Applied Statistics I
(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

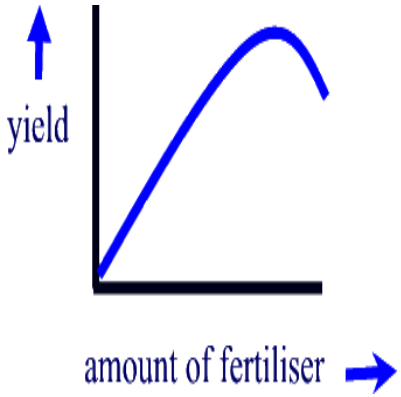
Joint distribution of data

Introduction

- Up to now we have considered distribution of only one variable.
- But in practice we meet problems with two or more variables.
- If we have problem of two variables, dependence between two variable is important.
- In statistics, **dependence** refers to any statistical relationship between two random variables or two sets of data.
- A **correlation** is a single number that describes the degree of dependence between two variables.

Correlation

1. Yield depend on the amount of fertilizer used.
2. Sales of a product depend on price charged.

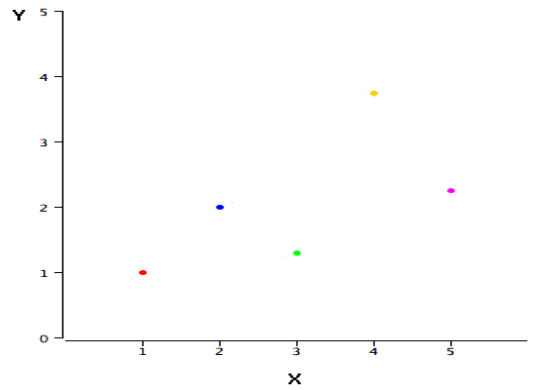


Cont...

- Possible correlations range from $+1$ to -1.
- A zero correlation indicates that there is no **linear** relationship between the variables.
- A correlation of -1 indicates a **perfect negative** correlation.
- A correlation of $+1$ indicates a **perfect positive** correlation.

Scatter plot

- A scatter plot is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data.
- The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.



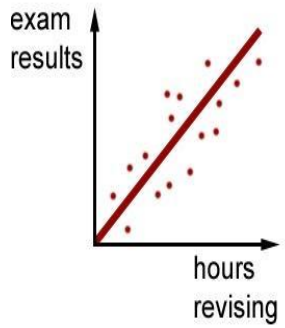
Different form of correlation

The correlation is described according to the following ways,

1. Positive or negative.
2. Simple, partial or multiple.
3. Linear or non linear.

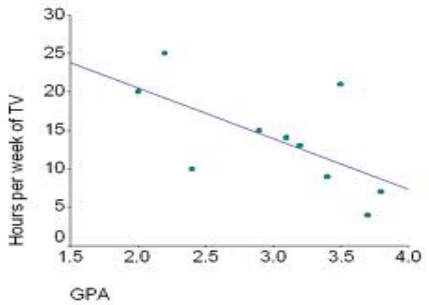
1.1 Positive correlation

- If an increase (or decrease) of values of one variable is associated with an increase (or decrease) in the corresponding vales of the other variable, they are said to be correlated.
- The correlation between these two variable is said to be **positive** or **direct**.
- **Eg:** People who do more revision get higher exam results.



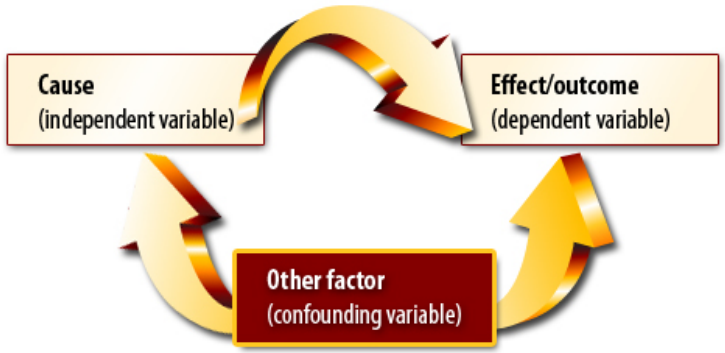
1.2 Negative correlation

- If an increase (or decrease) of values of one variable is associated with an decrease (or increase) in the corresponding vales of the other variable, they are said to be correlated.
- The correlation between these two variable is said to be **negative** or **inverse**.
- **Eg:** There is a negative correlation between TV viewing and class grades. Students who spend more time watching TV tend to have lower grades.



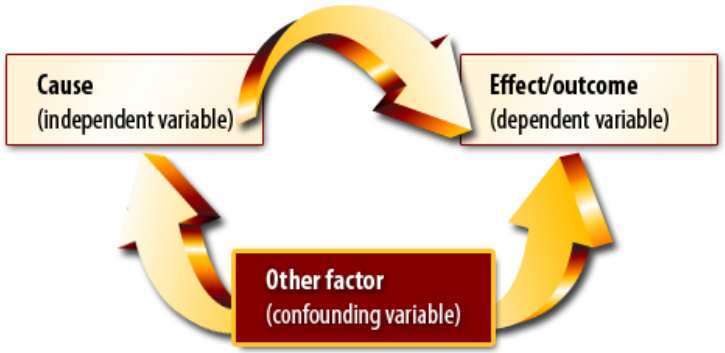
2.1 Simple correlation

- If we are studying only two variables the correlation between them is a **simple correlation**.
- In simple correlation, we measure the strength of the linear relationship between two variables, without taking into consideration the fact that both these variables may be influenced by a third variable.



2.2 Partial correlation

- **Partial correlation** analysis involves studying the linear relationship between two variables after excluding (or held constant) the effect of one or more independent factors.
- The correlation between X and Y, with the effects of Z removed (or held constant) is called the partial correlation of X and Y.

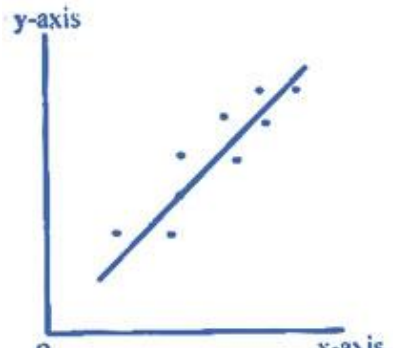


2.3 Multiple correlation

- In **multiple correlation**, we study the effects of all the independent variables simultaneously on a dependent variable.
- The correlation co-efficient between the yield of paddy (X_1) and the other variables, viz. type of seedlings (X_2), manure (X_3), rainfall (X_4), humidity (X_5) is the multiple correlation co-efficient $R_{1.2345}$.
- This co-efficient takes value between 0 and $+1$.

3.1 Linear correlation

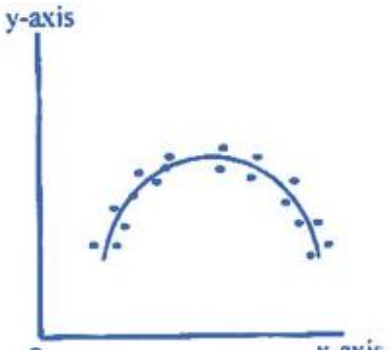
- Correlation is said to be linear if the ratio of change is constant.
- The amount of output in a factory is doubled by doubling the number of workers is the example of linear correlation.
- If all the points on the scatter diagram tends to lie near a line which are look like a straight line, the correlation is said to be linear.



Linear Correlation

3.1 Non linear correlation

- Correlation is said to be non linear if the ratio of change is not constant.
- If all the points on the scatter diagram tends to lie near a smooth curve, the correlation is said to be non linear (curvilinear).



Non Linear Correlation

Degree of correlation

- If the points form a straight line it indicates a **perfect correlation**.
- If the points form a band of some width it indicates **imperfect correlation** between two variables.
- The direction of the band shows the nature of the correlation (positive or negative).

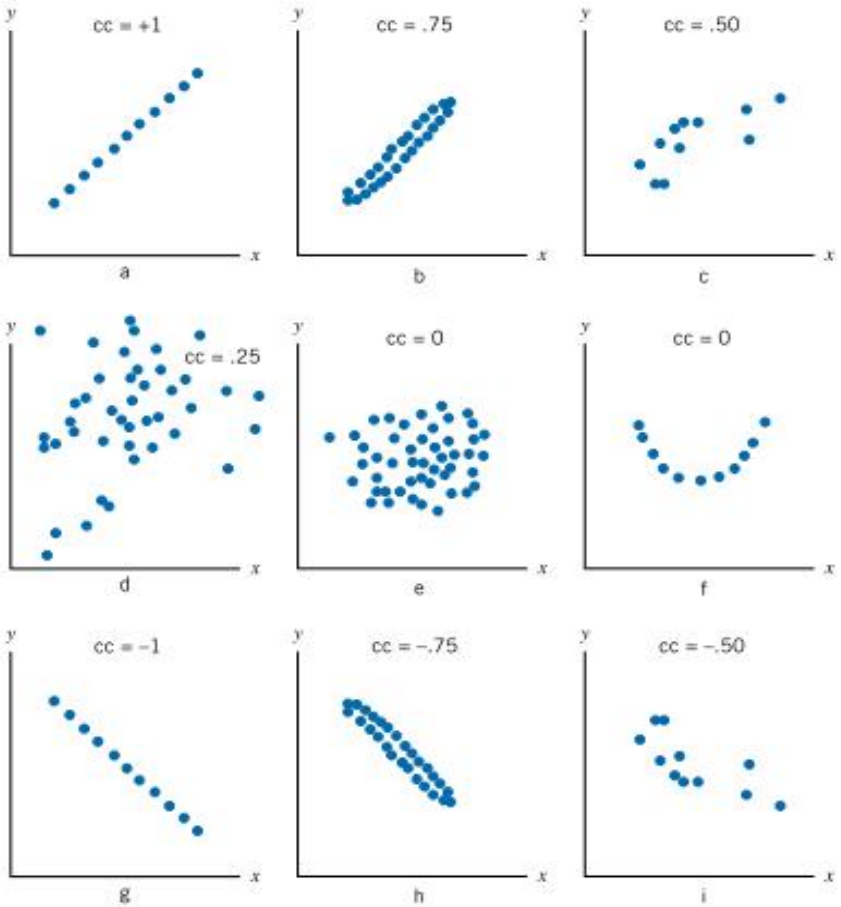
Cont...

- The width of the band gives an idea of the degree of correlation.
- The narrower the band the greater is the degree of correlation.
- When the points are scattered in all directions, it indicates that there is no correlation between the variables (i.e variables are uncorrelated).

Remark

$$\text{If two variables are independent} \Rightarrow \text{They are uncorrelated.}$$
$$\text{Uncorrelated} \not\Rightarrow \text{Independent.}$$

Scatter plots for different degree of correlations



Cont...

(a) represents a perfect correlation of $+1$, all the points fall on a perfectly straight line with a positive slope.

(b) represents a strong correlation where the behavior of one variable is similar, but not identical to the behavior of the other variable.

(c) a correlation of .50 represents a moderately strong positive relationship.

(d) relationship is weak, so the coefficient is only .25.

Cont...

(e) correlation is zero; the x and y variables are not linearly related.

(f) coefficient is also zero. This is because the correlation coefficient measures a linear association, while the relationship in Figure (f) is curvilinear.

Note:
Figures (g), (h), (i) are mirror images of Figures (a), (b), (c). All the correlation coefficients are negative.

Mathematical methods of measuring correlation

- The degree or level of correlation is measured with the help of **correlation coefficient**.
- For **population data**, the correlation coefficient is denoted by ρ.

Population covariance

- The joint variation of X and Y is measured by the **population covariance** of X and Y. The population covariance of X and Y denoted by $Cov(X, Y)$ is defined as:

$$COV(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N}.$$

Cont...

- The $Cov(X, Y)$ may be positive, negative or zero. The covariance has the same units in which X and Y are measured.
- When $Cov(X, Y)$ is divided by σ_X and σ_Y, we get the **population correlation coefficient** ρ,

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Cont...

- ρ is free of the units of measurement. It is a pure number and lies between -1 and $+1$.
- If $\rho = \pm 1$, it is called perfect correlation.
- If there is no correlation between X and Y, then $\rho = 0$.

Sample covariance

- The joint variation of X and Y is measured by the **sample covariance** of X and Y. The sample covariance of X and Y denoted by $Cov(X, Y)$ is defined as:

$$COV(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)}.$$

- For **sample data** the correlation coefficient denoted by r.

Cont...

- When $Cov(X, Y)$ is divided by S_X and S_Y, we get the **sample correlation coefficient** r,

$$r = \frac{Cov(X, Y)}{S_X S_Y}.$$

- r is free of the units of measurement.
- It is a measure of strength of the linear relation between X and Y variables.

Cont...

- On the other hand it is called as **Karl Pearsons coefficient of correlation** or **product-moment correlation coefficient** and denoted by,

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \text{ for } n \text{ pairs } (x_i, y_i).$$

Cont...

- It can also be written as,

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}.$$

Properties of r

- r is a pure number and lies between -1 and $+1$.
- The sign of r determines the nature of correlation.
- If r is positive it indicates a positive correlation.
- If r is negative it indicates a negative correlation.
- The magnitude of r determines the degree of correlation.

Cont...

- If $r = +1$, it indicates a positive perfect correlation.
- If $r = -1$, it indicates a negative perfect correlation.
- If $r = 0$, the data are uncorrelated.
- If $-1 < r < 0$, it indicates negative imperfect correlation.
- If $r < -0.85$, it indicates a strong negative correlation.
- If $0 < r < 1$, it indicates positive imperfect correlation.
- If $r > 0.85$, it indicates a strong positive correlation.

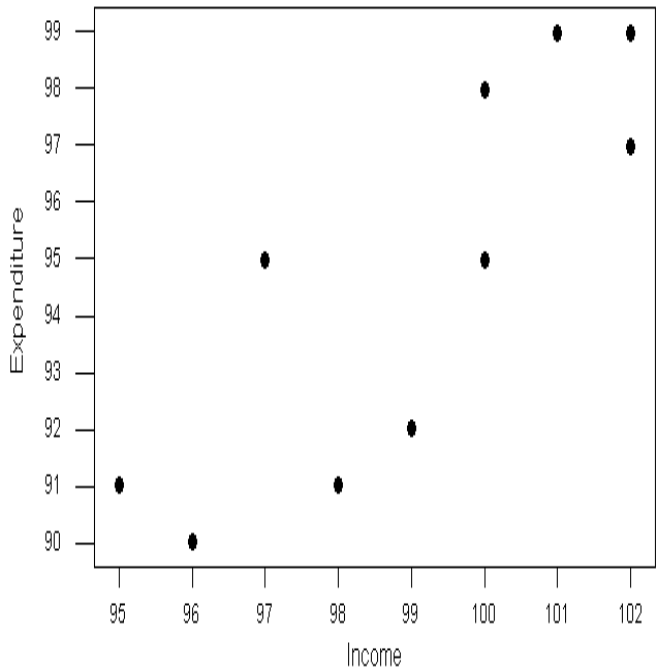
Example

The daily income and the daily expenditure of ten employees of factory are recorded as below.

Daily income	Daily expenditure
100	98
101	99
102	99
102	97
100	95
99	92
97	95
98	91
96	90
95	91

Describe the relation using Scatter plot and Karl Pearsons coefficient of correlation.

Solution


Expenditure
99
98
97
96
95
94
93
92
91
90
95
96
97
98
99
100
101
102
Income

Cont...

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
100	98	1	3	3	1	9
101	99	2	4	8	4	16
102	99	3	4	12	9	16
102	97	3	2	6	9	4
100	95	1	0	0	1	0
99	92	0	-3	0	0	9
97	95	-2	0	0	4	0
98	91	-1	-4	4	1	16
96	90	-3	-5	15	9	25
95	91	-4	-4	16	16	16
				64	54	111

Cont...

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{10} x_i}{10} = \frac{100+101+102+\ldots+95}{10} \\
&= \frac{990}{10} = 99 \\
\overline{y} &= \frac{\sum_{i=1}^{10} y_i}{10} = \frac{98+99+99+\ldots+91}{10} \\
&= 94.7 \simeq 95 \\
r &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \sum_{i=1}^{n}(y_i-\overline{y})^2}} \text{ for } n \text{ pairs } (x_i, y_i) \\
r &= \frac{64}{\sqrt{54 \times 111}} \\
&= 0.83
\end{aligned}$$

It indicates a positive imperfect correlation.

Rank correlation coefficient

- The product-moment correlation coefficient is used to measure the strength of the linear association between two variables.
- The product-moment correlation coefficient is less appropriate when the points on a scatter graph seem to follow a curve or when there are outliers on the graph.

Cont...

- The rank correlation coefficient is appropriate for those data and it is defined as,

$$\begin{aligned} \rho &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}; \text{ where} \\ d_i &= \text{difference of ranks of } i^{th} \text{ pair of observations} \\ n &= \text{number of observations.} \end{aligned}$$

Example 1

Use the raw data in the table below to calculate the correlation between the IQ of a person with the number of hours spent in front of TV per week.

IQ (x_i)	Hours of TV per week (y_i)
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Cont...

x_i	y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Cont...

By considering the last column, we can find the value of $\sum_{i=1}^{10} d_i^2$ as 194.

$$\begin{aligned} \rho &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 194}{10(10^2-1)} \\ &= -0.175757575 \end{aligned}$$

The correlation between IQ and hours spent watching TV is very low.

Thanks