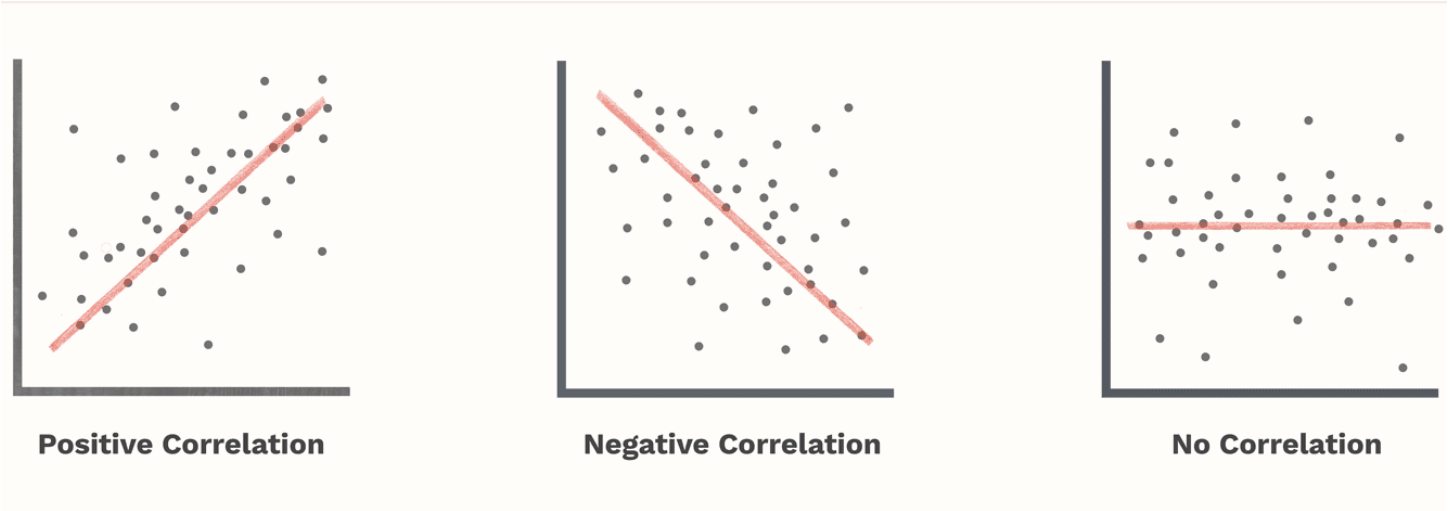
Mean FVC and mean FEV_1 by age group and sex



Correlation Analysis

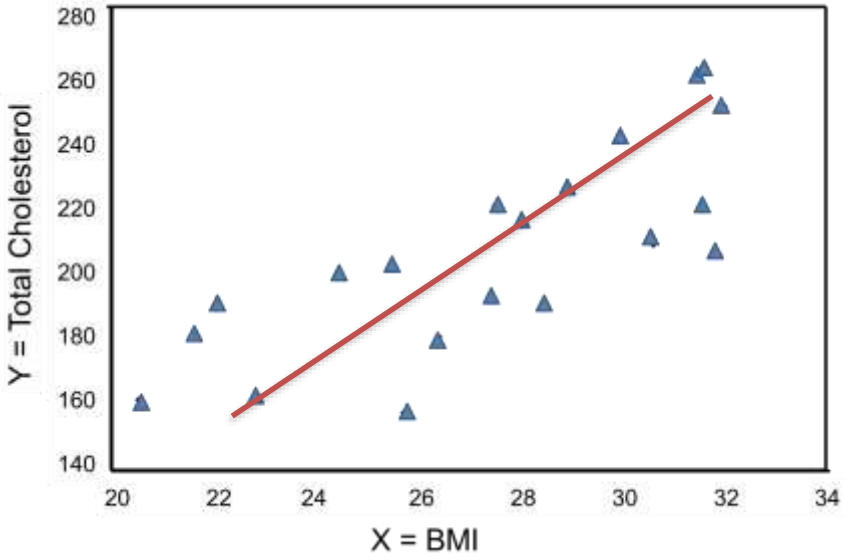
Types of correlation

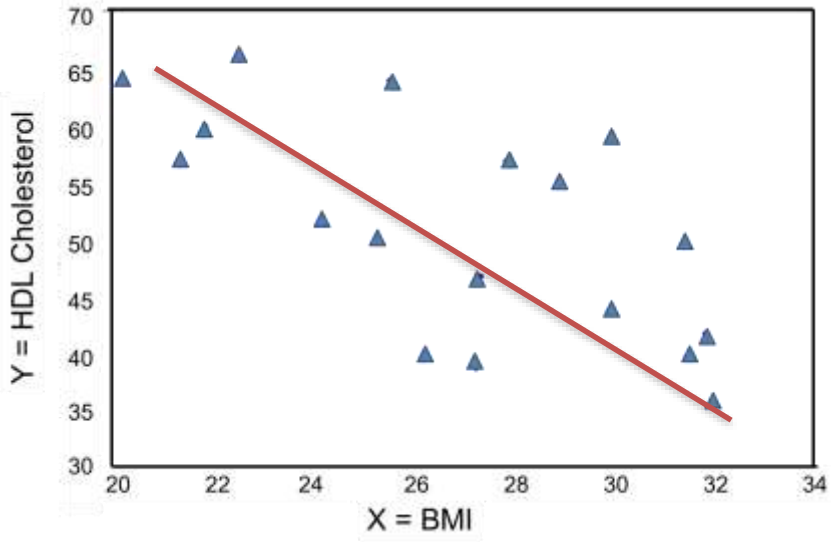
- Positive Correlation x↑ y↑
- Negative Correlation x↑ y↓
- No Correlation



Correlation Analysis

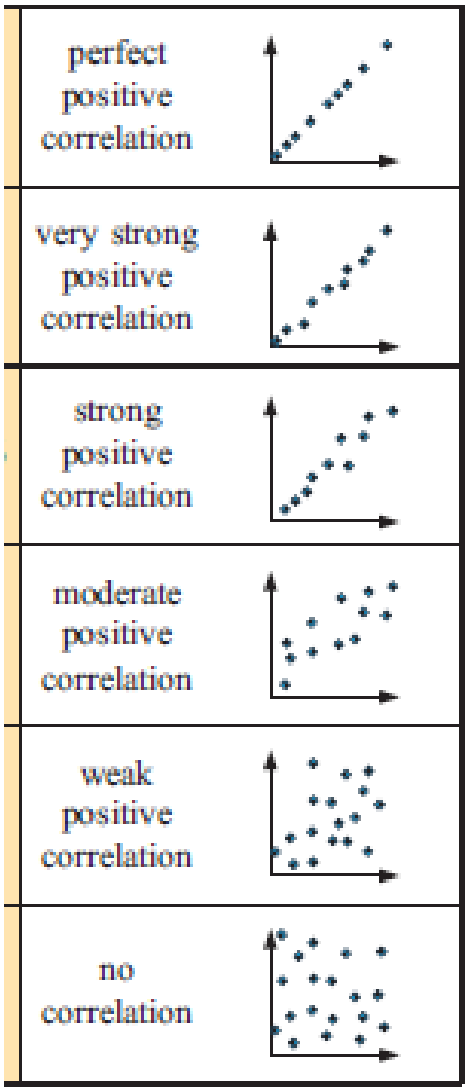
Types of correlation

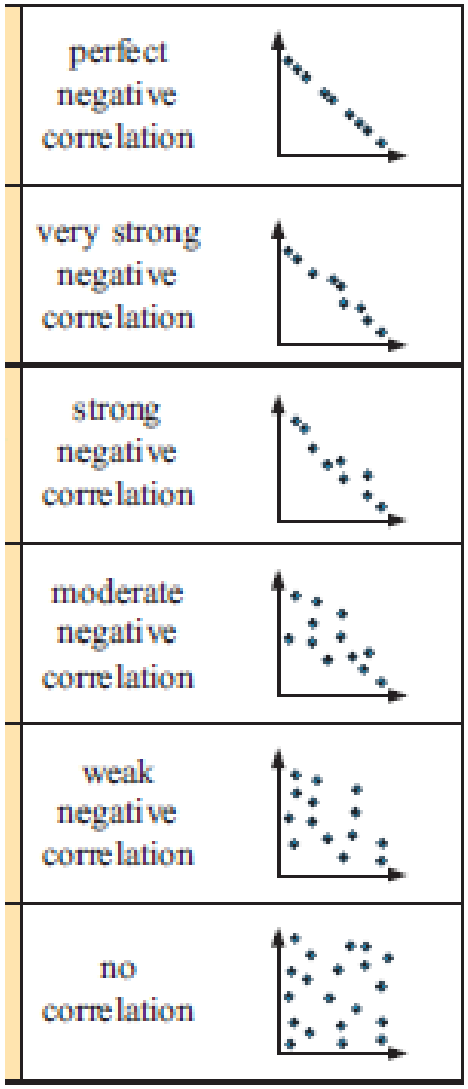




Correlation Analysis

Types of correlation



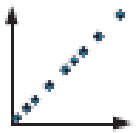
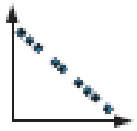
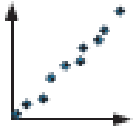
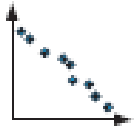
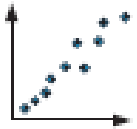
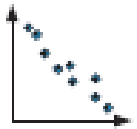
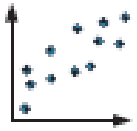
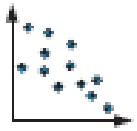
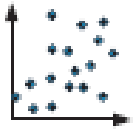
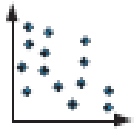
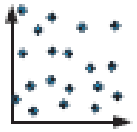
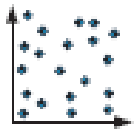


Correlation Analysis

Correlation coefficient

- It is a constant calculated using mathematical formula to quantify the relationship
- It explains the degree and direction of correlation
- The most commonly used correlation coefficient is Karl Pearson's correlation coefficient
- It is denoted by letter 'r'
- Its value ranges from +1 through 0 to -1
- The other correlation coefficients are
- Spearman's rank cc
- Kendall tau rank cc
- Cramer's correlation

Correlation Analysis

Positive correlation		Negative correlation	
$r = 1$	perfect positive correlation 	$r = -1$	perfect negative correlation 
$0.95 \leqslant r < 1$	very strong positive correlation 	$-1 < r \leqslant -0.95$	very strong negative correlation 
$0.87 \leqslant r < 0.95$	strong positive correlation 	$-0.95 < r \leqslant -0.87$	strong negative correlation 
$0.5 \leqslant r < 0.87$	moderate positive correlation 	$-0.87 < r \leqslant -0.5$	moderate negative correlation 
$0.1 \leqslant r < 0.5$	weak positive correlation 	$-0.5 < r \leqslant -0.1$	weak negative correlation 
$0 \leqslant r < 0.1$	no correlation 	$-0.1 < r \leqslant 0$	no correlation 

Pearson correlation coefficient

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

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

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

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

Step 1: *Make a chart.* Use the given data, and add three more columns: xy, x^2, and y^2.

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

Step 2: *Multiply x and y together to fill the xy column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

Step 3: *Take the square of the numbers in the x column, and put the result in the x^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

Step 4: *Take the square of the numbers in the y column, and put the result in the y^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	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

Step 5: *Add up all of the numbers in the columns and put the result at the bottom of the column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	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
Σ	247	486	20485	11409	40022

- **Step 6:** *Use the following correlation coefficient formula.*

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

- The answer is: **2868 / 5413.27 = 0.529809**

Correlation Analysis

Significance test

- To test whether the association is merely apparent, and might have arisen by chance use the *t* test

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- DF = n-2
- N = number of pairs
- If the calculated **t** value is more than table **t** value the correlation is significant

Correlation Analysis

- Calculated t- value is 1.2493
- DF = 4
- Table t-value is 2.766 at 0.05 probability
- The calculated t- value is less than table t- value at 0.05 probability and 4 degrees of freedom the correlation is statistically not significant

Correlation Analysis

Drawing Scatter Plot and calculating Correlation Coefficient in MS-Excel

Correlation Analysis
Summary

- Correlation analysis is a method of bivariate statistical analysis
- Used to study the relationship between two quantitative variables
- Studies the direction and magnitude of correlation
- Correlation analysis is done by scatter plot and correlation coefficient
- CC calculates precisely the degree of correlation

Correlation Analysis

Take home message

Correlation analysis is used to study the relationship between two quantitative variables and this can be done using MS-Excel software

Thank you