

## Types of variables and their association:

A **variable** is simply anything that varies, anything that assumes different values or categories. A **variable** (in statistics) is a characteristic, attribute, or measurement that can have different "values".

Unlike the variables encountered in a basic algebra classes, the values of variables in a statistics class may be numbers, but they are not required to be. They can be categories as well. Generally, a variable will describe the members of some population in some way.

Some examples of the types of variables encountered in a statistics class:

- a person's age in years
- the number of hairs on a person's head
- the temperature of a classroom in degrees Celsius
- exam's grade for a course
- a movie's rating (from 1 to 5 stars)
- the model of a car
- one's gender

A **random variable** is one whose values are determined by chance. In a more formal way, we can define the Random Variable as follows: -

A random variable (R.V.) is any variable whose value cannot be determined beforehand; meaning before the incident. Such variables are subject to chance but the values of these variables can be restricted towards certain sets of value.

A random variable is popular in nature meaning they are presents everywhere. (Below few examples):

- The temperature in a day;
- Age of the children;
- Profit per day;
- Sales per day etc.

**Data** are the values that a variable (or variables) actually assume. Data, in mathematical and scientific speak, is a group of information collected. This information could be anything, and can be used to prove or disprove a hypothesis (or scientific guess) during an experiment.

**Data** that can be collected can be height, weight, a person's opinion on a political issue, the number of people that catch a certain cold over a year, and so much more. Data comes in a number of different types, which determine what kinds of mapping can be used for them.

The basic distinction between continuous (or **Quantitative**) and categorical data (**Qualitative**) is that Quantitative data is data where the values can change continuously; examples include weight, price, profits, counts, etc. Basically, anything you can measure or count is **quantitative**. **Categorical** data, in contrast, is for those aspects of your data where you make a distinction between different groups, and

where you typically can list a small number of categories. This includes product type, gender, age group, etc.

Types of scales & levels of measurement in Statistics

### **Discrete and continuous variables:**

**Discrete variables** are variables in which there are no intermediate values possible. For instance, the number of phone calls you receive per day. You cannot receive 6.3 phone calls.

**Continuous variables** are everything else; any variable that can theoretically have values in between points (e.g., between 153 and 154 kg. for instance). This is not all that useful of a distinction between the two types of variables. What is really more important for statistical considerations is the level of measurement used. Understanding the level of measurement of a variable (or scale or measure) is the first and most important distinction one must make about a variable when doing statistics.

Measurement scales are used to categorize and/or quantify variables. The four scales of measurement that are commonly used in statistical analysis are: nominal, ordinal, interval, and ratio scales.

**Nominal Scale of Measurement:** variable measured on a "nominal" scale is a variable that does not really have any evaluative distinction. One value is really not any greater than another. A good example of a nominal variable is sex (or gender). Information in a data set on sex is usually coded as 0 or 1, 1 indicating male and 0 indicating female (or the other way around--0 for male, 1 for female). 1 in this case is an arbitrary value and it is not any greater or better than 0. There is only a nominal difference between 0 and 1. With nominal variables, there is a qualitative difference between values, not a quantitative one.

**Ordinal Scale of Measurement:** Something measured on an "ordinal" scale does have an evaluative concept. One value is greater or larger or better than the other. Product A is preferred over product B, and therefore A receives a value of 1 and B receives a value of 2. Another example might be rating your job satisfaction on a scale from 1 to 10, with 10 representing complete satisfaction. With ordinal scales, we only know that 2 is better than 1 or 10 is better than 9; we do not know by how much. It may vary. The distance between 1 and 2 maybe shorter than between 9 and 10.

i.e. If the categories of a variable can be ranked, such as from highest to lowest or from most to least or from best to worst, then that variable is said to be **ordinal**.

**Interval Scale of Measurement:** A variable measured on an interval scale gives information about more or betterness as ordinal scales do, but interval variables have an equal distance between each value. The distance between 1 and 2 is equal to the distance between 9 and 10. Temperature using Celsius or Fahrenheit is a good example, there is the exact same difference between 100 degrees and 90 as there is between 42 and 32. i.e. If the categories can be ranked, and if they also represent equal intervals, then the variable is said to be **interval**.

**Ratio Scale of Measurement:** Something measured on a ratio scale has the same properties that an interval scale has except, with a ratio scaling, there is an absolute zero point. Weight measured in Kg is an example. There is no value possible below 0; it is absolute zero.

Below is a table that specifies the criteria that distinguishes the four scales of measurement, and the following table provides examples for each scale.

Example:

Let's consider the following example, You have collected data of the students about their weight and height as follows: (Heights and weights are not collected independently. In the below table, one row represents the height and weight of the same person).

X = Height	Y = Weight
140	45
150	50
160	55
170	60
180	90
190	100

Random Variable X & Y

Here,

- **X: Height of the students**
- **Y: Weight of the students**

Is there any relationship between height and weight of the students? If we investigate closely we will see one of the following relationships could exist

- When **X** increases, **Y** also increases.
- When **X** increases, **Y** decreases.

Such relationships need to be quantified in order to use it in statistical analysis. So the question arises, **How do we quantify such relationships?** There are 3 ways to quantify such relationship:

1. Co-variance,
2. Pearsons Correlation Coefficient (PCC),
3. Spearman Rank Correlation Coefficient (SRCC).

We will be discussing the above concepts:

### Covariance

Covariance is pretty much similar to variance. Let's shed some light on the variance before we start learning about the Covariance.

Variance generally tells us how far data has been spread from it's mean. Since mean is considered as a representative number of a dataset we generally like to know how far all other points spread out

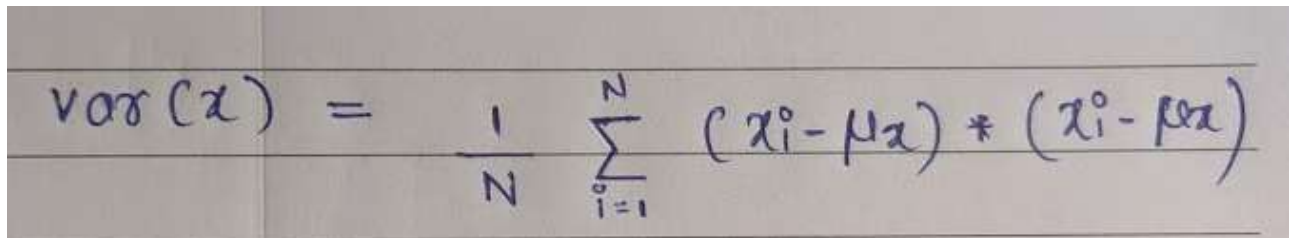
(Distance) from its mean. So basically it's average of squared distances from its mean. There are two types of variance:- Population variance and sample variance. Below table gives the formulation of both of its types.

<b>Population Variance</b>	<b>Sample Variance</b>
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
$\sigma^2$ = population variance $x_i$ = value of $i^{th}$ element $\mu$ = population mean $N$ = population size	$s^2$ = sample variance $x_i$ = value of $i^{th}$ element $\bar{x}$ = sample mean $n$ = sample size

Covariance is a measure of **how much two random variables vary together**. It's similar to variance, but where **variance tells you how a single variable varies**, co variance tells you **how two variables vary together**.

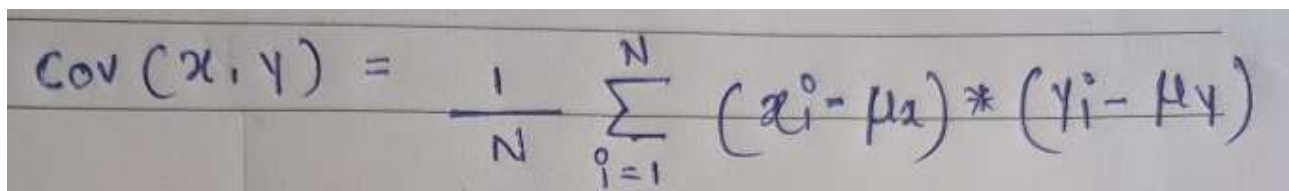
**Formulation of Covariance:**

As we have stated covariance is much similar to the concept called variance. Thus formulation of both can be close to each other.



$$\text{var}(x) = \frac{1}{N} \sum_{i=1}^N (x_i^o - \mu_x) * (x_i^o - \mu_x)$$

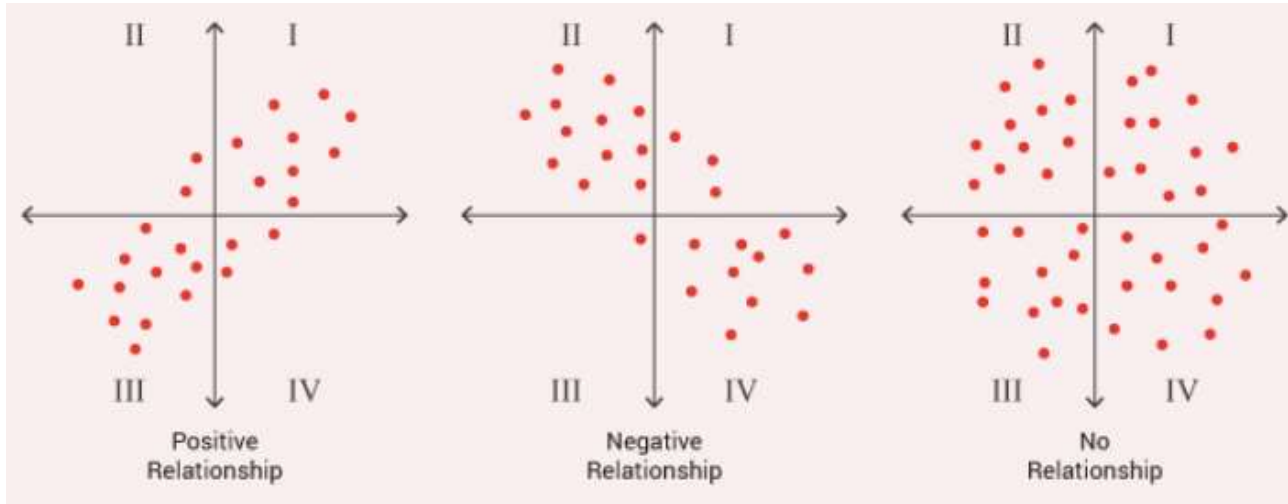
Variance of X



$$\text{cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i^o - \mu_x) * (y_i^o - \mu_y)$$

The covariance of X, Y

If you closely look at the formulation of variance and covariance formulae they are very similar to each other.



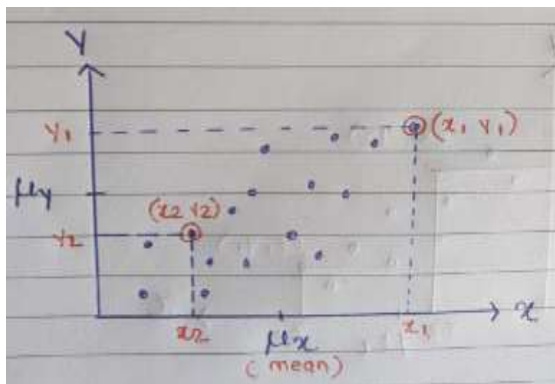
Basically we can say its measure of a linear relationship between two random variables. Based on the direction we can say there are 3 types of Covariance can be seen:-

1. Positive Covariance
2. Negative Covariance
3. Zero Covariance

**Positive Covariance:** We define there is a positive relationship between two random variables X and Y when  $Cov(X, Y)$  is positive.

- When X increases, Y also increases
- There should be a **directly proportional** relationship between two random variables.

Consider the following example.



In the above diagram, when X increases Y also gets increases. As we said earlier if this is a case then we term  $Cov(X, Y)$  is +ve.

Let's consider two points that denoted above i.e.  $(X_1, Y_1)$  and  $(X_2, Y_2)$ . The mean of both the random variable is given by  $\mu_x$  and  $\mu_y$  respectively.

- $(X_1 - \mu_x)$ , This operation returns a positive value as  $X_1 > \mu_x$ ;
- $(Y_1 - \mu_y)$ , This operation returns a positive value as  $Y_1 > \mu_y$

Thus multiplication of both positive numbers will be positive.

- $(X_2 - \mu_x)$ , This operation returns a negative value as  $X_2 < \mu_x$ ;
- $(Y_2 - \mu_y)$ , This operation returns a negative value as  $Y_2 < \mu_y$ .

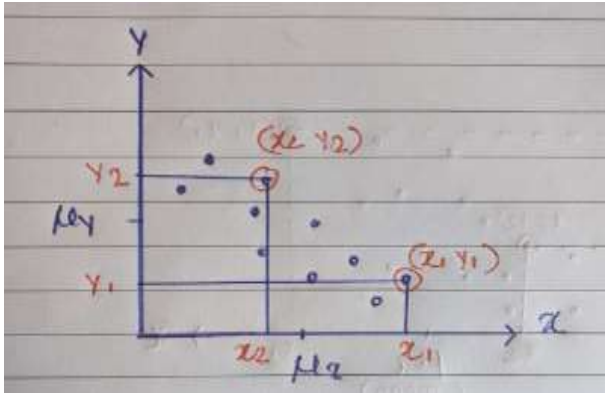
Thus multiplication of both negative numbers will be positive. Means if we have such a relationship between two random variables then covariance between them also will be positive.

### Negative Covariance

We define there is a negative relationship between two random variables X and Y when  $\text{Cov}(X, Y)$  is **-ve**.

- When **X** increases, **Y** decreases.
- When there is an **inversely proportional** relationship between two random variables.

Consider the following example,



In the above diagram, we can clearly see as X increases, Y gets decreases. This is the case of  $\text{Cov}(X, Y)$  is **-ve**. Let's check on two points  $(X_1, Y_1)$  and  $(X_2, Y_2)$ . The mean of both the random variable is given by  $\mu_x$  and  $\mu_y$  respectively.

- $(X_1 - \mu_x)$ , This operation returns a positive value as  $X_1 > \mu_x$ ;
- $(Y_1 - \mu_y)$ , This operation returns a negative value as  $Y_1 < \mu_y$

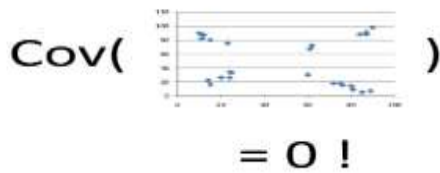
Thus multiplication of positive and negative numbers will be negative.

- $(X_2 - \mu_x)$ , This operation returns a negative value as  $X_2 < \mu_x$ ;
- $(Y_2 - \mu_y)$ , This operation returns a positive value as  $Y_2 > \mu_y$

Thus multiplication of positive and negative will be negative. Means if we have such a relationship between two random variables then covariance between them also will be negative.

### Zero Covariance

When there is **NO RELATIONSHIP** between two random variables. Then it is said to be **ZERO** covariance between two random variables. In this scenario, the data points scatter on X and Y axis such way that there is no linear pattern or relationship can be drawn from them.



This can also happen when both the random variables are independent of each other. However, the covariance between two random variables is ZERO that does not necessary means there is an absence of a relationship. A **Nonlinear relationship** can exist between two random variables that would result in a covariance value of ZERO!

### Properties of Covariance

- Covariance with itself is nothing but the variance of that variable.

$$\text{cov}(X, X) = \text{var}(X)$$

- When random variables are multiplied by constants (let's say a & b) then covariance can be written as follows:

$$\text{COV}(aX, bY) = ab \times \text{COV}(X, Y)$$

- Covariance between a random variable and constant is always ZERO;  $\text{cov}(X, a) = 0$
- $\text{Cov}(X, Y)$  is as same as  $\text{Cov}(Y, X)$

### Drawbacks of using Covariance

- When we say that the covariance between two random variables is **+ve** or **-ve** but we cannot gives the answer to **How much positive?** or **How much negative?** etc.
- Covariance is completely dependent on scales/units of numbers. Therefore it is difficult to compare the covariance among the dataset having different scales. This drawback can be solved using Pearsons Correlation Coefficient (PCC).

### Analyzing association between variables (Quantitative Variables):

When variables are related, the outcome variable is called the **Response Variable** or outcome (**Dependent Variable**) and the variable that defines the outcome variable is called the **Explanatory Variable** or predictor (**Independent Variable**).

The analysis studies how the outcome on the response variable depends on or explained by the value of explanatory variable.

**Sampling and combination of variables:**

Taking samples from population to infer some of the characteristics of the populations from which they come, what kind of relationship exists??

To answer these questions we need to consider the **Combination or Association between Variables**. There is an association between two variables if the distribution of the response variable changes in some way as the value of explanatory variable changes. There are various methods for analysing such relationship and then determines whether such association is strong enough to have practical importance.

**Measures of association:** A measure of association is a statistic that summarizes the strength of the statistical dependence between two variables or more. one of such measures is **Correlation**

**Linear Relationship:** is the simplest class of formulas that describe relationship between variables. A straight line is an example.

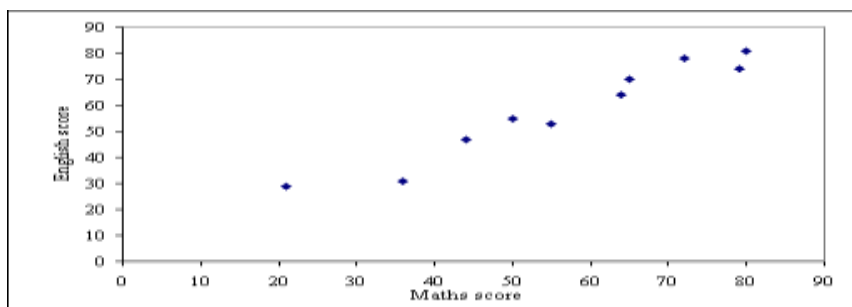
**Correlation and Correlation Coefficient:**

**Example:**

The following table shows the results of two examinations for a group of students; they are a maths exam and an English exam:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Maths score	72	65	80	36	50	21	79	64	44	55
English score	78	70	81	31	55	29	74	64	47	53

If we take a piece of graph paper and draw two axes. The horizontal axis will represent the score on the English exam. The vertical axis will represent the score on the Maths exam. For each student, we then mark a small dot at the co-ordinates representing their two scores.

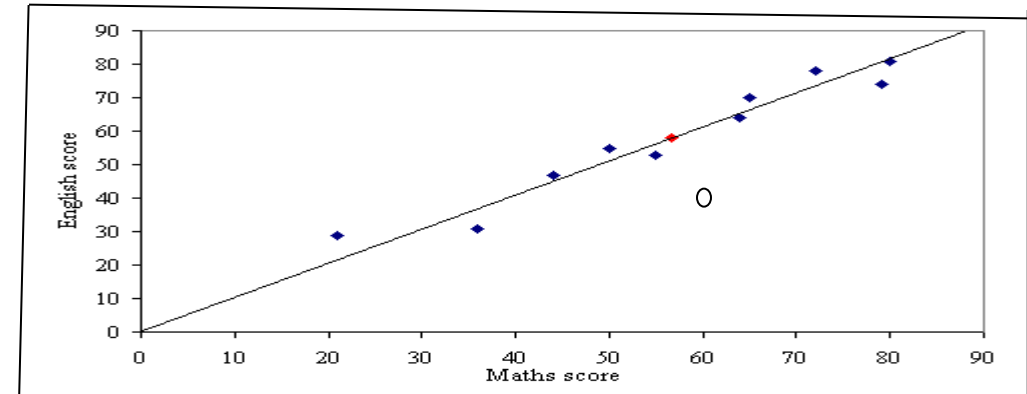


You can see that the points follow a fairly strong pattern. People who are good at maths tend to be good at English as well. The marks lie fairly close to an imaginary straight line that we can draw on



the graph. In the diagram below, in this straight line, it included another point (circled) which will be explained later.

The fact that the points lie close to the straight line is called a **Strong Correlation**. The fact that this line points upwards to right - indicating that the English mark tends to increase as the maths mark increases - is called a **Positive Correlation**.



The straight line that we draw through the points is called either **the line of best fit** or the **regression line**. It describes the relationship between the two variables (the quantities compared) mathematically. There is a standard way to draw this line to ensure that it fits as closely to the data points as possible. Later on, we will investigate exactly what that mathematical way is. For now, we only have to remember one thing:

The regression line goes through the point whose co-ordinates are the mean values of the variables. The arithmetic means are found by adding the relevant scores, and dividing by 10. This is because there are ten students in the table.

Mean maths score = 56.6

Mean English score = 58.2

and we can be sure that the line must go through the point (56.6, 58.2). This is the circled point in the graph above. You will notice that there is roughly the same number of data point lying above this line as there are below it.

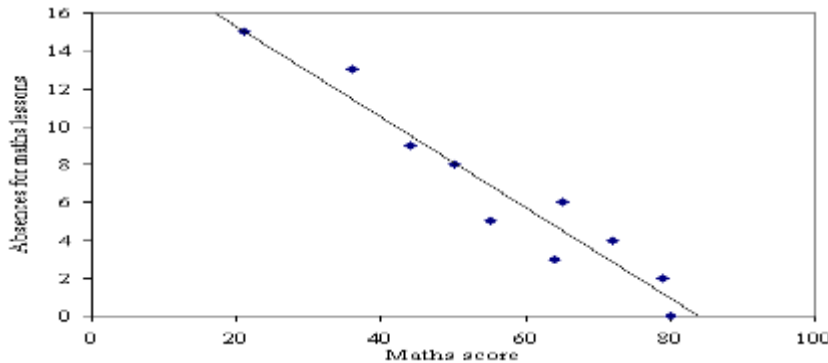
From the graph, if we look at the straight line, we can see that when the maths mark is 30, the English mark is approximately 28. Similarly, we can assume that anyone who got an English mark of 40, would also get a maths mark of about 40. However, there are limits on the predictions that we can make, as you will see later on when the regression line will be explained.

### **Negative Correlation**

In the following table, the maths marks for the ten students are duplicated and this time added the number of absences from maths lessons for each student:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Maths score	72	65	80	36	50	21	79	64	44	55
Absences	4	6	0	13	8	15	2	3	9	5

In this case, the scatter diagram looks like this.



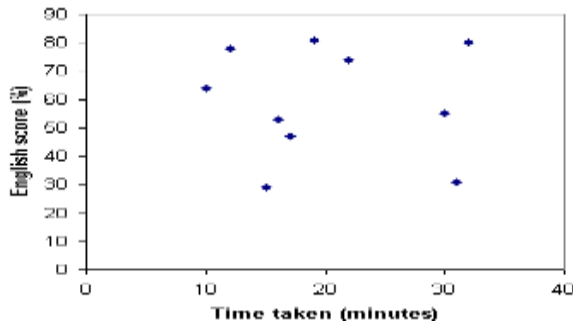
Again, there is a good correlation between the maths scores and the absences from maths lessons, except that as the number of absences increases, the maths score goes down. This is referred to as **Negative Correlation**. Again, we can use the line of best fit to make predictions. What score would a student have received if he had been absent 10 times? According to the graph, it would have been about 41. If a student received a mark of 30, how many times would you expect him to have been absent? From the graph, it seems to be about 13 times.

However, this graph shows well the limitations of making predictions. What score would someone have received if they had been absent for all 30 maths lessons? According to the graph, the score would be less than zero! Similarly, how many times would a student have had to be absent in order to gain a score of 90? Well, the line hits the horizontal axis when the score is just over 80, so in order to get a score of 90, a student would have to be absent a negative number of times.

**No correlation**

Finally, one more table, this time showing the English marks compared with the average length of time the students spend travelling to college each morning, recorded in minutes.

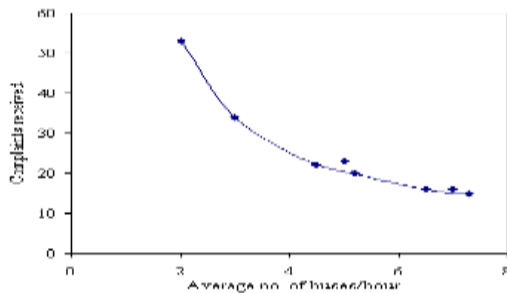
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Maths score	72	65	80	36	50	21	79	64	44	55
Absences	4	6	0	13	8	15	2	3	9	5



In this case, the scatter diagram shows no particular pattern. It is clear that we can't draw a straight line anywhere near the data points, and we say that there is no correlation between the length of time taken to travel to college and the final English mark that a student gets. We cannot predict the English mark of any student based on how long it takes him to get to college. Nor can we predict how long it takes a student to get to college given that student's English mark.

**Non-linear correlations**

A bus company wanted to discover if there was any relationship between the number of buses it ran and the number of complaints it received. It carried out a survey testing the average number of buses per hour for different days, and the number of complaints that it received on those days. Here are the results:



As you can see, there is a negative correlation between the number of buses per hour and the number of complaints, but in this case, a curved line fits the data better than a straight line. We are about to investigate the rule that lets you fit a straight line to the data points - it is enough to say at this point that similar rules exist which let you fit various curved lines to the data points as well.

**Note:** In summarizing the relationship between two quantitative variables, we need to consider:

1. Association/Direction (i.e. positive or negative)
2. Form (i.e. linear or non-linear)
3. Strength (weak, moderate, strong)

**Calculating the Correlation Coefficient**

By looking at the graph, it is possible to see whether there is a strong or weak correlation between two variables or whether the correlation is positive or negative. However, there is a mathematical

way of working it out (**Pearson's Correlation Coefficient**) represented by  $\rho$  or  $r$ , it is a single number which ranges between -1 (strong negative correlation) and +1 (strong positive correlation). Values close to 0 indicates a weak correlation with 0 itself indicates no correlation at all.

**Note:** the correlation coefficient does not reflect every type of relationship between numeric variables, but only the **Linear Relationship**. It tells to what degree the relationship between two variables can be expressed by a straight line.

**Note:** **Pearson's Correlation Coefficient** assumes ratio or scale variable, when the two variables are not ratio scales (Ordinal or Nominal) then **Kendall's or Spearman Correlation** is used.

**Definition:** Let X and Y be random variables with covariance  $\sigma_{xy}$  and standard deviations  $\sigma_x$  and  $\sigma_y$  respectively, the correlation coefficient for X and Y is given by r

$$r = \sigma_{xy} / \sigma_x \cdot \sigma_y$$

**Summary:**

- A correlation coefficient represents the strength of the linear relationship between two variables.
- A higher correlation coefficient, in absolute term, means that the points in the scatter plot lie closer to a straight line.
- Positive correlation means that if one of the variables increases the other also increases.
- A higher correlation coefficient means a stronger relationship between two variables.
- The question “is the relationship significant?” means that “are two variables independent? Or “is there a relationship between them?”. To test this hypothesis t-test is performed.

**Facts About r**

1. r measures the strength of a linear(only) relation between two quantitative variables.
2. r is strongly affected by outliers.
3. r ignores the distinction between response and explanatory variables.
4. r is not affected by changes in the unit of measurement.
5. a positive value of r means a positive association between the two variables and a negative value of r means negative association between the variables.