Government of India
Ministry of Earth Sciences
India Meteorological Department
Meteorological Training Institute
Lecture notes

On
Meteorological Statistics

For E-learning phase of Forecaster’s
Training Course

Prepared

By
Dr. D.M. Rase
Correlation Analysis

1. Introduction: -

The various statistical parameters such as averages, measures of dispersion, skewness and kurtosis etc supply ample information about the variable under study and help the analyst in drawing various inferences. However these do not explain the mechanism behind the variation. For ex. The daily temperature of a station may be caused by advection of moisture or cold / warm air, time of onset of sea breeze etc. To comprehend the variation of a variable it may be necessary to study as to how it varies jointly with other variables. Correlation analysis is an important statistical tool to achieve this objective.

2. Definition:

Whenever two variables X and Y are so related that a change in one is accompanied by change in the other in such a way that an increase in the one accompanied by an increase or decrease in the other, then two variables are said to be correlated.

For Ex:  
(i) Heights and weights of students in a class.
(ii) Amount of fertilizer per acre and the yield of grain per acre on a number of different plots.

3. Positive and negative correlation: - Correlation can be either positive or negative. When the values of two variables move in same direction so that increase in the value of one variable is associated with an increase the value of the other variable also, and a decrease in the value of one variable is associated with the decrease in the value of the other variable also, correlation is said to be positive.

If on the other hand, the values of two variables move in different directions, so that with an increase in the value of one variable the value of the other variable decreases, and with decrease in the value of one variable the value of other variable increases, correlation is said to be negative.

When an increase in one variable leads to increase in the other the correlation is positive or direct.

For ex:
(i) Rainfall and the grain output of a state.
(ii) Height and weight of students in a class.
(iii) Dew point and minimum temperature of a station.

When an increase in one variable leads to decrease in the other the correlation is negative or inverse.
For ex:
  i) Height and temperature in the troposphere.
  ii) Height and pressure.
  iii) Pressure and volume of a perfect gas.

4. **Linear and non-linear correlation**:

   When the variation in the values of two variables is in constant ratio, correlation is said to be linear. Thus if 10% increase in price each time the supply by 20%, there is a linear relationship in these two variables. Their relationship is of the type $Y = a + bx$. (equation of straight line). If the corresponding values of two such series are plotted on a graph paper a straight line would be obtained.

   In economic data the ratio of change in two variables is generally not constant. In such cases corresponding figure of two variables would not give a straight line. The correlation may be curvi-linear or non-linear. Thus, linear correlation is one where the ratio of variation in the related variables is constant and non-linear correlation is one where this ratio is fluctuating.

5. **Properties of Coefficient of Correlation**

   1. The Coefficient of Correlation lies between -1 to +1, symbolically $-1 \leq r \leq +1$
   2. The Coefficient of Correlation is independent of change of scale and origin of the variables $X$ and $Y$
   3. The Coefficient of Correlation is the geometric mean of two regression coefficient, symbolically $r = \sqrt{b_{xy} \cdot b_{yx}}$

6. **Methods of study of Correlation**:

   Various methods of ascertaining whether two variables are correlated or not are:

   (i) Scatter Diagram Method
   (ii) Karl Pearson’s Coefficient of Correlation
   (iii) Rank Method
   (iv) Regression

6.1 **Scatter Diagram Method**:

   In this method the given data are plotted on a graph paper in the form of dots. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the variables. The more closely the points come to a straight line, the higher the
degree of relationship. If all the points lie on the straight line, the correlation is said to be
perfectively positive or negative i.e. \( r = +1 \) or \( -1 \).

By applying this method we can get an idea about the direction of correlation and also
whether it is high or low. But we can not establish the exact degree of correlation.

| Positive linear correlation | Curvilinear correlation | No correlation |

Fig. 1 Scatter diagrams which reveal different types of correlations
6.2 Karl Pearson’s Coefficient of Correlation (KPCC) Or Product moment Coefficient of Correlation

Assumptions:

The Karl Pearson’s Coefficient of Correlation is based on the following assumptions:

i) There is linear relationship between the variables i.e. when two variables are plotted on a scatter diagram a straight line will be formed by the points so plotted.

ii) The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc. are affected by such forces that a normal distribution is formed.

iii) There is a cause and effect relationship between the forces affecting the distribution of the items in the two series. For example, there is no relationship between income and height because the forces that affect these variables are not common.

Of the several mathematical methods measuring correlation, Karl Pearson’s Coefficient of Correlation (KPCC) is most widely used in practice. The Pearson’s Coefficient of Correlation denoted by \( r \). The formula for computing Coefficient of Correlation \( r \) is

\[
 r = \frac{\Sigma XY}{N \sigma_x \sigma_y}
\]

Here \( X = x-\bar{x}, \; Y = y-\bar{y} \)

\( \sigma_x \) = Standard deviation of series \( x \)

\( \sigma_y \) = Standard deviation of series \( y \)

\( N \) = Number of pairs of observations

\( r \) = Karl Pearson’s Coefficient of Correlation

Another expression for \( r \) is

\[
 r = \frac{\Sigma XY}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}}
\]

Where \( X = x-\bar{x}, \; Y = y-\bar{y} \)
Merits and Limitations of the Karl Pearson’s Coefficient of Correlation (KPCC)

Amongst the mathematical methods used for measuring the degree of relationship, KPCC method is most popular. The Coefficient of Correlation summerizes in one figure not only the degree of correlation but also the direction, i.e. whether correlation is positive or negative.

However, the utility of this coefficient depends on the skill of user.

The chief limitations of this method are:

1 Coefficient of Correlation always assumes linear relationship regardless of the fact whether that assumption is correct or not.
2 Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.
3 The value of the coefficient is unduly affected by the extreme items.
4 As compared with other methods this method takes more time to compute the value of correlation coefficient.

6.3 Rank Correlation Coefficient

The Karl Pearson’s method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal, or when the shape of the distribution is not known, there is a need for a measure of correlation that involves no assumption about the parameters of the population. It is possible to avoid making any assumptions about the populations being studied by ranking the observations according to size and biasing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest or it may be the smallest. Using the ranks rather than actual observations gives the coefficient of rank correlation.

This method of finding out co-variability or the lack of it between two variables was developed by the British psychologist, Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individuals in the group can be arranged in order there by obtaining for each individual a number indicating his/her rank in the group, Spearman’s rank Correlation Coefficient (R) is defined as

\[
R = 1 - \frac{6 \sum D^2}{N(N^3-N)}
\]

Where D refers to the difference of ranks between paired items in two series and N is the total number of pairs.
**Merits:**
1) This method is simpler to understand and easier to apply compared to the Karl Pearson’s method.
2) Where the data are of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with greater advantage.
3) This is only method that can be used where we are given the ranks and not the actual data.
4) Even where actual data are given, rank method can be applied for ascertaining rough degree of correlation

**De Merits**
1) This method can not be used for finding out the correlation in a grouped frequency distribution
2) Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time.

**Utility of correlation Analysis in Meteorology:**

i) To comprehend the dispersion of a variable with the help of another variable. This helps in understanding the behavior of a variable. In meteorology temperature, pressure humidity, rainfall etc. mutually interact. Objective study of their joint variation can be undertaken with correlation analysis.

ii) Once two variables are well correlated, one variable can be predicted (estimated) on the basis of other.
Regression analysis

1. Introduction: -

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable from the given value of the other variable. For example if we know that the yield of rice and rainfall are closely related we may find out the amount of rain required to achieve a certain production figure.

2. Definition:-

The statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called ‘Regression’.

3. Convention:-

The variable, which is the basis of prediction, is called independent variable (say X) and the variable that is to be predicted is referred to as the dependent variable (say Y).

4. Difference between Correlation and Regression Analysis: -

i) The Correlation coefficient is a measure of degree of co-variability between X and Y, where as the objective of regression analysis is to study the ‘nature relationship’ between the variables so that we may be able to predict the value of one on the basis of another.

ii) The cause and effect relation is clearly indicated through regression analysis than by correlation.

5. Linear Regression: -

Two variables are said to have linear relationship when change in the independent variable (say X) by one unit leads to constant absolute change in the dependent variable (say Y).

When two variables have linear relationship the regression lines can be used to find out the values of dependent variable.

6. Regression Lines: -

When we plot two variables (say X and Y) on a scatter diagram and draw two lines of best fit, which pass through the plotted points, these lines are called regression lines. In linear regression, these lines are straight ones. These regression lines are based on two equations
called regression equations, which will give best estimate of one variable when the other is exactly known or given. We have two regression lines as the regression of X on Y and the regression of Y on X. The regression line of Y on X gives most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y. However, when there is either perfect positive or perfect negative correlation between the two variables the two regression lines will coincide i.e. we will have only one line. The regression lines cut each other at the point of average of X and Y, i.e. if from the point where both the regression lines cut each other a perpendicular is drawn on the X-axis, we will get the mean value of X and if from that point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.

7. **Regression Equations**:-

Regression Equations also known as estimating equations are algebraic expressions of the regression lines. Since there are two regression lines, there is two-regression equations - regression equation of X on Y is used to describe the variation in the values of X for given changes in Y and regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

8. **Regression Equation of Y on X**:-

Regression equation of Y on X is expressed as follows:

\[ Y = a + bX \]

In this equation a and b are constants (fixed numerical values) which determine the position of the line completely. These constants are called the parameters of the line. If the value of either or both of them is changed, another line is determined. The parameter ‘a’ determines the level of the fitted line (i.e. the distance of the line directly above or below the origin). The parameter ‘b’ determines the slope of the line, i.e. change in X. The symbol Y stands for the value of Y computed from the relationship for a given X.

If the values of the constants ‘a’ and ‘b’ are obtained, the line is completely determined. To obtain these values the method of ‘Least Squares’ is used, which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviations of the actual Y values from the computed Y values is the least, or in other words, in order to obtain a line which fits the points best

\[ \Sigma (Y - \overline{Y}_c)^2 \] should be minimum. Such a line is known as the line of ‘best fit’
A straight line fitted by least squares has the following characteristics:

i) It gives the best fit to the data in the sense that it makes the sum of the squared deviations from the line,
   \[ \Sigma (Y - Y_c)^2 \], smaller than they would be from any other straight line. This property accounts for the name ‘Least Squares’

ii) The deviations above the line equal those below the line, on the average. This means that the total of the positive and negative deviations is zero, or \( \Sigma (Y - \bar{Y}_c) = 0 \).

iii) The straight line goes through the overall mean of the data (X, Y)

iv) when the data represent a sample from a larger population, the least square line is a ‘best’ estimate of the population regression line.

With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters a and b such that the least squares requirement is fulfilled:

\[ \Sigma Y = Na + b \Sigma X \]
\[ \Sigma XY = a \Sigma X + b \Sigma X^2 \]

These equations are usually called the normal equations. In the equations \( \Sigma X, \Sigma Y, \Sigma XY, \Sigma X^2 \) indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values.

9. **Regression Equation of X on Y:**

Regression equation of X on Y is expressed as follows:

\[ X = a + b Y \]

To determine the values of a and b the following two normal equations are to be solved simultaneously.

\[ \Sigma X = Na + b \Sigma Y \]
\[ \Sigma XY = a \Sigma Y + b \Sigma Y^2 \]

10. **Another form of Regression equations** (Deviations taken from Arithmetic Mean of X and Y)
The above method of finding out regression equations is tedious. The calculations can very much be simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means. In such a case the two regression equations are written as follows:

11. Regression Equation of X on Y:-

\[(X - \bar{X}) = r (\frac{\bar{x}}{\bar{y}}) (Y - \bar{Y})\]

Where \(\bar{X}\) is the mean of X-series and \(\bar{Y}\) is the mean of Y-series.

\(r (\frac{\bar{x}}{\bar{y}})\) is known as the regression coefficient on X on Y.

The regression coefficient of X on Y is denoted by the symbol \(b_{xy}\). It measures the change in X corresponding to a unit change in Y.

\[b_{xy} = r (\frac{\bar{x}}{\bar{y}}) = \frac{\Sigma xy}{\Sigma y^2}\] Where \(x = X - \bar{X}\) and \(y = Y - \bar{Y}\)

11. Regression Equation of Y on X:-

\[(Y - \bar{Y}) = r (\frac{\bar{y}}{\bar{x}}) (X - \bar{X})\]

\(r (\frac{\bar{y}}{\bar{x}})\) is the regression coefficient of Y on X is denoted by the symbol \(b_{yx}\). It measures the change in Y corresponding to a unit change in X.

Ex: Estimate the value of ISMR from given data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISMR (percentage departure from normal %)</td>
<td>-0.4</td>
<td>9.9</td>
</tr>
<tr>
<td>Nino 3.4 SST (anomaly(^0)C)</td>
<td>-0.2</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Correlation coefficient between ISMR & SST = -0.38

**Solution:**

By using above equation we can estimate the value of ISMR as follows:

Here ISMR = Y and SST = X hence, \(\bar{X} = -0.2\), \(\bar{Y} = -0.4\), \(\bar{x} = 0.64\), \(\bar{y} = 9.9\) and \(r = -0.37\)

Substituting the values we get
Y \ (-0.4) = -0.38(9.9/0.64)(X\ (\ -0.2))

Y= - 5.72X – 1.54 i.e. ISMR= -5.72 SST-1.54

12. Relation between regression coefficient and correlation coefficient:

The square root of the product of two regression coefficient gives the value of
Correlation coefficient. In other words the correlation coefficient is a Geometric Mean of
two regression coefficient. Symbolically, \( r = \sqrt{b_{xy} \times b_{yx}} \)

Proof:
\( b_{yx} = r \left( \frac{\bar{y}}{\bar{x}} \right) \) and \( b_{xy} = r \left( \frac{\bar{x}}{\bar{y}} \right) \)
\( b_{yx} \times b_{xy} = r \left( \frac{\bar{y}}{\bar{x}} \right) \times r \left( \frac{\bar{x}}{\bar{y}} \right) = r^2 \)
\( r^2 = \sqrt{b_{yx} \times b_{xy}} \)

Note:-

The following points should be noted about the regression coefficients:

i) Both regression coefficients will have the same sign, i.e. either they will have positive
or negative.

ii) Since the value of coefficient of correlation cannot exceed one, one of the regression
coefficients must be less than one or, in other words, both regression coefficients
cannot be greater than one.

iii) The coefficient of correlation will have the same sign as that of regression
coefficients, i.e. if regression coefficient have a negative sign, coefficient of
correlation (r) will also be negative and if regression coefficient have a positive sign,
coefficient of correlation (r) will also be positive.

For example:
If \( b_{yx} = -0.8 \) and \( b_{xy} = -1.2 \), \( r = \sqrt{-0.8 \times -1.2} = -0.98 \) and not + 0.98.
Since \( b_{yx} = r \left( \frac{\bar{y}}{\bar{x}} \right) \), we can find out any of the four values given other three.

***********************************
Sampling Theory

Sampling is an important concept in Statistics. Statistics is collection, compilation, analysis and interpretation of data. When we deal with large quantum of data, we often use the word Population or Universe. A population could be finite, infinite, existent or hypothetical. A part or small section selected from the population is called a sample and the process of such selection is called sampling.

Often it is not possible to explore a large population even if it is finite. In some cases, we may not at all be able to access the entire population. Collection of data of a very large population is time consuming, expensive and frequently error prone, whereas a sample of the population could be studied with great precision, accuracy at a much lower cost. Sampling is used for testing a hypothesis about the parent population from where sample is drawn. It is also used to estimate the population parameters given sample parameters.

Sampling is a vast subject and only a basic and preliminary portion would be covered in this note.

Random sampling

Random sampling is the fundamental assumption on which most of the sampling theory is based. In a random sampling each individual of the same population has the same chance of being selected and the selection of one item in no way influences the selection of the others. Thus a random sample is a sample of no bias.

A few examples of population, sample, random samples:

Suppose a school has 500 students. We want to select a random sample of size 25. The roll numbers of all the students are written in separate cards and put inside a box which is thoroughly shaken and 25 cards picked up. This is an example of a perfect random sampling. Here the set of 500 students is the population.

Now consider the ISMR data for the period 1901-2000, which constitutes a set of 100 values. Is it a population or sample? Strictly it is a sample only, as the population should be deemed to contain ISMR values backwards for several centuries. But we do not have ISMR values prior to 1871. Thus the ISMR population can be deemed as a hypothetical population. The sample data of 1901-2000 is not at all a random sample as data is taken continuously for several years. In a restricted way, the data of 1901-2000 (or 1871-2012) can be considered as a population also.
**Sampling distribution**

Suppose, we have a large population of size \(N\) and that we would like to select a random sample of size \(n\). Now this can be done in \(k = \binom{N}{n}\) different ways. Suppose \(N=1, 00,000\) and \(n=1000\), \(\binom{N}{n}\) is obviously a very large number. Now let us consider a statistic such as AM. For each of the \(k\) samples, we can compute the AM and the set of these \(k\) AMs will have a frequency distribution which is called the sampling distribution and in this case that of AM. Similarly, we can have sampling distribution of median, standard deviation and of any other statistical parameter. Sampling theory revolves substantially on the characteristics of the sampling distribution.

**Sampling distribution of AM – large samples**

Suppose there is a population with mean \(\mu\) and SD \(\sigma\), normally distributed. Let \((x_1, x_2, \ldots, x_n)\) be a random sample drawn from the population \(N(\mu, \sigma)\). The AM of the sample mean \(\bar{x}\) is

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}
\]

(1)

In (1) \(x_1, x_2, \ldots, x_n\) all take all the values of the population and so

\[
AM(\bar{x}) = \sum \frac{AM(x_i)}{n} = \frac{n\mu}{n} = \mu
\]

(2)

Thus the sample mean \(\bar{x}\) is distributed with the population mean \(\mu\) itself. Now,

\[
Var(\bar{x}) = Var\left(\frac{x_1 + x_2 + \ldots + x_n}{n^2}\right)
\]

(3)

As \((x_1, x_2, \ldots, x_n)\) is a random sample, but each \(x_i\) assuming all the values, the variables \(x_1, x_2, \ldots, x_n\) can all be considered independent and if \(r_{ij}\) is the CC between \(x_i\) and \(x_j\) then \(r_{ij} = 0\). Thus \(Var(\sum x_i) = \sum Var(x_i), i = 1, \ldots n\)

\[
Var(x_i) = \sigma^2
\]

(4)

Thus (3) becomes

\[
Var(\bar{x}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \text{ i.e. } SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}
\]

(5)

The result (5) should be considered interesting. It shows that when the population is distributed with mean and SD as \((\mu, \sigma)\) the sampling distribution of random samples of size \(n\)
is distributed with \((\mu, \frac{\sigma}{\sqrt{n}})\). The SD of \(\bar{x}\) is called the Standard Error (SE). The reliability of the sampling process improves if the SE is small. From (5), it follows that SE can be made smaller if \(n\) is taken as large, but the decrease is not proportional. For eg. If \(n=25\), \(SE = \frac{\sigma}{5}\). If \(n=100\), \(SE = \frac{\sigma}{10}\). In this case, to reduce the SE by 50%, the sample size has to be made bigger by 400%.

Now what is the distribution of \(\bar{x}\) the sample mean? It can be shown that if \(X\) and \(Y\) follow ND so is \(X + Y\). Thus \(\bar{x}\) is also normally distributed i.e. \(N (\mu, \frac{\sigma}{\sqrt{n}})\).

**Principle behind sampling tests**

Let us consider a hypothetical distribution of height of adult males in a large community distributed normally with \(N (66, 4)\), unit in inches. We know that 95% of area lies within the ordinates \(\pm 1.96\) for the ND \(N(0,1)\), and 99% lies within \(\pm 2.58\). For the given example of \(N (66, 4)\), the 95% and 99% intervals are \((58.2, 73.8)\) and \((55.7, 76.3)\) respectively. Now if an individual’s height is 6’2” i.e.74”, then we can say that he does not belong to the population as per the 95% limits. Is this conclusion correct? Yes by 95%. But there is a probability of 5% that we may declare a person belonging to the community as not belonging. This 5% is called the Level of Significance (LS) or Type I Error.

If we use the 99% limits then a person of height 74” would comfortably fit into the population. The LS here is 1% and the Type I error is 0.01. Thus chance of a genuine person being considered as non-genuine is considerably decreased. But chance of a non genuine person being taken as genuine increase, which is called Type II error. Discussion on this error is beyond the scope of this note.

Let us again return to \(N (66, 4)\). We select a random sample of size 100 and find that the sample mean is 67. Now the sample mean \(\bar{x}\) is distributed with AM 66 but \(SD = \frac{4}{\sqrt{100}} = 0.4\) i.e. \(N (66, 0.4)\). The difference between population and sample mean is 1 which is 2.25 SE. Thus under 5% LS, we conclude that the sample does not belong to the population but if 1% LS is used it belongs to the population. By and large the above methodology is the principle behind the sampling tests.
Central Limit Theorem

We have seen that if a population is normal, then the sample mean also follows Normal Distribution (ND). Now, what if the distribution of the population is not known? It can be shown that \( \bar{x} \) follows ND provided \( n \) is large even if the parent population were not normal. This remarkable result is from Central Limit Theorem, dealt in theoretical statistics books.

Student’s t distribution

Let us consider \( y = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \) \( \ldots \) (6)

We have observed that this follows \( N(0,1) \) even if the parent population is non-normal provided \( n \) is large. It can be taken that the sample is large when \( n \geq 30 \). If \( n < 30 \), then \( y \) in (6) follows \( t \)-distribution. ‘\( t \)’ is defined by

\[
t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \ldots \ldots \ldots (7)
\]

where \( s \) is the SD of the sample but defined by the modified formula

\[
s = \sqrt{\frac{\sum(x-x)^2}{n-1}} \ldots \ldots \ldots (8)
\]

The quantity \( v = n-1 \) is called degrees of freedom (df). The \( t \)-distribution frequency curve extends from \( -\infty \) to \( +\infty \). The area under the \( t \)-curve for various df are given in Table 1. Here, 2\textsuperscript{nd} and 4\textsuperscript{th} columns correspond to \( \alpha =0.01 \) (2 tails) and \( \alpha =0.05 \) (2 tails) only need to be considered. (Other columns can be ignored as of now).

Fig. 2 depicts the shape of the \( t \)-curve, for various df.

Tests of significance

Difference between Population and Sample means

i) Large sample \( n \geq 30 \) : Suppose \( \mu \) is the population mean and \( \sigma \) the SD. Let \( \bar{x} \) be the sample mean and \( n \) the sample size. \( SE = \frac{\sigma}{\sqrt{n}} \). If \( \sigma \) is not known \( s \), the sample SD can be computed and \( SE \) can be taken as \( \frac{s}{\sqrt{n}} \). Now compute \( y = \frac{\bar{x} - \mu}{s / \sqrt{n}} \).
If $|y| \leq 1.96$, the difference between population and sample mean is not significant at 5% LS.

If $|y| > 1.96$, the difference is significant.

If $|y| \leq 2.58$, difference not significant at 1% LS.

If $|y| > 2.58$, difference significant at 1% LS.

![Shape of the t-curve for various df](image)

**Fig. 2** Shape of the $t$-curve for various df

**Table 1** Area under the $t$-curve for various degrees of freedom (df)
(ii) Small samples \( n<30 \):

Compute \( t_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \)

The \( df \) \( \nu = n-1 \). Get the theoretical value of \( t \) from Table 1 for the desired LS.

If \( |t_c| > t \) difference is significant.

If \( |t_c| \leq t \) difference not significant.

Example: Suppose \( n=20 \), Then \( \nu = 19 \). For 19 df the value of \( t \) is 2.09 for 5% LS and 2.86 for 1% LS, which are higher than 1.96 and 2.58 when the sample is large.

**Difference between two sample means**

(i) For large samples: Let \( \mu, \sigma \) be the population mean and SD respectively.
Two random samples of sizes $n_1$ and $n_2$ are drawn from the populations with mean and SD as $(\bar{x}_1, s_1)$ and $(\bar{x}_2, s_2)$. Let $z = \bar{x}_1 - \bar{x}_2$ be the difference of means. Theoretically, $z$ should be distributed with mean 0. We have to test whether the difference between the two means is significant or not. We have $\bar{z} = 0$

$$SE(z) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \ldots \ldots \quad (9)$$

Compute $y = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \ldots \ldots \quad (10)$

If $\sigma$ is not known same is estimated from $s_1$ and $s_2$. In that case we get

$$y = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad \ldots \ldots \quad (11)$$

If $|y| \geq 1.96/2.58$ difference significant at 5% / 1% LS.

If $|y| < 1.96/2.58$ difference not significant at 5% / 1% LS.

(ii) For small samples

$$t_c = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \quad \ldots \ldots \quad (12)$$

The df here is $v = n_1 + n_2 - 2$. Get the theoretical values of $t$ for $v$ df.

If $|t_c| \geq t$ significant ; If $|t_c| < t$ not significant ; L.S. as appropriate.

Fiducial intervals for population mean given sample mean

Frequently the parent population’s characteristics are seldom known and we have to infer them only from the sample. Suppose $\bar{x}$ is the sample mean, $s$ the SD and $n$ the sample size. The 95% fiducial or confidential limits for population mean are

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \quad \text{and} \quad \text{the 99% limits are} \quad \bar{x} \pm 2.58 \frac{s}{\sqrt{n}}.$$
If \( n < 30 \), we should obtain the value of \( t \) for \( n-1 \) df for 5% and 1% levels. If \( t_5 \) and \( t_1 \) are the theoretical values then \( \bar{x} \pm t_5 \frac{s}{\sqrt{n}} \) and \( \bar{x} \pm t_1 \frac{s}{\sqrt{n}} \) are the confidential limits at 5% and 1% LS respectively.

*To Test the Significance of an observed Correlation Coefficient*

Let us consider the CC between ISMR and SST / SOI. The CC’s computed are -0.3752 and 0.3449 respectively based on 50 year data 1951-2000. Whereas these values are the actual CC values, we can also consider the CC’s as computed from a sample drawn from a still large population of values of ISMR, SST and SOI. In that case, whether CC is significant or not is a question which is addressed.

Generally, we denote the population CC by \( \rho \) and the sample CC by \( r \). \( r \) is computed from the sample data. A null hypothesis assumption of \( \rho = 0 \), i.e. the two variables \( x \) and \( y \) are unrelated is made. Thus \( x \) and \( y \) are independent. Then, we proceed to test whether the null hypothesis is true or false. We compute the statistic \( t \) from the formula

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

(13)

The \( df = n - 2 \). Suppose \( t_c \) is the computed value and \( t_{0.05} \) and \( t_{0.01} \) are theoretical values of \( t \) are 5 and 1% LS respectively. If \( n \geq 30 \), then \( t_{0.05} = 1.96 \) and \( t_{0.01} = 2.58 \).

If \( |t| \geq t_c \), then \( \rho \neq 0 \), i.e. the CC is significant.

If \( |t| < t_c \), then \( \rho = 0 \), i.e. the CC is not significant.

Thus the testing of significance of an observed CC only attempts to find out whether there is ‘some’ relation or not. It may be noted that significance may not imply a high degree of relationship which has to be derived from the observed CC only

*Sampling in meteorological data*

Application of conventional theory of sampling to meteorological data poses several problems. Samples drawn from such a population are by and large not random. Most of the meteorological data are likely to be in the form of time series and choosing a random sample generally does not serve much purpose. The weather shows persistence with reference to time and space. Suppose, we choose a random sample of 250 rain gauges out of nearly 5000 rain gauges of India from IMD’s network. Can we use the 250 stations data as a sample and proceed
to estimate the statistics of population mean, say that of India? Perhaps not, even if the stations are picked up at random, the weather parameters would still show dependence.

Meteorologists have therefore devised several ingenious methods to modify the various sampling tests and still apply them to meteorological data. Again consider ISMR data of 1901-2000. This is a yearly data which is derived once a year. The CC between two consecutive years is near 0. We can therefore consider this sample as random for all practical purposes.

**Practical exercises in sampling**

The ISMR is distributed with mean 891.5 mm and SD 85.6 mm during the period 1901-2000. Considering this as a large sample derive the 95 % and 99% confidence intervals of the population mean of ISMR.

*Solution:* Here $\bar{x} = 891.5$, $s=85.6$, $n=100$. Thus this is a large sample. As the population SD is not known, we take $\sigma = s = 85.6$, $SE = \frac{s}{\sqrt{n}} = \frac{85.6}{10} = 8.6$ mm. The 95 % limits are $891.5 \pm 1.96 \times 8.6 = (874.7, 908.3)$. The 99 % limits are $891.5 \pm 2.58 \times 8.6 = (869.4, 913.6)$. Here the population is hypothetical, very large and spread over several hundred years. Sample is assumed random though strictly it is not.

The ISMR (1901-2000) is distributed with mean 891.5 mm and SD 85.6 mm. The ISMR values for the 12 year period 2001-12 are given in the following Table (unit mm).

i) Derive the mean of ISMR for the period 2001-12.

ii) Use *t*-test to test the significance of difference between the mean of (i) with the population mean.

iii) Repeat (i) presuming that population SD is not known.

<table>
<thead>
<tr>
<th>Year</th>
<th>Rain (mm)</th>
<th>Year</th>
<th>Rain (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>822.0</td>
<td>2007</td>
<td>942.3</td>
</tr>
<tr>
<td>2002</td>
<td>720.3</td>
<td>2008</td>
<td>876.3</td>
</tr>
<tr>
<td>2003</td>
<td>912.0</td>
<td>2009</td>
<td>697.2</td>
</tr>
<tr>
<td>2004</td>
<td>768.5</td>
<td>2010</td>
<td>909.3</td>
</tr>
<tr>
<td>2005</td>
<td>879.9</td>
<td>2011</td>
<td>909.3</td>
</tr>
<tr>
<td>2006</td>
<td>887.9</td>
<td>2012</td>
<td>829.1</td>
</tr>
</tbody>
</table>
Solution:

i) Mean of 2001-12 i.e, $\bar{x} = 846.2$ mm

ii) The population SD $\sigma$ is given and is 85.6 mm. $SE = \frac{\sigma}{\sqrt{n}} = \frac{85.6}{\sqrt{12}} = 24.7$. As this is a small sample, $t$-test is to be applied. Compute $y = \frac{\mu - \bar{x}}{SE} = \frac{891.5 - 846.2}{24.7} = 1.83$

DF is 11 and theoretical values of $t$ are 2.201 at 5 % LS, 3.106 at 1 % LS (Table 1). Thus we conclude that the difference between $\mu$ and $\bar{x}$ which though constitutes 5 % of $\mu$ is still not significant.

iii) If $\sigma$ is not known, we have to take $s = \sigma$, but $s$ to be computed by the formula as given in eq. (8). $s$ is obtained as 79.9. $SE$ now is $\frac{s}{\sqrt{n}} = 23.1$. Again difference is not significant at both 5 % and 1 % LS.

The conclusion from the above exercise could be stated as under. Though the ISMR has been subdued during 2001-12 with 3 failures in 2002, 04 and 09 and without any excess rainfall years, the deficit has been within sampling limits. It need not be taken that ISMR has decreased during 2001-2012 (compared to 1901-2000).

Ex: ISMR rainfall of Konkan & Goa (KG) and Kerala. For the period 1901-1980 the statistics are as under:

<table>
<thead>
<tr>
<th>Subdivision</th>
<th>Mean rainfall (mm)</th>
<th>SD (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>2766.9</td>
<td>464.0</td>
</tr>
<tr>
<td>Kerala</td>
<td>2062.3</td>
<td>390.6</td>
</tr>
</tbody>
</table>

Examine the hypothesis that both the above series have come from the same population.

Solution: The null hypothesis is that both the ISMR series of KG and Kerala belong to the same population. Now

$\bar{x}_1 = 2766.9, \ s_1 = 464.0, \ \bar{x}_2 = 2062.3, \ s_2 = 390.6, \ n_1 = n_2 = 80$

Thus the samples are large. The samples are not inter-correlated.
\[ SE = \sqrt{\frac{s_1^2 + s_2^2}{n_1}} = 67.8 \]

\[ y = (\bar{x}_1 - \bar{x}_2) / SE = 704.6/67.8 = 10.4 \]

Thus the difference is highly significant. The two rainfall series (i.e., samples) have not come from the same population.

**General exercises**

i) \( \mu = 100 \), \( \sigma = 10 \), \( n = 17 \), \( \bar{x} = 94 \). Test the significance of the difference between \( \mu \) and \( \bar{x} \).

ii) \( n = 50 \), \( r = 0.37 \). Test the significance of the correlation coefficient \( r \). If \( t \) value for 48 degrees of freedom \( (t_{0.05}) \) = 1.96

References:

Gupta S.P. Statistical Methods

Lecture Notes for Integrated Meteorological Training Course by Dr Y.E.A. Raj, 2013

WMO, 1966 Statistical analysis in Meteorology