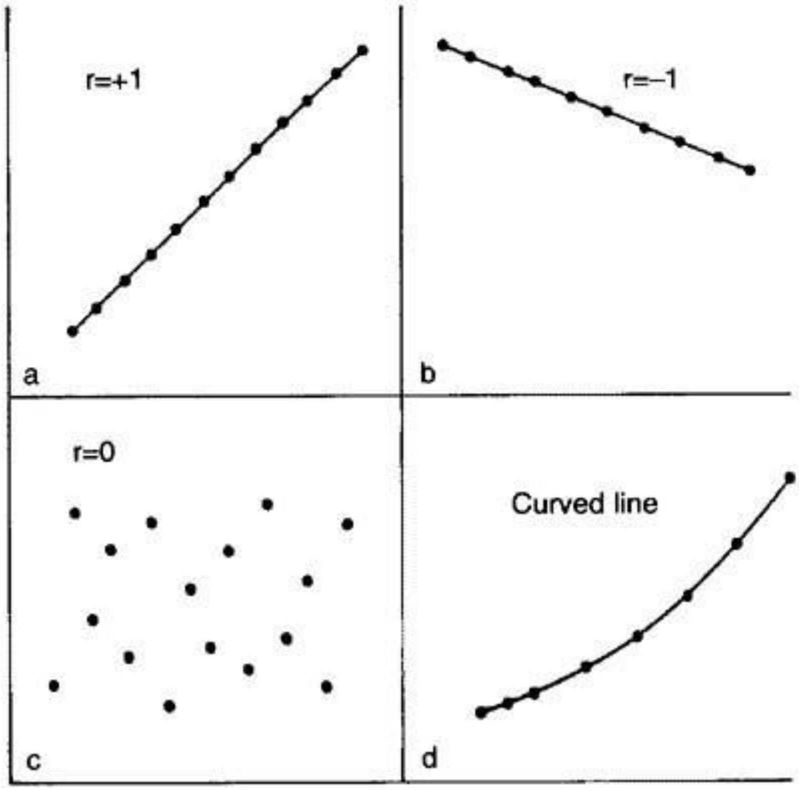
# Correlation and regression

The word correlation is used in everyday life to denote some form of association. In statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarize the association.

## Correlation coefficient

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Figure 11.1 gives some graphical representations of correlation.



## Calculation of the correlation coefficient

A paediatric registrar has measured the pulmonary anatomical dead space (in ml) and height (in cm) of 15 children. The data are given in table 11.1 and the scatter diagram shown in figure 11.2 Each dot represents one child, and it is placed at the point corresponding to the measurement of the height (horizontal axis) and the dead space (vertical axis). The registrar now inspects the pattern to see whether it seems likely that the area covered by the dots centres on a straight line or whether a curved line is needed. In this case the paediatrician decides that a straight line can adequately describe the general trend of the dots. His next step will therefore be to calculate the correlation coefficient.

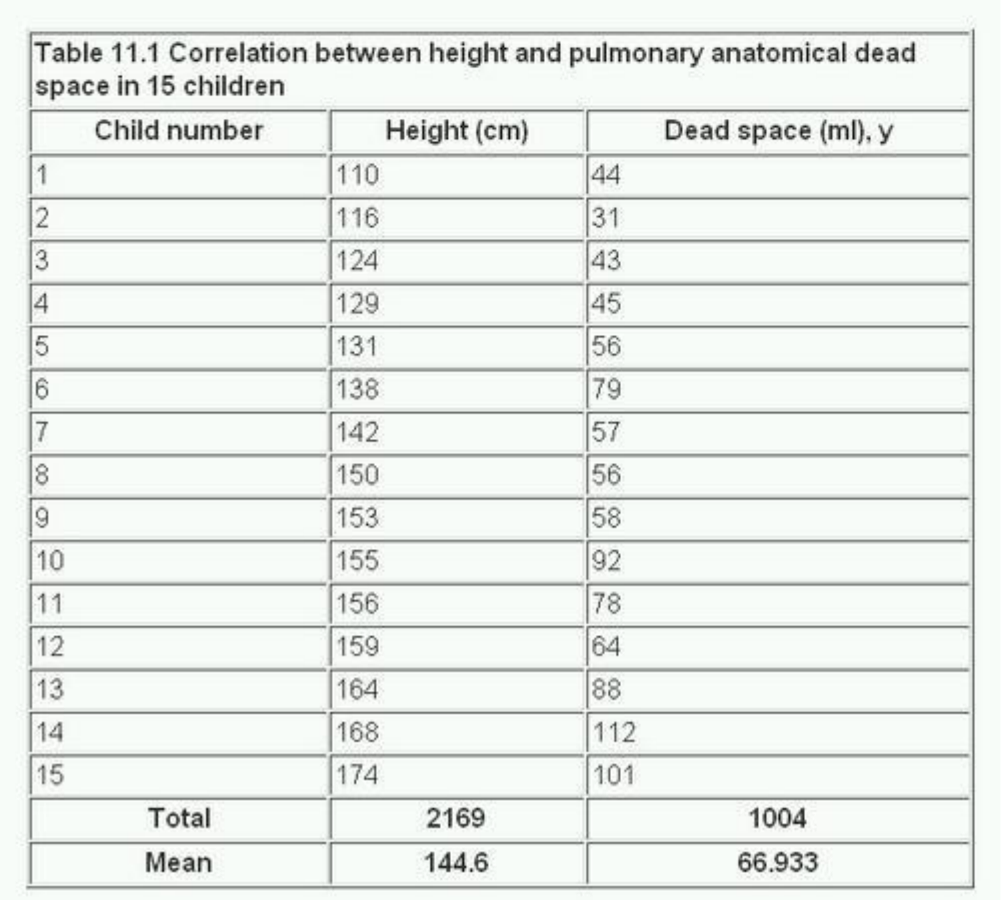


Table 11.1 Correlation between height and pulmonary anatomical dead space in 15 children

| Child number | Height (cm) | Dead space (ml), y |
|---|---|---|
| 1 | 110 | 44 |
| 2 | 116 | 31 |
| 3 | 124 | 43 |
| 4 | 129 | 45 |
| 5 | 131 | 56 |
| 6 | 138 | 79 |
| 7 | 142 | 57 |
| 8 | 150 | 56 |
| 9 | 153 | 58 |
| 10 | 155 | 92 |
| 11 | 156 | 78 |
| 12 | 159 | 64 |
| 13 | 164 | 88 |
| 14 | 168 | 112 |
| 15 | 174 | 101 |
| Total | 2169 | 1004 |
| Mean | 144.6 | 66.933 |

When making the scatter diagram (figure 11.2 ) to show the heights and pulmonary anatomical dead spaces in the 15 children, the paediatrician set out figures as in columns (1), (2), and (3) of table 11.1 . It is helpful to arrange the observations in serial order of the independent variable when one of the two variables is clearly identifiable as independent. The corresponding figures for the dependent variable can then be examined in relation to the increasing series for the independent variable. In this way we get the same picture, but in numerical form, as appears in the scatter diagram.

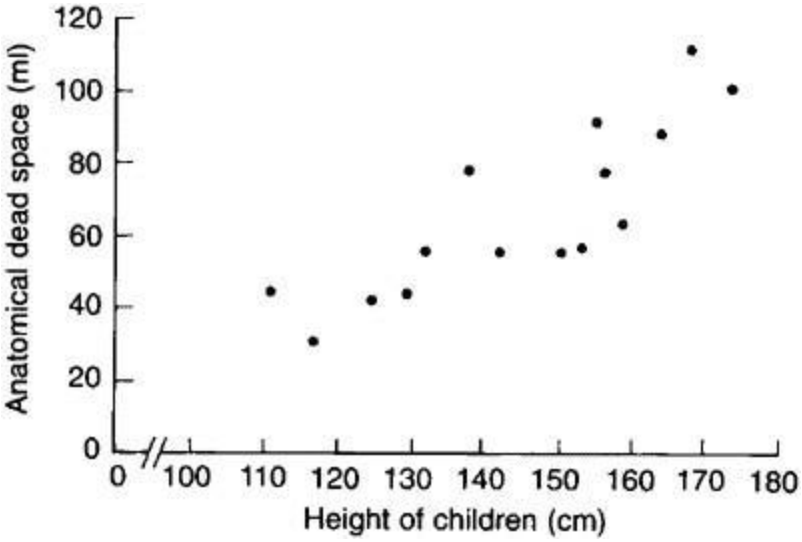


Figure 11.2 Scatter diagram of relation in 15 children between height and pulmonary anatomical dead space.

The calculation of the correlation coefficient is as follows, with x representing the values of the independent variable (in this case height) and y representing the values of the dependent variable (in this case anatomical dead space). The formula to be used is:

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

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

which can be shown to be equal to:

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD(x)SD(y)}$$

## Calculator procedure

Find the mean and standard deviation of x, as described in $\bar{x}, SD(x)$

$$\bar{x} = 144.6, SD(x) = 19.3769$$

Find the mean and standard deviation of y: $\bar{y}, SD(y)$

$$\bar{y} = 66.93, SD(y) = 23.6476$$

Subtract 1 from n and multiply by SD(x) and SD(y), (n - 1)SD(x)SD(y)

$$14 \times 19.3679 \times 23.6976 (6412.0609)$$

This gives us the denominator of the formula. (Remember to exit from "Stat" mode.)

For the numerator multiply each value of x by the corresponding value of y, add these values together and store them.

110 x 44 = *Min*

116 x 31 = *M+*

etc.

This stores $\Sigma xy \ (150605)$ in memory. Subtract $n\bar{x}\bar{y}$

MR - 15 x 144.6 x 66.93 (5426.6)

Finally divide the numerator by the denominator.

r = 5426.6/6412.0609 = 0.846.

The correlation coefficient of 0.846 indicates a strong positive correlation between size of pulmonary anatomical dead space and height of child. But in interpreting correlation it is important to remember that correlation is not causation. There may or may not be a causative connection between the two correlated variables. Moreover, if there is a connection it may be

## Spearman rank correlation

A plot of the data may reveal outlying points well away from the main body of the data, which could unduly influence the calculation of the correlation coefficient. Alternatively the variables may be quantitative discrete such as a mole count, or ordered categorical such as a pain score. A non-parametric procedure, due to Spearman, is to replace the observations by their ranks in the calculation of the correlation coefficient.

This results in a simple formula for Spearman's rank correlation, Rho.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where d is the difference in the ranks of the two variables for a given individual. Thus we can derive table 11.2 from the data in table 11.1 .

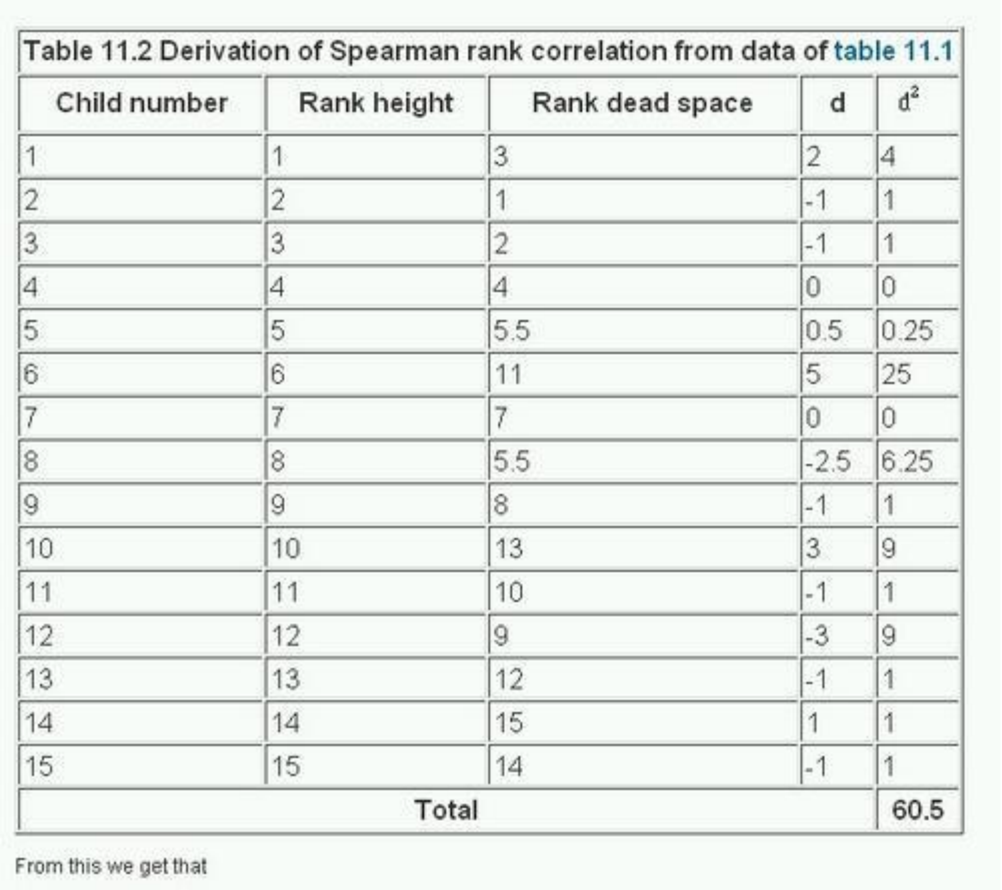


Table 11.2 Derivation of Spearman rank correlation from data of table 11.1

| Child number | Rank height | Rank dead space | d | $d^2$ |
|---|---|---|---|---|
| 1 | 1 | 3 | 2 | 4 |
| 2 | 2 | 1 | -1 | 1 |
| 3 | 3 | 2 | -1 | 1 |
| 4 | 4 | 4 | 0 | 0 |
| 5 | 5 | 5.5 | 0.5 | 0.25 |
| 6 | 6 | 11 | 5 | 25 |
| 7 | 7 | 7 | 0 | 0 |
| 8 | 8 | 5.5 | -2.5 | 6.25 |
| 9 | 9 | 8 | -1 | 1 |
| 10 | 10 | 13 | 3 | 9 |
| 11 | 11 | 10 | -1 | 1 |
| 12 | 12 | 9 | -3 | 9 |
| 13 | 13 | 12 | -1 | 1 |
| 14 | 14 | 15 | 1 | 1 |
| 15 | 15 | 14 | -1 | 1 |
| Total | | | | 60.5 |

From this we get that

From this we get that

$$r_s = 1 - \frac{6 \times 60.5}{15 \times (225 - 1)} = (0.8920)$$

In this case the value is very close to that of the Pearson correlation coefficient. For n> 10, the Spearman rank correlation coefficient can be tested for significance using the t test given earlier.

## The regression equation

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. For instance, in the children described earlier greater height is associated, on average, with greater anatomical dead Space. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. The first of these is its distance above the baseline; the second is its slope. They are expressed in the following *regression equation* :

With this equation we can find a series of values of $y_{\text{fit}}$ the variable, that correspond to each of a series of values of x, the independent variable. The parameters α and β have to be estimated from the data. The parameter signifies the distance above the baseline at which the regression line cuts the vertical (y) axis; that is, when y = 0. The parameter β (the *regression coefficient*) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards.

The regression equation is often more useful than the correlation coefficient. It enables us to predict y from x and gives us a better summary of the relationship between the two variables. If, for a particular value of x, x i, the regression equation predicts a value of y fit , the prediction error is $y_1 - y_{\text{fit}}$ . It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error $\Sigma(y_1 - y_{\text{fit}})$ of zero because the positive and negative terms exactly cancel. To remove the negative signs we square the differences and the regression equation chosen to minimise the sum of squares of the prediction errors, $S^2 = \Sigma(y_1 - y_{\text{fit}})^2$ We denote the sample estimates of Alpha and Beta by a and b. It can be shown that the one straight line that minimises $s_2$ , the least squares estimate, is given by

$$\mathbf{b} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$\mathbf{a} = \bar{y} - \mathbf{b}\bar{x}$$

it can be shown that

$$\mathbf{b} = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)SD(x)^2}$$

which is of use because we have calculated all the components of equation (11.2) in the calculation of the correlation coefficient.

The calculation of the correlation coefficient on the data in table 11.2 gave the following:

$$\Sigma xy = 150605, SD(x) = 19.3679, \bar{y} = 66.93, \bar{x} = 144.6$$

Applying these figures to the formulae for the regression coefficients, we have:

$$b = \frac{150605 - 15 \times 66.93 \times 144.6}{14 \times 19.3679^2} = \frac{5426.6}{5251.6} = 1.033\text{ml/cm}$$

$$a = 66.39 - (1.033 \times 144.6) = -82.4$$

Therefore, in this case, the equation for the regression of y on x becomes

$$y = -82.4 + 1.033x$$

This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.033 ml *over the range of measurements made*.

The line representing the equation is shown superimposed on the scatter diagram of the data in figure 11.2. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:

If x = 110, y = (1.033 x 110) - 82.4 = 31.2

If x = 140, y = (1.033 x 140) - 82.4 = 62.2

If x = 170, y = (1.033 x 170) - 82.4 = 93.2

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them.

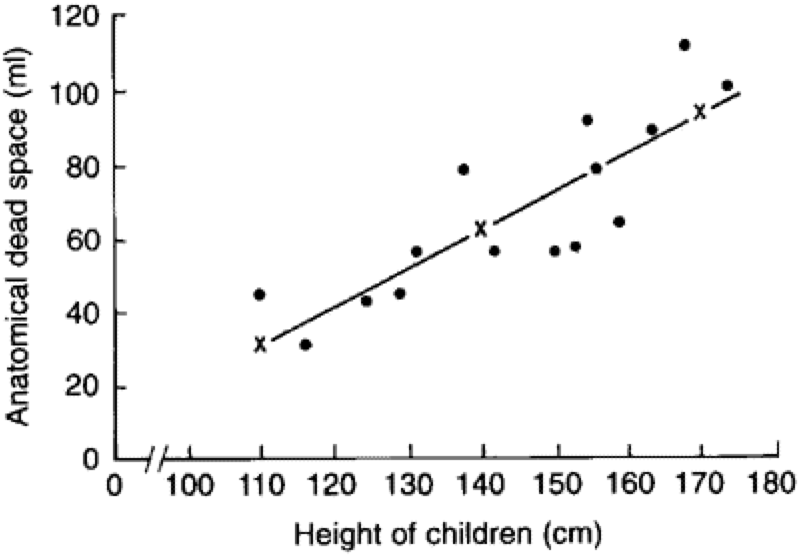


Figure 11.3 Regression line drawn on scatter diagram relating height and pulmonaiy anatomical dead space in 15 children

The standard error of the slope SE(b) is given by:

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x - \bar{x})^2}}$$

where $S_{res}$ is the residual standard deviation, given by:

$$S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n - 2}}$$

This can be shown to be algebraically equal to

$$\sqrt{((SD((y)^2(1 - r^2)(n - 1) / (n - 2)}$$

We already have to hand all of the terms in this expression. Thus $S_{res}$ is the square root of $23.6476^2(1 + -0.846^2)14 / 13 = \sqrt{171.2029} = 13.08445$. The denominator of (11.3) is 72.4680. Thus SE(b) = 13.08445/72.4680 = 0.18055.

We can test whether the slope is significantly different from zero by:

t = b/SE(b) = 1.033/0.18055 = 5.72.

Again, this has n - 2 = 15 - 2 = 13 degrees of freedom. The assumptions governing this test are:

1. That the prediction errors are approximately Normally distributed. Note this does not mean that the x or y variables have to be Normally distributed.
2. That the relationship between the two variables is linear.
3. That the scatter of points about the line is approximately constant - we would not wish the variability of the dependent variable to be growing as the independent variable increases. If this is the case try taking logarithms of both the x and y variables.

Note that the test of significance for the slope gives exactly the same value of P as the test of significance for the correlation coefficient. Although the two tests are derived differently, they are algebraically equivalent, which makes intuitive sense.

We can obtain a 95% confidence interval for b from

$$b - t_{0.05} \times SE(b) \text{ to } b + t_{0.05} \times SE(b)$$

where the tstatistic from has 13 degrees of freedom, and is equal to 2.160.

Thus the 95% confidence interval is

l.033 - 2.160 x 0.18055 to l.033 + 2.160 x 0.18055 = 0.643 to 1.422.

Regression lines give us useful information about the data they are collected from. They show how one variable changes on average with another, and they can be used to find out what one variable is likely to be when we know the other - provided that we ask this question within the limits of the scatter diagram.

Source: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

Q1. Find the Karl Pearson’s Coefficient of correlation for the following data and interpret it.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X | 7 | 5 | 4 | 11 | 10 | 12 | 14 | 9 |
| Y | 14 | 8 | 8 | 19 | 16 | 19 | 20 | 16 |

Q2. Find the Karl Pearson’s Coefficient of correlation for the following data.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X | 14 | 8 | 10 | 11 | 9 | 13 | 5 |
| Y | 14 | 9 | 11 | 13 | 11 | 12 | 4 |

Q3. Find the Spearman’s Rank Correlation Coefficient for the following data.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 15 | 32 | 25 | 30 | 35 | 20 | 19 | 22 | 27 | 31 |
| y | 50 | 70 | 65 | 72 | 90 | 58 | 53 | 57 | 68 | 74 |

Q4. Find the Spearman’s Rank Correlation Coefficient for the following data.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 12 | 15 | 13 | 20 | 15 | 14 | 19 | 13 | 21 | 18 |
| y | 25 | 21 | 15 | 18 | 20 | 17 | 20 | 16 | 20 | 22 |

Q5. Calculate the Spearman’s coefficient of Correlation between ages of husband (x) and ages of Wife (y), both are expressed in years, from the following data.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X | 60 | 30 | 37 | 30 | 42 | 37 | 55 | 45 |
| Y | 50 | 25 | 33 | 27 | 40 | 33 | 50 | 42 |

Q6. From the following data, find the regression equation of y on x and hence estimate y when x = 13.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| x | 14 | 10 | 15 | 11 | 9 | 12 | 6 |
| y | 8 | 6 | 4 | 3 | 7 | 5 | 9 |

Q7. From the following data, find the regression equation of x on y and hence estimate x when y = 10 .

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| x | 11 | 7 | 9 | 5 | 8 | 6 | 10 |
| y | 16 | 14 | 12 | 11 | 15 | 14 | 17 |

Q8. Given the following data, find the regression of x on y and estimate x when y = 35.The Correlation Coefficient is 0.65 .

| | X | Y |
|---|---|---|
| Mean | 43 | 37 |
| S.D | 3.1 | 2.8 |

Q9. Given the two regression equations 2x-y-15 = 0 and 3x – 4y + 25 = 0.
Find (i) the mean values of x and y
(ii) the coefficient of Correlation r .

Q10. Given the two regression equations 5x – 6y + 90 = 0 and 15x -8y -180 = 0. Find the mean values of x and y, the coefficient of correlation r.

Q11. Define Correlation and explain various types of correlation.

Q12. Explain scatter diagram and types of correlation.

## Moving Average

A moving average is a technique to get an overall idea of the trends in a data set; it is an average of any subset of numbers. The moving average is extremely useful for **forecasting long-term trends**.

An average represents the "middling" value of a set of numbers. The moving average is exactly the same, but **the average is calculated several times for several subsets of data.** For example, if you want a two-year moving average for a data set from 2000, 2001, 2002 and 2003 you would find averages for the subsets 2000/2001, 2001/2002 and 2002/2003. Moving averages are usually plotted and are best *visualized*.

### Calculating a 5-Year Moving Average Example

**Sample Problem:**Calculate a five-year moving average from the following data set:

| Year | Sales ($M) |
|---|---|
| 2003 | 4 |
| 2004 | 6 |
| 2005 | 5 |

| | |
|---|---|
| 2006 | 8 |
| 2007 | 9 |
| 2008 | 5 |
| 2009 | 4 |
| 2010 | 3 |
| 2011 | 7 |
| 2012 | 8 |

The mean (average) sales for the first five years (2003-2007) is calculated by finding the mean from the first five years (i.e. adding the five sales totals and dividing by 5). This gives you the **moving average for 2005 (the center year)** = 6.4M:

| Year | Sales ($M) |
|---|---|
| 2003 | 4 |
| 2004 | 6 |
| 2005 | 5 |
| 2006 | 8 |
| 2007 | 9 |

(4M + 6M + 5M + 8M + 9M) / 5 = 6.4M

The average sales for the **second subset of five years (2004 – 2008)**, centered around 2006, is **6.6M**:
(6M + 5M + 8M + 9M + 5M) / 5 = 6.6M

The average sales for the **third subset of five years (2005 – 2009)**, centered around 2007, is **6.6M**:
(5M + 8M + 9M + 5M + 4M) / 5 = 6.2M
source: https://www.statisticshowto.com/moving-average/

Q1. Find three yearly moving averages and draw these on the graph paper. Also represent the original time series on the graph.

| Year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|
| Production (in thousand units) | 12 | 15 | 20 | 18 | 25 | 32 | 30 | 40 | 44 |

Q2. Find five yearly moving averages and draw these on the graph paper. Also represent the original time series on the graph.

| Year | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales | 51 | 53 | 56 | 57 | 60 | 55 | 59 | 62 | 68 | 70 |

Q3. Find the moving averages of length 4 for the following data. Represent the given data and the moving averages on the graph paper.

| Year | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales(in thousand units) | 60 | 69 | 81 | 86 | 78 | 93 | 102 | 107 | 100 | 109 |

Q4. Fit straight line trend by the method of least squares for the following data Representing production in thousands units . plot the data and the trend line on a graph paper. Hence or otherwise estimate the trend for the year 1991.

| year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|
| Production (in thousands Units ) | 14 | 15 | 17 | 16 | 17 | 20 | 23 |

Q5. Fit a straight line trend to the following time series, representing sales in Lakhs of Rs. of a company for the years 1998 to 2005.Plot the given data as well as the trend line on the graph paper. Hence or otherwise estimate trend for the year 2006.

| Year | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|---|
| Sales(in Lakhs of Rs.) | 31 | 33 | 30 | 34 | 38 | 40 | 45 | 49 |

Q6. Find the seasonal component of the time series, using method of seasonal indices .

| **Season**<br>**Year** | **i** | **ii** | **iii** | **iv** |
|---|---|---|---|---|
| **2003** | 33 | 37 | 32 | 31 |
| **2004** | 35 | 40 | 36 | 35 |
| **2005** | 34 | 38 | 34 | 32 |

| | | | | |
|---|---|---|---|---|
| **2006** | 36 | 41 | 35 | 36 |
| **2007** | 34 | 39 | 35 | 32 |

Q7. For the following data calculate

a. Laspeyre’s Index No.
b. Fishers Index No.

| Commodity | Base Year | | Current Year | |
|---|---|---|---|---|
| | Price | Quantity | Price | Quantity |
| A | 4 | 10 | 5 | 12 |
| B | 3 | 8 | 6 | 10 |
| C | 2 | 8 | 3 | 9 |
| D | 5 | 4 | 8 | 5 |

Q8. For the following data calculate

a. Paasche’s Index No.
b. Fishers Index No.

| Commodity | Base Year | | Current Year | |
|---|---|---|---|---|
| | Price | Quantity | Price | Quantity |
| A | 4 | 10 | 5 | 12 |
| B | 3 | 8 | 6 | 10 |
| C | 2 | 8 | 3 | 9 |
| D | 5 | 4 | 8 | 5 |

Q9. From the following data calculate the cost of living index number for 2006 by the family budget method.

| Group | Price in 2001 | Price in 2006 | Weight |
|---|---|---|---|
| Food | 15 | 36 | 60 |
| Clothing | 48 | 96 | 5 |
| Lighting & Fuel | 30 | 90 | 10 |
| Rent | 60 | 180 | 15 |
| Miscellaneous | 45 | 90 | 10 |

Q10. From the following data calculate the cost of living index number for 2006 by the Aggregative Expenditure Method..

| Commodity | Quantity (Year 2000) | Price in 2000 | Price in 2006 |
|---|---|---|---|
| Rice | 10 | 12 | 18 |
| Wheat | 15 | 9 | 15 |
| Milk | 5 | 18 | 24 |

| | | | |
|---|---|---|---|
| Sugar | 6 | 15 | 24 |
| Pulses | 8 | 30 | 36 |
| Oil | 4 | 48 | 72 |

Q11. For the following data , calculate the chain base Index Number for the following Data:

| Year | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|
| Prices | 23 | 28 | 35 | 45 | 52 |

Q12. For the following data , calculate the chain base Index Number for the following Data:

| | **Average Sales** | | | |
|---|---|---|---|---|
| **Commodity** | **2002** | **2003** | **2004** | **2005** |
| **A** | 35 | 39 | 42 | 45 |
| **B** | 38 | 45 | 52 | 60 |
| **C** | 42 | 51 | 56 | 65 |

Q13. What is a time series? Describe the various components of time series.

Q14. What are seasonal variations? Explain briefly with examples.

## Binomial Distribution

- Applied to single variable discrete data where results are the numbers of “successful outcomes” in a given scenario.
  e.g.: no. of times the lights are red in 20 sets of traffic lights,
  no. of students with green eyes in a class of 40,
  no. of plants with diseased leaves from a sample of 50 plants

- Used to calculate the probability of occurrences *exactly, less than, more than, between* given values
  e.g. the “probability that the number of red lights will be exactly 5”
  “probability that the number of green eyed students will be less than 7”
  “probability that the no. of diseased plants will be more than 10”

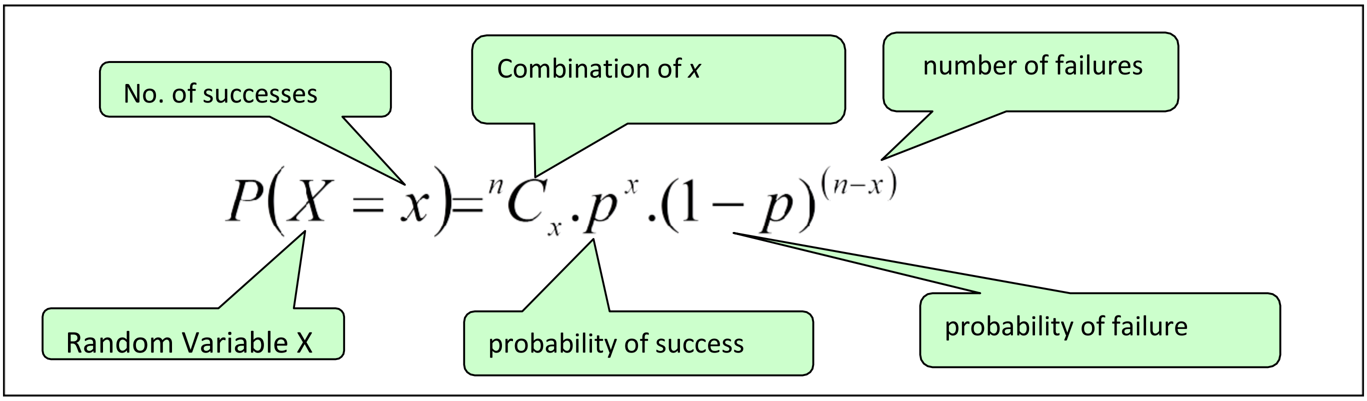
- Parameters, statistics and symbols involved are:

| | population parameter symbol | sample statistic symbol |
|---|---|---|
| probability of success | $\pi$ | $p$ |
| sample size | $N$ | $n$ |

- Other symbols:
  $X$ , the number of successful outcomes wanted

  $^nC_x$ , or $^nC_r$ : the number of ways in which $x$ “successes” can be chosen from sample size $n$ . The formula.

- Formula used: $^{n}C$ key on your calculator can be used directly in the



$$P(X = x) = {}^{n}C_{x}.p^{x}.(1-p)^{(n-x)}$$

## Binomial Distribution

- Applied to single variable discrete data where results are the numbers of "successful outcomes" in a given scenario.
  e.g.: no. of times the lights are red in 20 sets of traffic lights,
  no. of students with green eyes in a class of 40,
  no. of plants with diseased leaves from a sample of 50 plants

- Used to calculate the probability of occurrences *exactly, less than, more than, between* given values
  e.g. the "probability that the number of red lights will be exactly 5"
  "probability that the number of green eyed students will be less than 7"
  "probability that the no. of diseased plants will be more than 10"

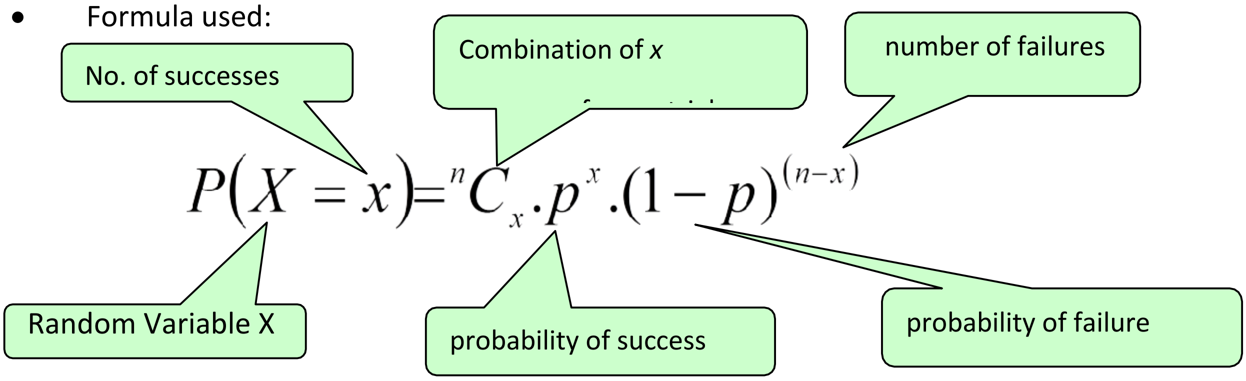
- Parameters, statistics and symbols involved are:

| | population parameter symbol | sample statistic symbol |
|---|---|---|
| probability of success | $\boldsymbol{\pi}$ | $\boldsymbol{p}$ |
| sample size | $N$ | $\boldsymbol{n}$ |

- Other symbols:
  $\boldsymbol{X}$ , the number of successful outcomes wanted

  $^{n}C_{x}$, or $^{n}C_{r}$ : the number of ways in which $\boldsymbol{x}$ "successes" can be chosen from sample size $\boldsymbol{n}$ . The $^{n}C_{r}$ key on your calculator can be used directly in the formula.

- Formula used:



Read as "the probability of getting ' $\boldsymbol{x}$ ' successes is equal to the number of ways of choosing ' $\boldsymbol{x}$' successes from $n$ trials *times* the probability of success to the power of the number of successes required *times* the probability of failure to the power of the number of resulting failures."

## Poisson Distribution

This is often known as the *distribution of rare events*. Firstly, a Poisson process is where DISCRETE events occur in a CONTINUOUS, but finite interval of time or space. The following conditions must apply:

- For a small interval the probability of the event occurring is proportional to the size of the interval.
- The probability of more than one occurrence in the small interval is negligible (i.e. they are rare events). Events must not occur simultaneously
- Each occurrence must be independent of others and must be at random.
- The events are often defects, accidents or unusual natural happenings, such as earthquakes, where in theory there is no upper limit on the number of events. The interval is on some continuous measurement such as time, length or area.

The parameter for the Poisson distribution is $\lambda$ (lambda). It is the average or mean number of occurrences over a given interval.

The probability function is:

$$\boxed{p(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!} \quad \text{for } x = 0,\ 1.\ 2,\ 3...}$$

## Normal Distribution

- Applied to single variable continuous data
  e.g. heights of plants, weights of lambs, lengths of time

- Used to calculate the probability of occurrences *less than, more than, between* given values
  e.g. "the probability that the plants will be less than 70mm",
  "the probability that the lambs will be heavier than 70kg",
  "the probability that the time taken will be between 10 and 12 minutes"

- Standard Normal tables give probabilities - you will need to be familiar with the Normal table and know how to use it.
  First need to calculate how many standard deviations above (or below) the mean a particular value is, i.e., calculate the value of the "standard score" or "Z-score".
  Use the following formula to convert a raw data value, $X$ , to a standard score, $Z$ :

$$Z = \frac{(X - \mu)}{\sigma}$$

Source: https://library2.lincoln.ac.nz/documents/Normal-Binomial-Poisson.pdf

## PROBLEMS BASED ON BINOMIAL, POISSON AND NORMAL DISTRIBUTION

Q1. An unbiased cubical dice is thrown 5 times and the number appearing on its uppermost face is noted. Find the probability that the number of times an even number appears is

a) 3 times
b) At least 4 times

Q2. It is observed that 60% of students of a class are vegetarians . If 7 students from the class are selected at random, find the probability that :-

a) 3 are vegetarians
b) 4 or 5 are vegetarians

Q3. If the mean and variance of a Binomial distribution are 4 and 2.4 respectively , Find probability of

(i) 5 successes (ii) 8 successes

Q4. A manufacture of ball pens knows that 5% of his products is defective. If he sells pens in boxes of 100 and guarantees not more than 10 pens are defective, What is the approximate probability that the box will contain

a) At least one defective pen
b) 2 or more defective pens

(Given: $e^{-5} = 0.006738 = 0.0067$)

Q5. It is observed that 1% of mangoes in a box are bad. find the Probability that in a box of 100 mangoes , number of bad mangoes is

(i) only 1
(ii) less than 2

(given : $e^{-1} = 0.3679$ )

Q6. Akash receives , on an average , 5 messages per day. Find the probability that on a specific day , he will receive

(i) only 2 (ii) only 3

(Given : $e^{-5} = 0.0067$ )

Q7. If X follows a normal distribution with mean 120 and variance 1600,

Find

a) $P( X \leq 140)$
b) $P( X \geq 110)$

Q8. If X is a normal variate with mean 40 and standard deviation 8, find

(i) $P( x \geq 42)$ (ii) $P(x \leq 39 )$

Q9.In an intelligence test administered to 1000 persons, the average I.Q. was 80 with a standard deviation of 15.

a) How many persons had their I.Q. between 70 and 110?
b) What was the percentage of persons with I.Q. above 100?

Q10.1500 candidates appeared for a certain examination. The mean marks were 58 with a standard deviation of 5 marks. Assuming the distribution of marks to be Normal. Find

a) The proportion of the students securing more than 63 marks
b) The number of students securing marks between 60 and 68

Q11.1500 candidates appeared for a certain examination. The mean marks were 58 with a standard deviation of 5 marks. Assuming the distribution of marks to be Normal. Find

a) The percentage of students with marks below 53.
b) The number of students securing marks between 60 and 68

Q12. Write any five properties of Normal Distribution.

Q13. Write any five properties of Poissson Distribution.

Q14. Write any five properties of Binomial Distribution.