Measurement of Dispersions: Scatter Diagram

What is Scatter Diagram?

A scatter diagram or *scatter graphs* or *scatter plot* is a chart type that is used to observe and visually display the relationship between variables. The values of the variations are represented by dots or points. A scatter diagram or scatter plot is a collection of points using a Cartesian coordinates to display the values of two variables for a set of data. The value of one variable determine the position on the horizontal axis and the value of the other variable determine the position on the vertical axis.

When to Use a Scatter Diagram:

1. When you have paired numerical data.
2. When you try to determine whether the two variables are related , such as:
    - when determining whether two effects that appear to be related both occur with the same cause;
    - when trying to identify potential root causes of problems.
3. Use a scatter diagram or scatter plot to determine whether or not two variables have a relationship or correlation.

Scatter Diagram Application and Uses

1. Demonstration of the relationship between two variables

The most common use of the scatter plot is to display the relationship between two variables and observe the nature of the relationship. The relationships observed can either be positive or negative, non-linear or linear, and/or, strong or weak. The data points or dots, which appear on a scatter plot, represent the individual values of each of those data points and also allow pattern identification when looking at the data holistically.

2. Identification of correlational relationships

Scatter plot will enable the identification of correlational relationships. Scatter plots tend to have *independent variables* on the horizontal axis and *dependent variables* on the vertical axis. It allows the observer to know or get an idea of what the possible vertical value may be, provided there is information on the horizontal value.

3. Identification of data patterns

Data pattern identification is also possible with scatter plots. Data points can be grouped together based on how close their values are, and this also makes it easy to identify any outlier points when there are data gaps.

Seeing as scatter plots aid in the identification of correlations between variables, the nature of the correlations can also be estimated based on a specific confidence level.

- Positive correlation depicts a rise, and it is seen on the diagram as data points slope upwards from the lower-left corner of the chart towards the upper-right.
- Negative correlation depicts a fall, and this is seen on the chart as data points slope downwards from the upper-left corner of the chart towards the lower-right.
- Data that is neither positively nor negatively correlated is considered uncorrelated (null).

Scatter Diagram Consideration

- Even if the scatter diagram shows a relationship, do not assume that one variable caused the other. Both may be influenced by a third variable.
- When the data are plotted, the more the diagram resembles a straight line, the stronger the relationship.
- If a line is not clear, statistics determine whether there is reasonable certainty that a relationship exists. If the statistics say that no relationship exists, the pattern could have occurred by random chance.
- If the scatter diagram shows no relationship between the variables, consider whether the data might be stratified.
- If the diagram shows no relationship, consider whether the independent (x-axis) variable has been varied widely. Sometimes a relationship is not apparent because the data do not cover a wide enough range.

Scatter Diagram Procedure

1. Collect pairs of data where a relationship is suspected.
2. Draw a graph with the *independent variable* on the horizontal axis and the *dependent variable* on the vertical axis. For each pair of data, put a dot or a symbol where the x-axis value intersects the y-axis value. (If two dots fall together, put them side by side, touching, so that you can see both.)
3. Look at the pattern of points to see if a relationship is obvious.

What is an independent variable?

A variable is something you are trying to measure. It can be practically anything, such as objects, amounts of time, feelings, events, or ideas.

In research, **variables** are any characteristics that can take on different values, such as height, age, species, or exam score. In scientific research, we often want to study the effect of one variable on another one. For example, you might want to test whether students who spend more time studying get better exam scores.

Variables are divided into two types:

- Independent variables
- Dependent variables

The variables in a study of a cause-and-effect relationship are called the **independent and dependent variables**.

- The **independent variable** is the **cause**. Its value is *independent* of other variables in your study.
- The **dependent variable** is the **effect**. Its value *depends* on changes in the independent variable.

An easy way to think of independent and dependent variables is, when you are conducting an experiment, the **independent variable** is **what *you* change**, and the **dependent variable** is **what changes because of that**.

The independent variable is the **cause** and the dependent variable is the **effect**.

Independent Variable and Dependent Variable in Business

The prices of raw materials, labor wage rates and facility rental rates are **independent** expense variables. The prices of raw materials, such as food commodities, metals and minerals, do not change, regardless of how much a small business spends on them. Labor wage rates and facility rental rates are other examples of independent expense variables. They affect the cost structure of a small business, but the owner cannot change market wage rates or rental rates by himself.

In the business context, **profit** is a **dependent variable** because it depends on the economy, sales and expenses. Product quality depends on the manufacturing and design processes. The number of employees laid off during a recession depends partly on declining business revenues. Government tax revenue depends on customer income, business profits, capital gains and other variables.

Independent variables that affect sales include customer demographics, store location and weather. Customer demographics include age, occupation, family status, income level and gender. These factors affect what a customer needs, which affects sales and ultimately profits. A store located in a densely populated metropolitan area may have higher sales than a store in a sparsely populated rural area. Similarly, customers may go shopping when the weather is pleasant, but few would venture outside in stormy or snowy weather. Some variables have a circular relationship with sales. For example, sales depend on advertising, but the level of advertising expenses also depends on sales.

	Independent variable	Dependent variable
1	Number of stores	Total of employees
2	Number of employees per store	Total revenue
3	Average salary per employee	Salaries
4	Average rent per store	Total expenses

Example 1

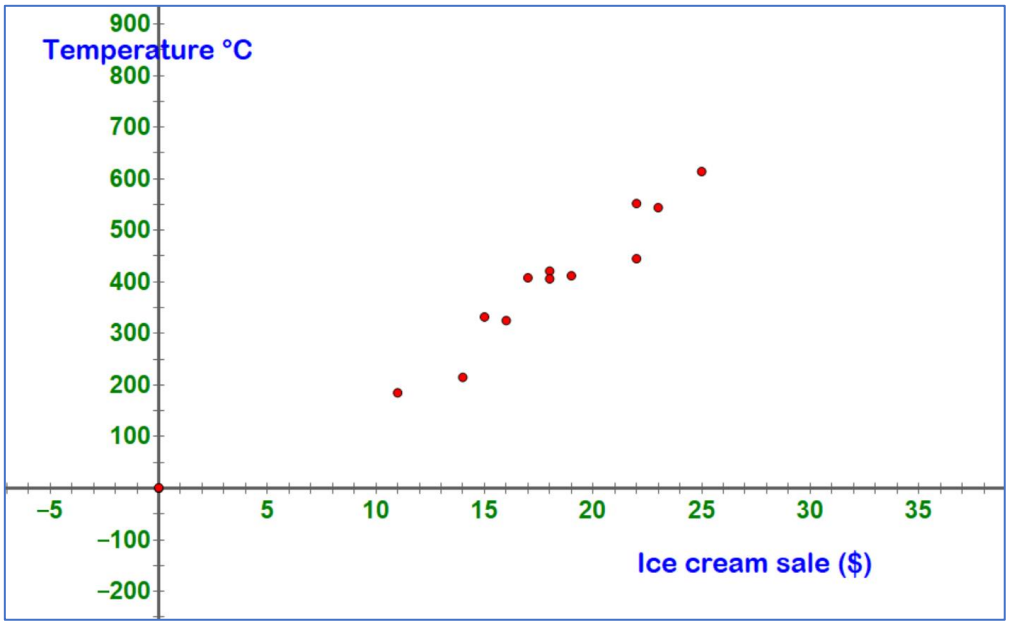
Ice Cream Sales

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days. Draw the scatter plot to show the Ice Cream Sales of the local shop.

Temperature °C Vs Ice Cream Sales	
Temperature °C	**Ice Cream Sales**
14	215
16	325
11	185
15	332
18	406
22	522
19	412
25	614
23	544
18	421
22	445
17	408

Solution

A Scatter Diagram of the Ice Cream Sales of the local shop



Correlation Coefficient (r) or Pearson Product Moment Correlation Coefficient

Correlation coefficients measure the strength of and the direction of a linear relationship between two variables. The linear correlation coefficient called the Pearson product-moment correlation coefficient, measures the strength of the linear association between two variables.

Correlation

When the two sets of data are strongly linked together we say they have a High Correlation. The word Correlation is made of

CO - meaning " *together* " and " *relation* "

- Correlation is Positive when the values **increase together**, and
- Correlation is Negative when **one value increase**, and the **other decreases**.

Symbol of Correlation Coefficient:

Topics	Population (Parameter)	Sample (Statistic)
Correlation Coefficient	ρ (rho) or R	r
Size	N	n

Correlation Coefficient (r) or Pearson product moment correlation coefficient

Sample Correlation Coefficient or Sample Pearson Correlation Coefficient (r_{xy}) formula:

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

$$r_{xy} \text{ or } r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\ \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Positive correlation:

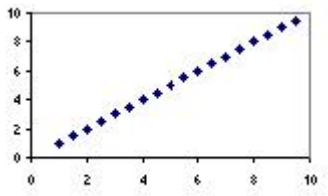
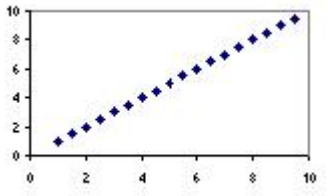
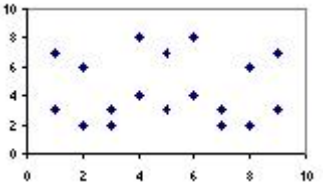
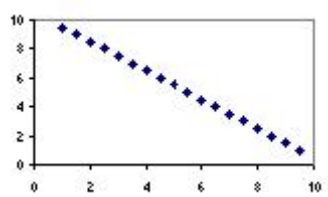
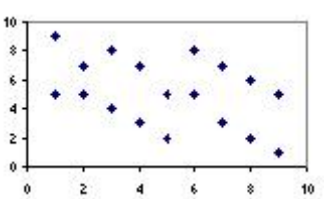
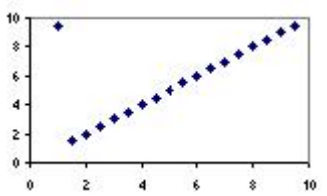
If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit.

Negative correlation:

If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit.

No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

The scatterplots below show how different patterns of data produce different degrees of correlation.

 Maximum positive correlation r = 1	 Strong positive correlation r = 0.80	 Zero Correlation r = 0
 Maximum negative correlation r = -1	 Moderate negative correlation r = -0.43	 Strong correlation and outlier r = 0.71

Example 2

The data in the table below display the income of eight sample families and their expenses.

Income (10,000 Baht)	1	3	4	6	8	9	11	14
Expenses (10,000 Baht)	1	2	4	4	5	7	8	9

Find a) A scatter diagram to determine the relationship between the income and their expenses;

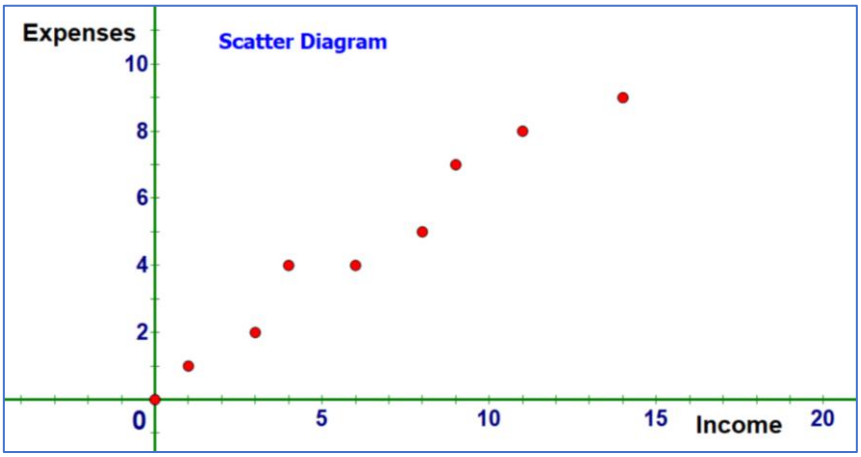
b) Correlation Coefficient (r) of the sample family.

Solution

a) A scatter diagram to determine the relationship between the income and their expenses

Let independent variable (x) = income

dependent variable (y) = expenses



b) Correlation Coefficient (r) of the sample family.

$$r_{xy} \text{ or } r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\ \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Family	Income (x) (10,000 Baht)	Expenses (y) (10,000 Baht)	xy	x^2	y^2
1	1	1	1	1	1
2	3	2	6	9	4
3	4	4	16	16	16
4	6	4	24	36	16
5	8	5	40	64	25
6	9	7	63	81	49
7	11	8	88	121	64
8	14	9	126	196	81
n = 8	$\sum x = 56$	$\sum y = 40$	$\sum xy = 364$	$\sum x^2 = 524$	$\sum y^2 = 256$

From table we have

n = 8	$\sum x = 56$	$\sum y = 40$
$\sum xy = 364$	$\sum x^2 = 524$	$\sum y^2 = 256$

$$r_{xy} \text{ or } r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r_{xy} = \frac{8(364) - (56)(40)}{\sqrt{8(524) - (56)^2}\sqrt{8(256) - (40)^2}}$$

$$= \frac{2912 - 2240}{\sqrt{4192 - 3136}\sqrt{2048 - 1600}}$$

$$= \frac{672}{\sqrt{1056}\sqrt{448}}$$

$$= \frac{672}{(32.49)(21.16)}$$

$$= \frac{672}{687.4884}$$

$$r_{xy} = 0.977$$

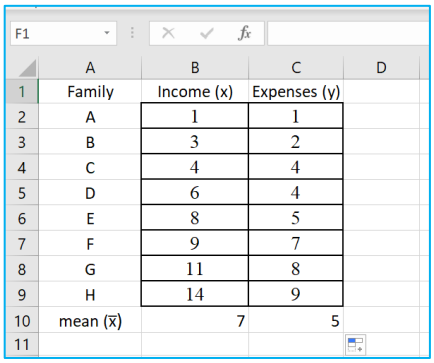
The two variables income and expenses of the sample correlation coefficient of the sample family has a strong positive correlation with $r_{xy} = 0.977$

Correlation Coefficient (r) or Pearson Product Moment Correlation Coefficient using Excel.

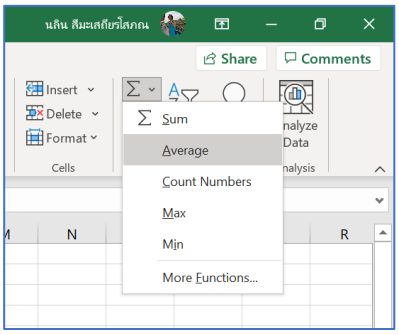
Correlation Coefficient (r) Method 1:

Step 1 Enter all data in Excel software program as shown below

Step 2: Find the mean by using the AVERAGE function: **=AVERAGE(B2:B9)**



	A	B	C	D
1	Family	Income (x)	Expenses (y)	
2	A	1	1	
3	B	3	2	
4	C	4	4	
5	D	6	4	
6	E	8	5	
7	F	9	7	
8	G	11	8	
9	H	14	9	
10	mean ($\bar{x}$)	7	5	
11				



The average (mean) $\bar{x}$ = 7 goes to any empty cell, say B10.

Using the same process to find the average of y ($\bar{y}$) , you shall get $\bar{y}$ = 5.

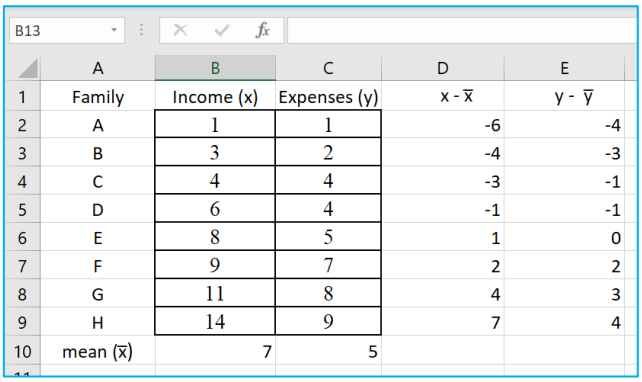
Using the following formula to find value of correlation coefficient (r_{xy})

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Step 3: Find the values of $x - \bar{x}$ and put the results in Col D:

Subtract the mean (average) from each number of the variable *x* (income) in Col B:

- move cursor to column D2
- Type: = **B2-B10** (mean value is in col B10, we will lock as a constant value)
- Click Enter. (You shall see the value of $x - \bar{x}$ = -6 in column D2)
- move cursor to the corner of column D2 and drag until col D9.
- The differences of $x - \bar{x}$ go to column D, beginning in D2.



	A	B	C	D	E
1	Family	Income (x)	Expenses (y)	x - x̄	y - ȳ
2	A	1	1	-6	-4
3	B	3	2	-4	-3
4	C	4	4	-3	-1
5	D	6	4	-1	-1
6	E	8	5	1	0
7	F	9	7	2	2
8	G	11	8	4	3
9	H	14	9	7	4
10	mean (x̄)	7	5		

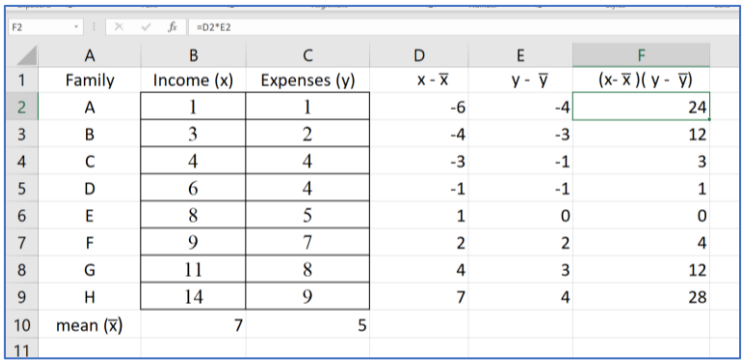
Step 4: Find the values of $y - \bar{y}$ and put the results in Col E:

Subtract the mean (average) from each number of the variable y (expenses) in Col C:

- move cursor to column E2
- Type: = **C2-C10** (mean value is in col C 10, we will lock as a constant value)
- Click Enter. (You shall see the value of $y - \bar{y}$ = -4 in column E2)
- move cursor to the corner of column E2 and drag until col E9.
- The differences of $y - \bar{y}$ go to column E, beginning in E2.

Step 5: Find the Find the values of $(x - \bar{x})(y - \bar{y})$ and put the results in Col F:

- move cursor to column F2
- Type: = **D2*E2** and Click Enter.
- move cursor to the corner of column F2 and drag until col F9.
- The value of $(x - \bar{x})(y - \bar{y})$ go to column F, beginning in F2.



	A	B	C	D	E	F
1	Family	Income (x)	Expenses (y)	x - x̄	y - ȳ	(x- x̄)(y - ȳ)
2	A	1	1	-6	-4	24
3	B	3	2	-4	-3	12
4	C	4	4	-3	-1	3
5	D	6	4	-1	-1	1
6	E	8	5	1	0	0
7	F	9	7	2	2	4
8	G	11	8	4	3	12
9	H	14	9	7	4	28
10	mean (x̄)	7	5			
11						

Step 6: Find the value of $(x - \bar{x})^2$ and put the results in Col G

Square each difference and put the results to column G, beginning in G**2**:

- Move cursor to column G**2**
- Type: **=D2^2**
- Click Enter. (You shall see the value of $(x - \bar{x})^2 = 36$ in column G2)
- move cursor to the corner of column G2 and drag until col G9

B13

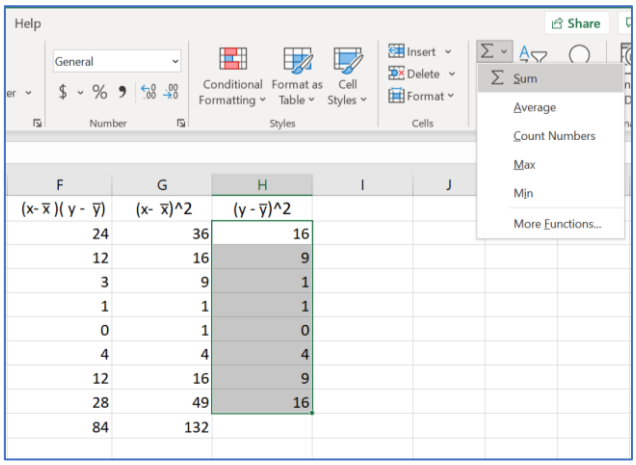
	A	B	C	D	E	F	G	H
1	Family	Income (x)	Expenses (y)	x - x̄	y - ȳ	(x- x̄)(y - ȳ)	(x- x̄)^2	(y - ȳ)^2
2	A	1	1	-6	-4	24	36	16
3	B	3	2	-4	-3	12	16	9
4	C	4	4	-3	-1	3	9	1
5	D	6	4	-1	-1	1	1	1
6	E	8	5	1	0	0	1	0
7	F	9	7	2	2	4	4	4
8	G	11	8	4	3	12	16	9
9	H	14	9	7	4	28	49	16
10	mean (x̄)	7	5					
11								

Step 7: Find the value of $(y - \bar{y})^2$ and put the results in Col H

Square each difference and put the results to column H, beginning in H**2**:

- Move cursor to column H**2**
- Type: **=E2^2**
- Click Enter. (You shall see the value of $(y - \bar{y})^2 = 16$ in column H2)
- move cursor to the corner of column H2 and drag until col H9

Step 8: Find the summation by using the SUM function: =SUM(F2:F9)



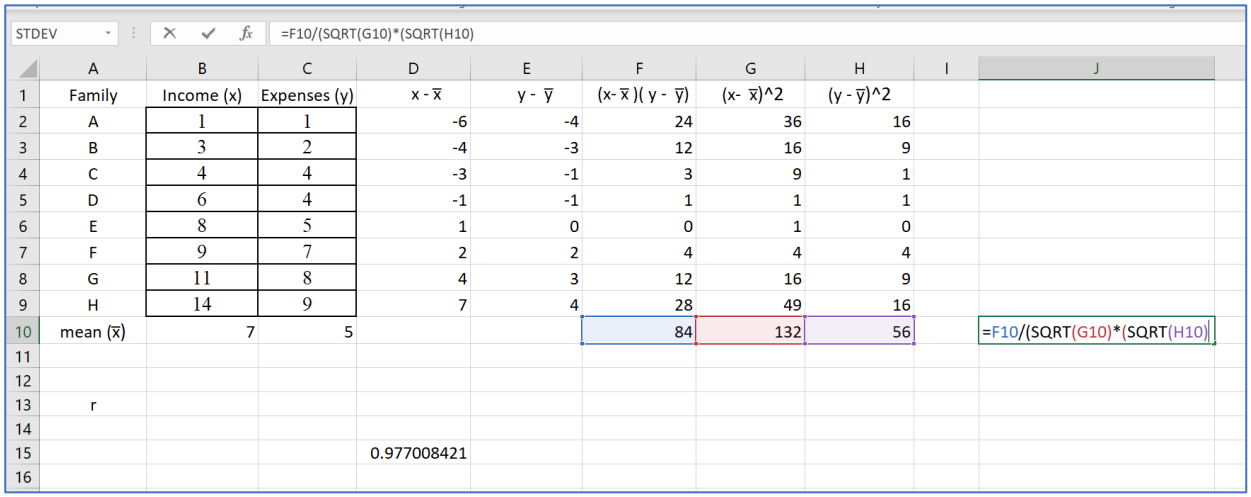
The summation goes to cell F9= 84,

Using the same process to find the summation of column G and Column H

Step 9: Calculate the correlation with the following formula

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

- Move cursor to column J10 or any empty space
- Type: = **F10/(SQRT(G10) *(SQRT(H10)**
- Click Enter. (You shall see the value of correlation coefficient (r) = 0.977008421



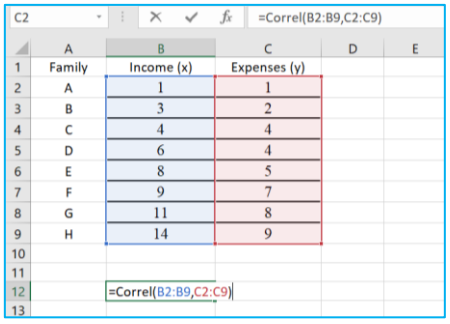
	A	B	C	D	E	F	G	H	I	J
1	Family	Income (x)	Expenses (y)	x - $\bar{x}$	y - $\bar{y}$	(x- $\bar{x}$)(y - $\bar{y}$)	(x- $\bar{x}$)^2	(y - $\bar{y}$)^2		
2	A	1	1	-6	-4	24	36	16		
3	B	3	2	-4	-3	12	16	9		
4	C	4	4	-3	-1	3	9	1		
5	D	6	4	-1	-1	1	1	1		
6	E	8	5	1	0	0	1	0		
7	F	9	7	2	2	4	4	4		
8	G	11	8	4	3	12	16	9		
9	H	14	9	7	4	28	49	16		
10	mean ($\bar{x}$)	7	5			84	132	56		=F10/(SQRT(G10)*(SQRT(H10)
11										
12										
13	r									
14										
15				0.977008421						
16										

Correlation Coefficient (r) Method 2:

Step 1 Enter all data in Excel software program as shown below

Step 2: Find the correlation coefficient (r) by using the **CORREL function**:

- Move cursor to column B12 or any empty space
- Type = Correl(B2:B9, C2:C9)
- Press Enter
- You shall get the correlation coefficient (r) = 0.977008421 on the screen.



	A	B	C	D	E
1	Family	Income (x)	Expenses (y)		
2	A	1	1		
3	B	3	2		
4	C	4	4		
5	D	6	4		
6	E	8	5		
7	F	9	7		
8	G	11	8		
9	H	14	9		
10					
11					
12		=Correl(B2:B9,C2:C9)			
13					

Exercise:

1. As shown in the table below, a person's target heart rate during exercise changes as the person gets older. Find correlation coefficient of a person's target heart rate during exercise.

Age (Years)	Target Heart Rate (beats per minute)
20	135
25	132
30	129
35	125
40	122
45	119
50	115

2. A study investigated the relationship between the amount of daily food waste measured in pounds and the number of people in a household. The data in the table displays the results of the study.

Number of people in household	Food waste (pounds)
2	3.4
3	2.5
4	8.9
4	4.7
4	3.5
4	4.0
5	5.3
5	4.6
5	7.8
6	3.2
8	12.0

   a) Draw a scatter diagram of the data above.
   b) Find correlation coefficient of the given data.

3. A teacher gives two tests to his students. The following are the scores of the mathematics and science subject of ten students.

Student	Mathematics	Science
1	15	20
2	12	15
3	10	12
4	14	18
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

   a) Draw a scatter diagram of the students' scores of the two tests.
   b) Compute the Pearson correlation coefficient, r, between the scores on the two tests.
   c) Compute the correlation coefficient, r, between the scores on the two tests by using Excel software program.