M.Sc. GEOGRAPHY LAB MANUAL

2nd Semester

Prepared By

Pure & Applied Science Dept.

Geography

MIDNAPORE CITY COLLEGE

MIDNAPORE CITY COLLEGE

DEPARTMENT OF PURE AND APPLIED SCIENCES

GEOGRAPHY LAB MANUAL

(MA/MSC, SEMESTER – II)

CONTENTS

COURSE NO.	UNIT	PAGE NO.
GEO 295: STATISTICAL TECHNIQUES	GEO 295.1: BASIC STATISTICS IN GEOGRAPHY	1 - 92
	GEO 295.2: ADVANCED QUANTITATIVE METHODS	93 - 127
GEO 296: REMOTE SENSING AND COMPUTER APPLICATION	GEO 296.1: PRINCIPLES OF REMOTE SENSING AND AERIAL PHOTOGRAPHY	128 - 189
	GEO 296.2: COMPUTER BASICS AND APPLICATIONS	190 - 216

GEO 295.1: BASIC STATISTICS IN GEOGRAPHY

1. Measurement in Geography: Nominal, ordinal, interval and ratio measurement.

2. Concept of covariance, correlation and regression: Bivariate analysis - linear, exponential, Product moment correlation, Spearman's Rank correlation, correlation matrix, partial correlation, residuals - mapping of residuals.

3. Probability distribution: addition and Law of multiplication, concept of probability distributions (binomial distributions, normal probability distribution), properties of normal curve.

4. Hypothesis testing: Formulation, Rejection rule, one and two tailed tests, significance level, and degrees of freedom, type I and type II errors, Standard Error. Different types of significance test for various purposes. Chi- square test, student's t- test.

5. Sampling techniques for geographical analysis.

---

1. Measurement in Geography: Nominal, ordinal, interval and ratio measurement.

The field of statistics is the science of learning from data. Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results. Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions. Statistics allows you to understand a subject much more deeply.

Measurement in Geography

Measurement refers to the assignment of numbers in a meaningful way, and understanding measurement scales is important to interpreting the numbers assigned to people, objects, and events.

Measurement scale, in statistical analysis, the type of information provided by numbers. Each of the four scales (i.e., nominal, ordinal, interval, and ratio) provides a different type of information.

Levels of Measurements

There are four different scales of measurement. The data can be defined as being one of the four scales. The four types of scales are:

(1) Nominal Scale
(2) Ordinal Scale
(3) Interval Scale
(4) Ratio Scale

1. Nominal Scale

A nominal scale is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

Characteristics of Nominal Scale

- A nominal scale variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
- It is qualitative. The numbers are used here to identify the objects.
- The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is "counting."

Example:

An example of a nominal scale measurement is given below:

What is your gender?

- M- Male
- F- Female

Here, the variables are used as tags, and the answer to this question should be either M or F.

2. Ordinal Scale

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the "order." Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

Characteristics of the Ordinal Scale

- The ordinal scale shows the relative ranking of the variables
- It identifies and describes the magnitude of a variable
- Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables
- The interval properties are not known
- The surveyors can quickly analyse the degree of agreement concerning the identified order of variables

Example:

- Ranking of school students – 1st, 2nd, 3rd, etc.
- Ratings in restaurants
- Evaluating the frequency of occurrences

- Very often
- Often
- Not often
- Not at all

- Assessing the degree of agreement
  - Totally agree
  - Agree
  - Neutral
  - Disagree
  - Totally disagree

3. Interval Scale

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

Characteristics of Interval Scale:

- The interval scale is quantitative as it can quantify the difference between the values
- It allows calculating the mean and median of the variables
- To understand the difference between the variables, you can subtract the values between the variables
- The interval scale is the preferred scale in Statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.

Example:

(1) Likert Scale
(2) Net Promoter Score (NPS)
(3) Bipolar Matrix Table

4. Ratio Scale

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.

Characteristics of Ratio Scale:

- Ratio scale has a feature of absolute zero
- It doesn't have negative numbers, because of its zero-point feature
- It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, and divided. Mean, median, and mode can be calculated using the ratio scale.
- Ratio scale has unique and useful properties. One such feature is that it allows unit conversions like kilogram – calories, gram – calories, etc.

Example:

An example of a ratio scale is:

What is your weight in Kgs.?

- Less than 55 kgs.
- 55 – 75 kgs.
- 76 – 85 kgs.
- 86 – 95 kgs.
- More than 95 kgs.

---

2. Concept of covariance, correlation and regression: Bi-variate analysis - linear, exponential, Product moment correlation, Spearman's Rank correlation, correlation matrix, partial correlation, residuals - mapping of residuals.

Concept of covariance, correlation and regression

Covariance and Correlation are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

Covariance

It is the relationship between a pair of random variables where change in one variable causes change in another variable.

It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

It is used for the linear relationship between variables.

It gives the direction of relationship between variables.

Formula –

For Population:

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

For Sample

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

Here,
x' and y' = mean of given sample set
n = total no of sample
x_i and y_i = individual sample of set

Example

Correlation

It shows whether and how strongly pairs of variables are related to each other.

Correlation takes values between -1 to +1, wherein values close to +1 represent strong positive correlation and values close to -1 represents strong negative correlation.

In this variable are indirectly related to each other.

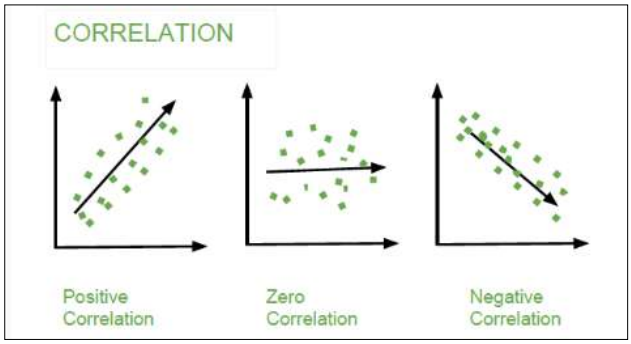
It gives the direction and s

Formula

$$Corr(x, y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

Here,
x' and y' = mean of given sample set
n = total no of sample
x_i and y_i = individual sample of set

Example



Difference between Covariance versus Correlation

Covariance	Correlation
Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.
Involve the relationship between two variables or data sets	involve the relationship between multiple variables as well
Lie between -infinity and +infinity	Lie between -1 and +1
Measure of correlation	Scaled version of covariance
provide direction of relationship	provide direction and strength of relationship
dependent on scale of variable	independent on scale of variable

have dimensions	dimensionless

Coefficient of Correlation

Population Correlation Coefficient:

1. The measure of joint or mutual variation in a bivariate population with two variables x and y, is called 'covariance of x and y':

$$\sigma_{xy} = \frac{\sum(x - \mu_x)(y - \mu_y)}{N}$$

2. In order to make comparison, the covariance must be standardised by dividing $(x - \mu_x)$ and $(y - \mu_y)$ by their SDs σ_x and σ_y respectively. This expression is called 'coefficient of correlation'; the 'population coefficient of correlation' is denoted by 'ρ' (rho):

$$\rho = \frac{\sum\left(\frac{x - \mu_x}{\sigma_x}\right)\cdot\left(\frac{y - \mu_y}{\sigma_y}\right)}{N}$$
$$= \frac{1}{\sigma_x \cdot \sigma_y} \cdot \frac{\sum(x - \mu_x)(y - \mu_y)}{N}$$
$$= \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Sample Correlation Coefficient:

1. The sample covariance of x and y, S_{xy}, measures the tendency for x and y to increase or decrease together in the sample:

$$S_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

2. The 'sample coefficient of correlation' is denoted by 'r'. It is also known as 'Karl Pearson's product moment coefficient of correlation'. The coefficient of correlation always lies between –1 and +1 respectively, i.e., $-1 \le r \le +1$:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

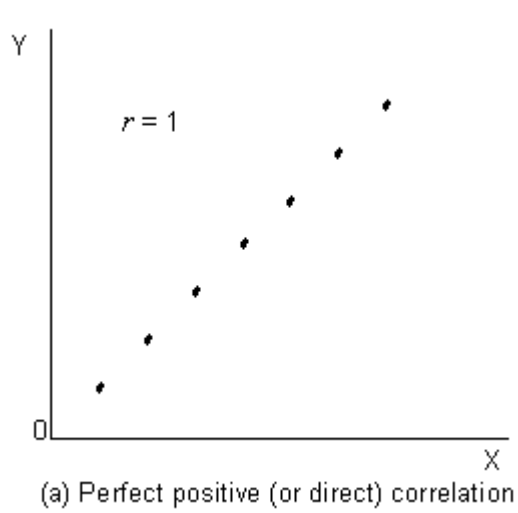
$$r = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \cdot \sqrt{n \cdot \sum y^2 - (\sum y)^2}}$$

3. (a) If r = –1, all the points on the scatter diagram lie on the regression line of negative slope. It is called a 'perfect negative correlation'.

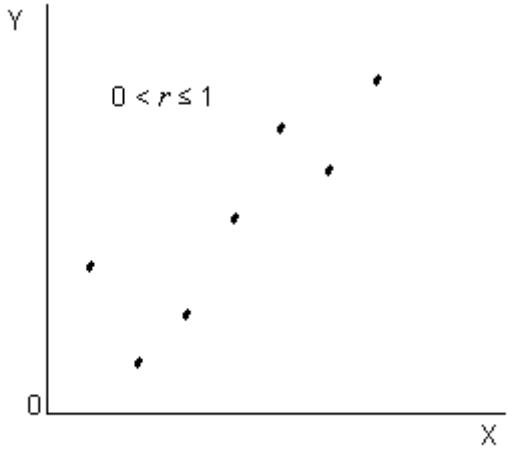
(b) If r = 1, all the points on the scatter diagram lie on the regression line of positive slope. It is called a 'perfect positive correlation'.

(c) If r = 0, all the points on the scatter diagram are spread throughout the diagram indicating no correlation between x and y.

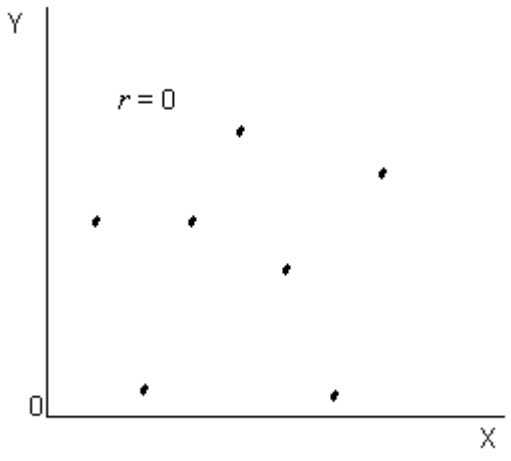
"Correlation coefficient is a measure of the closeness of linear relationship between the two variables."



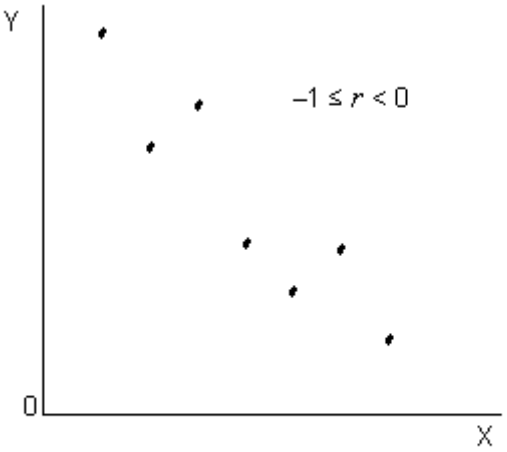
(a) Perfect positive (or direct) correlation



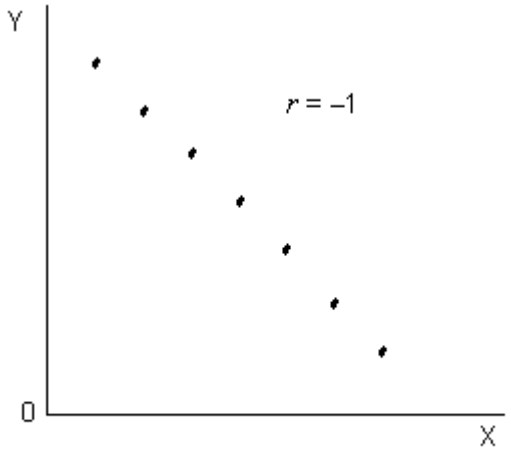
(b) Positive Correlation



(c) No correlation



(d) Negative Correlation



(e) Perfect negative (or inverse) correlation

Correlation Coefficient and Regression Coefficient:

1. The two regression coefficients b and d of the two regression lines can also be stated as follows:

$$b = \frac{S_{xy}}{S_x^2} \qquad \text{and} \qquad d = \frac{S_{xy}}{S_y^2}$$

2. Since $r = \frac{S_{xy}}{S_x \cdot S_y}$, therefore, $S_{xy} = r \cdot S_x \cdot S_y$.
3. The regression coefficients b and are related to correlation coefficient r by:

$$b = r \cdot \frac{S_y}{S_x} \qquad \text{and} \qquad d = r \cdot \frac{S_x}{S_y}$$

or

$$r = b \cdot \frac{S_x}{S_y} \qquad \text{and} \qquad r = d \cdot \frac{S_y}{S_x}$$

or

$$r = \sqrt{b \cdot d}$$

Where $S_x^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{n \cdot \sum x^2 - (\sum x)^2}{n(n-1)}$

$$S_y^2 = \frac{\sum(y-\bar{y})^2}{n-1} = \frac{n\cdot\sum y^2 - (\sum y)^2}{n(n-1)}$$

$$S_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{n\cdot\sum xy - \sum x\cdot\sum y}{n(n-1)}$$

Properties of Coefficient of Correlation:

1. The correlation coefficient is symmetrical with respect to x and y, i.e., $r_{xy} = r_{yx}$
2. The correlation coefficient is the geometric mean of the two regression coefficients, i.e.: $r = \sqrt{b \times d}$.
3. The correlation coefficient is a pure number and does not depend upon the units employed. For e.g., if the correlation coefficient between the heights and weights of students is computed as 0.98, it will be expressed simply as 0.98 (neither as 0.98 inches nor 0.98 pounds).
4. The correlation coefficient is independent of origin and unit of measurement. By this we mean that if we take deviations of x and y from some suitable origins or transform x and y into u and v respectively, it will not affect the correlation coefficient. Symbolically:

$$r_{xy} = r_{uv}$$

5. The correlation coefficient lies between –1 and +1, i.e., it cannot be less than –1 and greater than +1:

$$-1 \le r \le +1$$

Example:

x	3	1	1	2	4	2	3	5	2	3
y	2	4	3	2	1	2	1	3	2	1

Required:

(a) Covariance of x and y,

(b) Standard deviation of x and y,

(c) Coefficient of correlation, and

(d) Scatter diagram.

Solution:

(a) Covariance of x and y:

x	y	$x - \mu_x$	$x - \mu_y$	$(x - \mu_x)(x - \mu_y)$	$(x - \mu_x)^2$	$(x - \mu_y)^2$
3	2	0.4	–0.1	–0.04	9	4
1	4	–1.6	1.9	–3.04	1	16
1	3	–1.6	0.9	–1.44	1	9
2	2	–0.6	–0.1	0.06	4	4
4	1	1.4	–1.1	–1.54	16	1
2	2	–0.6	–0.1	0.06	4	4
3	1	0.4	–1.1	–0.44	9	1
5	3	2.4	0.9	2.16	25	9
2	2	–0.6	–0.1	0.06	4	4
3	1	0.4	–1.1	–0.44	9	1
26	21			–4.6	82	53

$$\mu_x = \frac{26}{10} = 2.6$$

$$\mu_y = \frac{21}{10} = 2.1$$

$$\sigma_{xy} = \frac{\sum(x - \mu_x)(y - \mu_y)}{N}$$

$$= \frac{-4.6}{10} = -0.46$$

(b) Standard deviation of x and y:

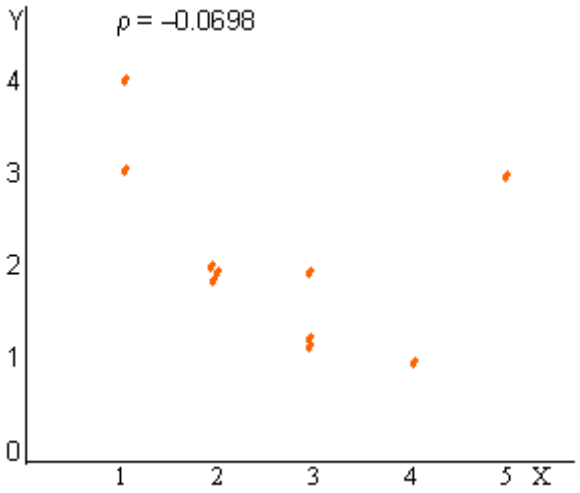
$$\sigma_x = \sqrt{\frac{\sum(x - \mu_x)^2}{N}} = \sqrt{\frac{82}{10}} = \sqrt{8.2} = 2.864$$

$$\sigma_y = \sqrt{\frac{\sum(y - \mu_y)^2}{N}} = \sqrt{\frac{53}{10}} = \sqrt{5.3} = 2.302$$

(c) Coefficient of correlation:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{-0.46}{(2.864)(2.302)} = -0.0698$$

(d) Scatter diagram:



Example:

Calculate:

(a) Covariance of x and y,

(b) Variances of x and y,

(c) Coefficient of correlation, and

(d) Coefficient of determination.

For the following sample data:

x	1	2	4	6	8	10	14	15	18	20
y	10	20	30	40	50	60	70	80	90	100

Solution:

(a) Covariance of x and y:

x	**y**	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
1	10	–8.8	–45	396	77.44	2025
2	20	–7.8	–35	273	60.84	1225
4	30	–5.8	–25	145	33.64	625
6	40	–3.8	–15	57	14.44	225
8	50	–1.8	–5	9	3.24	25

10	60	0.2	5	1	0.04	25
14	70	4.2	15	63	17.64	225
15	80	5.2	25	130	27.04	625
18	90	8.2	35	287	67.24	1225
20	100	10.2	45	459	104.04	2025
98	550			1820	405.6	8250

$$\bar{x} = \frac{\sum x}{n} = \frac{98}{10} = 9.8$$
$$\bar{y} = \frac{\sum y}{n} = \frac{550}{10} = 55$$

$$S_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$
$$= \frac{1820}{10 - 1} = 202.22$$

(b) Variances of x and y:

$$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{405.6}{9}} = \sqrt{45.07} = 6.713$$
$$S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} = \sqrt{\frac{8250}{9}} = \sqrt{916.67} = 30.277$$

(c) Coefficient of correlation:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$
$$= \frac{202.22}{(6.713)(30.277)} = 0.9949$$

(d) Coefficient of determination:

$$b = r \cdot \frac{S_y}{S_x} \quad \text{and} \quad d = r \cdot \frac{S_x}{S_y}$$
$$b = 0.9949 \times \frac{30.277}{6.713} \quad and \quad d = 0.9949 \times \frac{6.713}{30.277}$$
$$b = 4.48720 \quad \text{and} \quad d = 0.22059$$

$r^2 = b \times d$

$r^2 = 4.48720 \times 0.22059 = 0.9898 = 98.98\%$

Probable Error:

1. The probable error is about two-third of the standard error:

$$P.E. = \frac{2}{3}(S.E.) \qquad \text{or} \qquad 3(P.E.) = 2(S.E.)$$

2. Assuming ρ = 0, the sampling distribution of r has standard error:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}$$

3. In a standard normal distribution, z = ± 0.6745 will contain 50% of the area under curve, symbolically:

$$P\ (-0.6745 \le z \le 0.6745) = 0.5$$

4. Thus, the probable error r is:

$$P.E. = 0.6745 \times \sigma_r$$

or

$$P.E. = 0.6745 \times \sqrt{\frac{1-r^2}{n-2}}$$

5. Probabilities of r can now be calculated using P.E. as a unit of deviation:

$$P\ (-P.E. \le r \le P.E.) = 0.5$$

$$P\ (-3P.E. \le r \le 3P.E.) = 0.9544$$

Rank Correlation:

1. If observations on two variables are given in the form of ranks rather than some numerical measurements, it is possible to compute a coefficient of correlation between ranks of the two variables. This correlation coefficient is called 'Rank Correlation Coefficient'.
2. As this formula was presented by Spearman in 1904, it is also known as 'Spearman's Rank Correlation Coefficient':

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)}$$

Where $d_i = x_i - y_i$ (the difference between the rankings).

3. In order to test that there is no correlation between the two rankings, critical values of r_s at $\alpha = 0.05$ are given below:

Number of ranks (n)	Critical value (r_s)
5	1.0
6	0.89
7	0.79
8	0.74
9	0.74
10	0.65
20	0.45
25	0.40
50	0.28

Example:

Ranks of 9 students in a class in History (x) and Geography (y) are as follows:

Students	I	II	III	IV	V	VI	VII	VIII	IX
x	1	9	7	4	5	3	8	2	6
y	4	5	6	3	7	2	8	1	9

Calculate Spearman's Rank Correlation Coefficient and test its significance.

Solution:

Students	x	y	d = x – y	d^2
I	1	4	–3	9
II	9	5	4	16
III	7	6	1	1
IV	4	3	1	1
V	5	7	–2	4
VI	3	2	1	1
VII	8	8	0	0
VIII	2	1	1	1
IX	6	9	–3	9
Total	45	45	0	42

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} \qquad \text{Where } d_i = x_i - y_i$$
$$= 1 - \frac{6 \times 42}{9(81 - 1)} = 0.65$$

Critical value of r_s for n = 9 and $\alpha = 0.05$ is 0.74

Since 0.65 is less than the critical value of 0.74, r_s is insignificant.

Regression

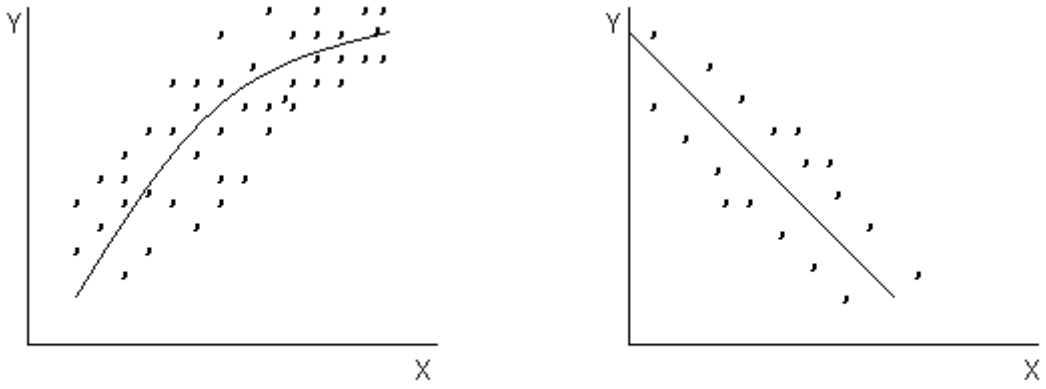
1. The term 'regression' was used by Sir Frances Galton in connection with the studies he made on the statures fathers and sons.
2. It is a technique which determines a relationship between two variables to estimate one of the variables (dependent) for a given value of the other variable (independent).
3. The variable whose value is to be estimated is called dependent variable (y) whereas the variable whose value is given is called independent variable (x).
4. Examples of dependent and independent variables are:

Independent	Dependent
Price	Demand
Rainfall	Yield
Credit sales	Bad debts
Volume of production	Manufacturing expenses

5. The values of the independent variable are assumed to be fixed. Hence it is not a random variable. On the other hand, the dependent variable, whose values are determined on the basis of the independent variable, is a random variable.
6. If x is the independent variable and y is the dependent variable then the relationship between x and y, described by a straight line (y = a + bx), is called 'linear relationship'.

Regression Lines:

1. If we plot the paired observations (X_1Y_1), (X_2Y_2), ……., (X_nY_n) on a graph, the resulting set of points is called a 'scatter diagram'.
2. A scatter diagram indicates a relationship between the variables X and Y and the dots of the scatter diagram tend to cluster around a curve or a line. Such a curve or line is known as 'curve of regression' or 'line of regression'.



(a) Curve of Regression **(b)** Line of Regression

Figure – Regression lines

Linear Regression Model:

1. For a fixed value of independent variable 'x', if the value of dependent variable 'y' is observed a large number of times, different values are possible each time because of the random error involved in the measurement process. The mean of these y-values is called the 'conditional mean of y given x' and is denoted by $\mu_{y/x}$.
2. The linear relationship between $\mu_{y/x}$ and x is called a 'population regression equation of y on x':

$$\mu_{y/x} = \alpha + \beta x$$

Where α and β are the parameters of the equation.

3. An observation y_i is the sum of a population mean $\mu_{y/x}$ and a component called 'Random Error ($\in$)' (read as "epsilon").

$$y_i = \mu_{y/x} + \in$$

or

$$y_i = \alpha + \beta x + \in$$

This equation is called a 'linear regression model of y on x' and $\in$ is the random variable with mean is equal to zero and variance σ_e^2.

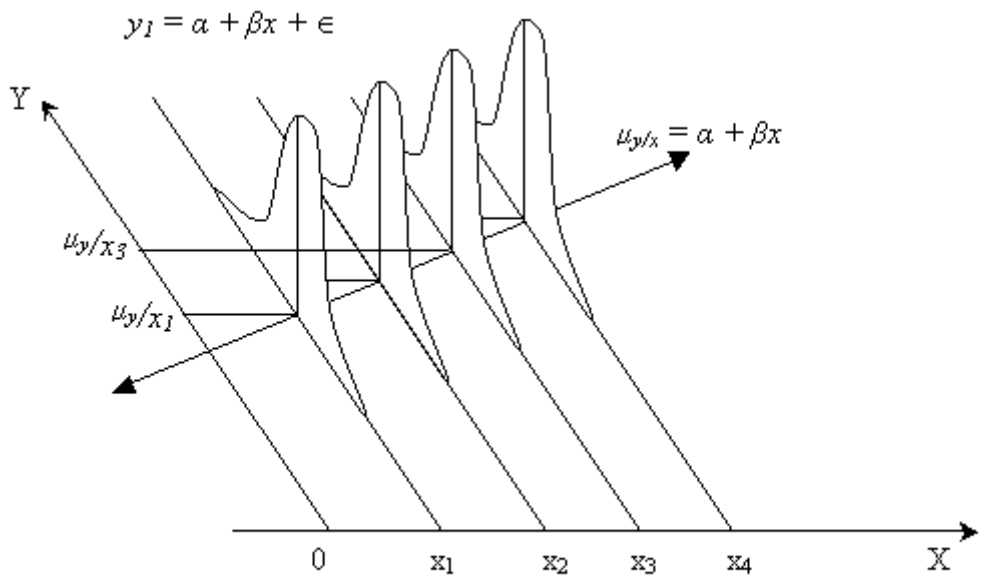


Figure – Linear Regression Model

4. In the above diagram, the line represents the line of regression of *Y* on *X*. The parameter *α*, which is the expected value of *Y* when *X* = 0, is called *Y*-intercept. The parameter *β* is slope of the population regression line and is known as the *'population regression coefficient'*. When the line slopes downward to the right, the value of *β* will be negative; it then represents the amount of decrease in *Y* for each unit increase in *X*.

5. In practice, the population regression line is unknown. Since the regression is defined by the *Y*-intercept *α* and the slope *β*, therefore, the task of estimating the population regression line involves obtaining the estimates of *α* and *β* (based on sample data). Thus the *'population regression line'* ($\mu_{y/x} = \alpha + \beta x$) is estimated by the *'sample regression line'* or *'sample regression equation'*:

$$\hat{y} = a + bx \quad \text{---------- (i)}$$

The problem of estimating the regression parameters α and β can be considered as fitting the best model on the scatter diagram. One method for this purpose is the 'method of least squares'.

Method of Least Squares:

1. According to the principle of least squares, a line or a curve is best fitted if the sum of squares of the deviations of estimated values of y from the observed values of y is minimum. Such line or a curve is called the 'least square curve' or 'least square line'. And the sum of squares is called the 'Error Sum of Squares (ESS)'. Therefore, the ESS is to be minimised and is represented by:

$$\text{ESS} = \sum (y_i - \hat{y})^2$$

Where ESS : Error sum of squares

y_i : observed values

$\hat{y}$: Estimated values, i.e., ($\hat{y} = a + bx$)

It is further elaborated as:

$$ESS = \Sigma(y_i - a - bx)^2$$

2. As we know that the statistic b is an estimator of β, is known as 'sample regression coefficient'. It measures changes in y per unit change in x. Therefore, it represents the slope of regression line. Mathematically it is represented as below:

$$b = \frac{n \cdot \sum xy - (\sum x) \cdot (\sum y)}{n \cdot \sum x^2 - (\sum x)^2}$$ ---------- (ii)(a)

$$b = \frac{S_{xy}}{S_{x^2}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$ ---------- (ii)(b)

3. The statistic a is the estimator of α, is called the 'sample regression constant', and it measures the y-intercept of the sample regression line:

$$a = \bar{y} - b\bar{x}$$ ---------- (iii)

4. Now assume 'y' to be 'independent' and 'x' to be 'dependent'. The 'regression equation of x on y' is as follows:

$$\hat{x} = c + dy$$ ---------- (i)

$$d = \frac{n \cdot \sum xy - (\sum x) \cdot (\sum y)}{n \cdot \sum y^2 - (\sum y)^2}$$ ----------(ii)(a)

$$d = \frac{S_{xy}}{S_{y^2}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}$$ ----------(ii)(b)

$$c = \bar{x} - d\bar{y}$$ ----------(iii)

Example:

A sample of paired observations is given as below:

X	2	4	6	7	9	10	11
Y	1	2	4	7	10	12	14

Required:

(a) Fit a line of regression to the data in the above table.

(b) Construct a scatter diagram and graph the fitted line on the scatter diagram, and

(c) Calculate error sum of squares.

Solution:

(a):

Regression Line of Y on X

x	y	xy	x^2	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
2	1	2	4	–0.438	1.438	2.068
4	2	8	16	2.594	–0.594	0.353
6	4	24	36	5.626	–1.626	2.644
7	7	49	49	7.142	–0.142	0.020
9	10	90	81	10.174	–0.174	0.030
10	12	120	100	11.69	0.31	0.096
11	14	154	121	13.206	0.794	0.630
49	50	447	407	49.994 ≈ 50	0.006 ≈ 0	5.841

$\hat{y} = a + bx$ ---------- (i)

$$b = \frac{n \cdot \sum xy - (\sum x) \cdot (\sum y)}{n \cdot \sum x^2 - (\sum x)^2}$$ ---------- (ii)

$$= \frac{7(447) - (49)(50)}{7(407) - (49)^2} = 1.516$$

$a = \bar{y} - b\bar{x}$ ---------- (iii)

$$= \frac{50}{7} - 1.516 \times \left(\frac{49}{7}\right) = -3.47$$

$$\hat{y} = -3.47 + 1.516x$$

For x = 2, $\hat{y} = -3.47 + 1.516(2) = -0.438$

x = 4, $\hat{y} = -3.47 + 1.516(4) = 2.594$

x = 6, $\hat{y} = -3.47 + 1.516(6) = 5.626$

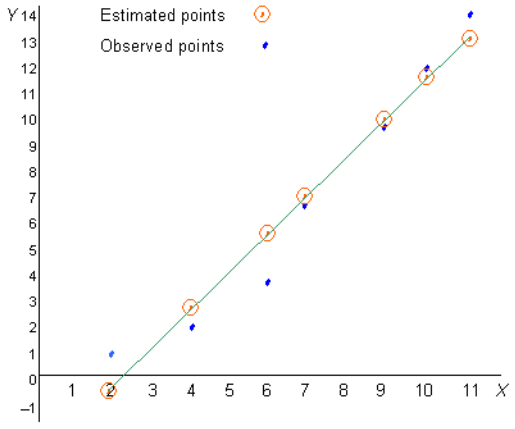
x = 7, $\hat{y} = -3.47 + 1.516(7) = 7.142$

x = 9, $\hat{y} = -3.47 + 1.516(9) = 10.174$

x = 10, $\hat{y} = -3.47 + 1.516(10) = 11.69$

x = 11, $\hat{y} = -3.47 + 1.516(11) = 13.206$

(b):



Scatter Diagram

(c) Error Sum of Squares (ESS):

ESS = $\sum(y_i - \hat{y})^2$

= 5.841

Coefficient of Determination:

1. A measure of variation in a sample of n values is given by the sample variance:

$$S_y^2 = \frac{\sum(y-\bar{y})^2}{n-1} = \frac{n\cdot\sum y^2 - (\sum y)^2}{n(n-1)}$$

It measures the variation in y about the sample mean $\bar{y}$. The term $\sum(y-\bar{y})^2$ is called 'Total Sum of Squares (TSS)'.

2. Another measure of variance in a sample of n paired values is called 'variance of estimate':

$$S_{y/x}^2 = \frac{\sum(y-\bar{y})^2}{n-2}$$

It measures the variation in y about the estimated regression line. The term $\sum(y-\bar{y})^2$ is called the 'Error Sum of Squares (ESS)':

$$ESS \leq TSS$$

3. The 'Regression Sum of Squares (RSS)' is the difference or excess of TSS over ESS:

$$RSS = TSS - ESS$$

Therefore, the TSS is partitioned into two components, i.e., ESS and RSS:

$$TSS = RSS + ESS$$

4. RSS is the variation in y reduced (or explained) by the regression equation and the ESS is the variation which remains (or unexplained) in y when regression line is filled. Thus, the total variation is divided into two, i.e., explained variation and unexplained variation.
5. RSS is used as a measure of reliability of the estimate obtained by the filled regression line. For this purpose the proportion of variation explained by the regression equation, called 'Coefficient of Determination' denoted by r^2, is calculated as:

$$r^2 = \frac{RSS}{TSS} = \frac{TSS-ESS}{TSS} = 1-\frac{ESS}{TSS}$$

$$= \left[1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}\right] \text{ or } \left[\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}\right]$$

Note that the minimum value of r^2 is zero (when RSS = 0 and ESS = TSS), and the maximum value of r^2 is +1 (when RSS = TSS and ESS = 0); therefore, r^2 lies between 0 and 1:

$$0 \leq r^2 \leq 1$$

6. Another formula is:

$$r^2 = \frac{a \cdot \sum y + b \cdot \sum xy - n \cdot \bar{y}^2}{\sum y^2 - n \cdot \bar{y}^2}$$

7. Coefficient of determination of two regression equations:

$$r^2 = b \times d$$

Example:

Take the previous example, and calculate the coefficient of determination.

Solution:

Coefficient of Determination

x	y	xy	x^2	y^2
2	1	2	4	1
4	2	8	16	4
6	4	24	36	16
7	7	49	49	49
9	10	90	81	100
10	12	120	100	144
11	14	154	121	196
49	50	447	407	510

$$\bar{y} = \frac{50}{7} = 7.143$$

$$r^2 = \frac{a \cdot \sum y + b \cdot \sum xy - n \cdot \bar{y}^2}{\sum y^2 - n \cdot \bar{y}^2}$$
$$= \frac{-3.47 \times (50) + 1.516 \times (447) - 7 \times (7.143)^2}{510 - 7 \times (7.143)^2}$$
$$= \frac{-173.5 + 677.652 - 357.157}{510 - 357.157}$$
$$= \frac{146.995}{152.843}$$
$$= 0.9617 \text{ or } 96.17\%$$

Residual analysis

Residual analysis is used to assess the appropriateness of a linear regression model by defining residuals and examining the residual plot graphs.

Residual

Residual (*e*) refers to the difference between observed value (y) vs. predicted value ($\hat{y}$). Every data point has one residual.

$$residual = observedValue - predictedValue$$

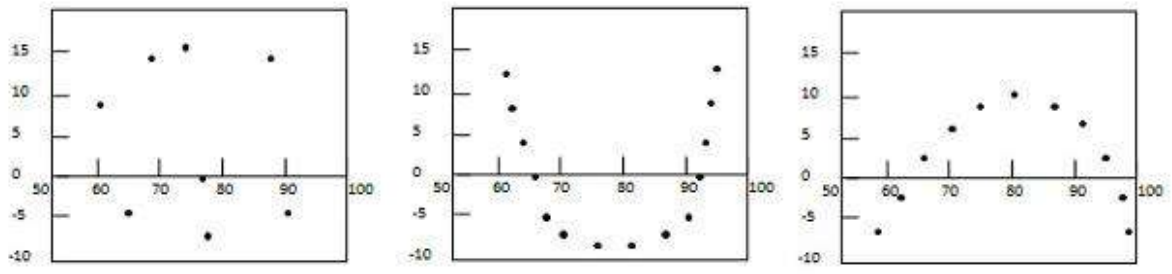
$$e = y - \hat{y}$$

Residual Plot

A residual plot is a graph in which residuals are on tthe vertical axis and the independent variable is on the horizontal axis. If the dots are randomly dispersed around the horizontal axis then a linear regression model is appropriate for the data; otherwise, choose a non-linear model.

Types of Residual Plot

Following example shows few patterns in residual plots.



I II III

In first case, dots are randomly dispersed. So, linear regression model is preferred. In Second and third case, dots are non-randomly dispersed and suggests that a non-linear regression method is preferred.

Example

Problem Statement:

Check where a linear regression model is appropriate for the following data.

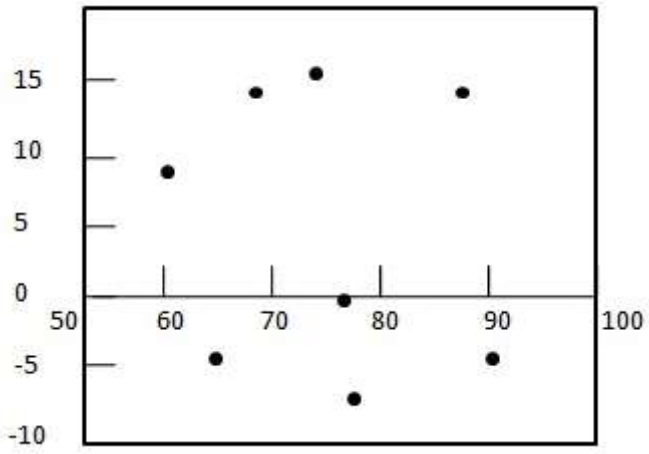
x	60	70	80	85	95
y (Actual Value)	70	65	70	95	85
$\hat{y}$ (Predicted Value)	65.411	71.849	78.288	81.507	87.945

Solution:

Step 1: Compute residuals for each data point.

x	60	70	80	85	95
y (Actual Value)	70	65	70	95	85
$\hat{y}$ (Predicted Value)	65.411	71.849	78.288	81.507	87.945
e (Residual)	4.589	-6.849	-8.288	13.493	-2.945

Step 2: - Draw the residual plot graph.



Step 3: - Check the randomness of the residuals.

Here residual plot exhibits a random pattern - First residual is positive, following two are negative, the fourth one is positive, and the last residual is negative. As pattern is quite random this indicates that a linear regression model is appropriate for the above data.

3. Probability distribution: addition and Law of multiplication, concept of probability distributions (binomial distributions, normal probability distribution), properties of normal curve.

Probability

Definition:

An experiment is any well-defined operation or procedure that results in one of two or more possible outcomes. An outcome is particular result of an experiment.

Counting Techniques:

(a) Tree Diagram,

(b) Multiplication Rule,

(c) Permutation, and

(d) Combination.

(a) Tree Diagram: Counting the number of possible outcomes of an experiment plays a major role in probability theory. These possible outcomes can be shown by the branches of a tree-like diagram called 'Tree-Diagram' or 'Branch Diagram'.

(b) Multiplication Rule: If an operation of an experiment can be performed in n_1 ways and if for each of these ways another operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 \times n_2$ ways, and the k^{th} in n_k ways, then all the k operations can be performed together in $n_1 \times n_2 \times n_3 \times \ldots\ldots \times n_k$ ways.

For example, a coin and a die tossed together, the possible outcomes will be:

$n_1 = 2$ (coin: two sides)

n_2 = 6 (die: 6 sides)

The two operations can result in ($n_1 \times n_2$) 12 ways.

(c) Permutations: A permutation is a group of items with a certain ordered arrangement. For example: ABC, ACB, CAB, CBA and BCA are different permutations. The rules of permutation are different under each of the following four situations:

Situation I:

(i) All the items are distinct,

(ii) Each item can occur only once in an arrangement, i.e., repetition not allowed or without replacement, and

(iii) Each item can occupy any place in an arrangement.

In the above situation, the number of permutations of n items arranged r at a time, denoted by ${}^nP_r$ is:

$$ {}^nP_r = \frac{n!}{(n-r)!} $$

Where, r ≤ n

Example:

How many three-digit numbers can be formed from the digits 1, 2, 4, 5 and 9 without replacement?

Solution:

n = 5 and r = 3

$$ {}^nP_r = \frac{n!}{(n-r)!} = \frac{5!}{(5-3)!} = 60 \text{ ways} $$

Situation 2:

(i) All the items are distinct,

(ii) An item can be repeated in an arrangement (i.e., repletion allowed or with replacement), and

(iii) Each item can occupy any place in an arrangement.

In the above situation, the number of permutations of n items arranged r at a time is:

$$^{n}P_{r} = (n)^{r}$$

Example:

Arrange the license plate with 3 alphabets and 3 digits with replacement.

Solution:

$$n_1 = {}^{n}P_{r} = {}^{26}P_{3} = 26^{3} = 17576$$
$$n_2 = {}^{n}P_{r} = {}^{10}P_{3} = 10^{3} = 1000$$
$$n_1 \times n_2 = 17576 \times 1000 = 17576000 \text{ ways}$$

Situation 3:

For n non-distinct items out of which n_1 are of one kind, n_2 are of another kind, ……, n_k are of another, and $n_1 + n_2 + \ldots\ldots\ldots\ldots + n_k = n$, the number of permutations of all n items is:

$$^{n}P_{n_1, n_2, n_3, \ldots\ldots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot n_3! \cdot \ldots\ldots\ldots \cdot n_k!}$$

Example:

Find the possible permutations of 7558.

Solution:

n = 4 (7558),

n_1 = 1 (one 7),

n_2 = 2 (two 5), and

$n_3 = 1$ (one 8).

$$^4P_{1,2,1} = \frac{4!}{1! \cdot 4!\ 1!} = 12 \text{ ways}$$

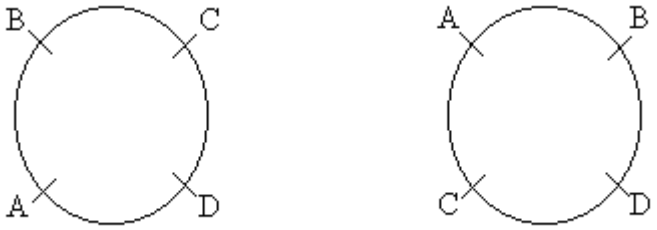
Situation 4:

When the items are arranged in a circle, two arrangements are not considered different, unless corresponding items of both are preceded or followed by a different item.

The number of permutations of n distinct items arranged in a circle is:

$$= (n - 1)!$$

The above lines have different permutations:



The above circles have same permutations.

Example:

Arrange 5 different trees in a circle:

Solution:

n = 5 trees

= (n – 1)!

= (5 – 1)! = 4! = 24 ways

(d) Combinations: A combination is a group of items without regard to the arrangement of items. ABC and BCA are two different permutations but are same combinations of three letters.

The number of combination of n distinct items taken r at a time, denoted by ${}^{n}C_{r}$, is:

$$ {}^{n}C_{r} = \frac{n!}{r! \cdot (n-r)!} $$

Example:

In how many ways five students can be selected from a group of 12 students?

Solution:

n = 12; r = 5

$$ {}^{n}C_{r} = {}^{12}C_{5} = \frac{12!}{5!\,(12-5)!} = 792 \text{ ways} $$

Basic Concepts of Probability Theory:

(a) Possibility Space,

(b) Event,

(c) Complementary Event,

(d) Mutually Exclusive Events,

(e) Composite Events, and

(f) Joint Events.

(a) Possibility Space: Also known as 'sample space' or 'outcome space'. A possibility space is a set of all possible outcomes of an experiment and is denoted by S.

If an experiment consists of a toss of a fair die and the numbers are of interest, the possibility space would be:

S = {1, 2, 3, 4, 5, 6}

If the interest is whether the number is even or odd, the possibility space would be:

S = {even, odd}

A possibility space may be represented by a rectangle. The number of outcomes in the possibility space is denoted by n(S).

(b) **Event:** A subset of a possibility space is called an event and is usually denoted by first few capital letters A, B, C, …… For example, a coin is tossed, the sample space is S = {H, T} and the subset A = {H} is the event when a head occurs.

Take another example, two coins are tossed. The sample space is S = {HH, HT, TH, TT} and the subset B = {HH, HT, TH} is the event that at least one head appears when two coins are tossed.

An event may be represented by a circle inside the rectangle of the possibility space.

An event is further divided into the following:

(i) **Simple Event:** is the subset if it contains only one outcome of the possibility space.

(ii) **Compound Event:** is the subset if it contains more than one outcomes of the possibility space.

(iii) **Null Event:** is a subset containing no outcomes. It is also called 'impossible event'.

(iv) **Sure Event:** It is also called 'certain event'. It is a subset containing all outcomes of the possibility space.

Number of possible events = 2^n

Example:

Two coins are tossed. List all the possible events or subsets of the possibility space.

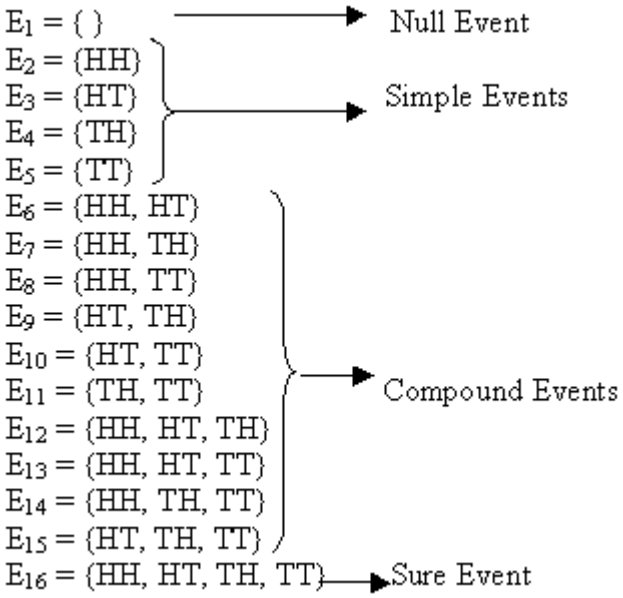
Solution:

S = {HH, HT, TH, TT}

n (S) = 4

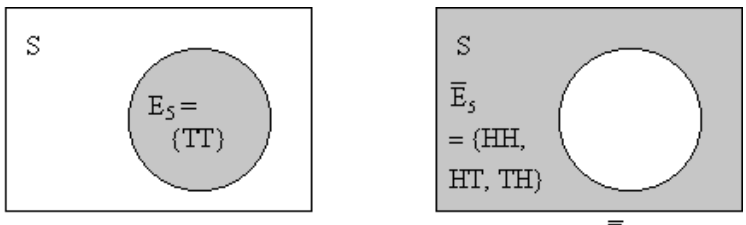
Number of possible events = $2^n = 2^4 = 16$ events

Possible Events of two tossed coins:



(c) Complementary Event: For an event A, the complementary event is defined as a set of those outcomes of the possibility space which are not in A. The complementary event of A is written as $\overline{A}$ (not A).

For example, the complementary events of E_5 = {TT} are $\overline{E}_5 = \{HH, HT, TH\}$. It is diagrammatically shown as below:

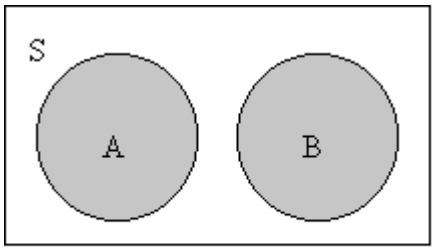


Event E_5 Event $\overline{E}_5$

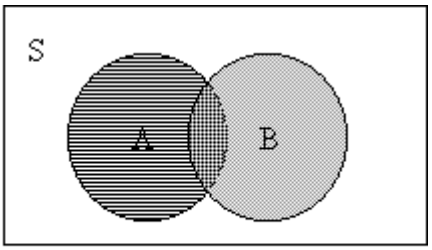
Venn Diagram showing the event E_5 and its complementary event $\overline{E}_5$

(d) **Mutually Exclusive Events:** Two events A and B are mutually exclusive if they have no outcomes in common and therefore they cannot happen together. Mutually exclusive events are also known as disjoint events.

Non-mutually exclusive events are the vice versa of the above definition and are also known as over-lapping events.



Mutually Exclusive Events



Non-Mutually Exclusive Events

For example, in case of two tossed coins, E_6 and E_7 are not mutually exclusive events, because both events have an outcome HH in common. However, E_6 and E_{11} are two mutually exclusive events because no single outcome is common.

(e) **Composite Events:** For two events A and B, a composite event is defined as a set of outcome of either A or B or both A and B and therefore at least one of two events must occur. The composite event of A and B is written as AUB, or A or B.

For example, the composite event of E_6 and E_7 is as follows:

$$E_6 \cup E_7 = \{HH, HT, TH\}$$

(f) **Joint Events:** For two events, A and B, a joint event is defined as a set of common outcomes of A and B and therefore both the events must occur together. The joint event of A and B is written as 'A and B' or 'A∩B'.

For example, the joint event of E_6 and E_7 is as follows:

$$E_6 \cap E_7 = \{HH\}$$

Take another example, the joint event of E_2 and E_{15} is as follows:

$$E_2 \cap E_{15} = \{\ \} \qquad \text{(i.e., Null Event)}$$

Probability Theory:

1. A probability is a numerical measure of the likelihood (or chance) that a particular event will occur.

2. The probability of any event must satisfy the following two conditions.

   (i) No probability is negative, P(Event) ≥ 0.

   (ii) No probability is greater than one, P(Event) ≤ 1.

3. There are three different approaches to assign probabilities to the events:

   (a) Classical or Mathematical Approach

   (b) Empirical or Relative Frequency Approach

   (c) Subjective Approach

(a) Classical Approach: It is the approach in which probabilities are assigned to the events before the experiment is actually performed and therefore, such probabilities are also called 'a priori' probabilities.

If the possibility space of the experiment is finite, and if each outcome of the possibility space is equally likely to occur, then the probability of event A:

$$= \frac{\text{number of outcomes in the event A}}{\text{number of outcomes in the possibility space S}}$$

$$P(A) = \frac{n(A)}{n(S)}$$

It is also referred to as 'axiomatic function of probability'.

Example:

Two coins are tossed once, what is the probability that two heads will appear?

Solution:

S = {HH, HT, TH, TT}

n (S) = 4

Event = Two head appear

i.e., n(A) = 1

Now the probability of event A is as calculated below:

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{4} = 0.25$$

(b) **Relative Frequency Approach:** This approach is applied when the possibility spaces are infinite, or the outcomes cannot be assumed equally likely.

If an experiment is represented 'n' times under uniform conditions and if 'm' times the outcome of the experiment is in favour of an event A, then the ratio $\left(\frac{m}{n}\right)$ approaches the probability of the event A as 'n' approaches infinity.

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

The ratio $\left(\frac{m}{n}\right)$ is considered as an estimate of the actual probability of event A and normally this estimate is called the probability of the event A and is written as:

$$P(A) = \frac{m}{n}$$

Since in this approach probabilities are assigned after performing a large experiment, therefore, they are also known as 'a posteriori' probabilities.

Example:

A die has been rolled 360 times and 'Six' has been observed 63 times. Estimate the probability of occurring a 'Six' when the die is to be rolled once again.

Solution:

m = 63

n = 360

$$P(A) = \frac{m}{n} = \frac{63}{360} = 0.175$$

(c) **Subjective Approach of Probability:** Subjective probability can be defined as the probability assigned to an event by an individual, based on whatsoever evidence is available. This evidence may be in the form of relative frequency of past occurrences, or it may be just an educated guess.

Probability of Complementary Events:

If A and $\overline{A}$ are complementary events in a probability S, then:

$$P(\overline{A}) = 1 - P(A)$$

Example:

A coin is tossed 3 times. Find the probability of getting at least one tail.

Solution:

Number of possible outcomes in S = n(S) = $2^3 = 8$

Let A = At least one tail appears

Then $\overline{A}$ = No tail appears = All heads appear

$n(\overline{A}) = 1$ (the only outcome with all heads)

$$P(\overline{A}) = \frac{n(\overline{A})}{n(S)} = \frac{1}{8}$$

Since $P(\overline{A}) = 1 - P(A)$

or $P(A) = 1 - P(\overline{A})$

therefore $P(A) = 1 - \frac{1}{8} = \frac{7}{8} = 0.875$

Independent and Dependent Events:

1. When two events are given, the occurrence of the first event may or may not have an effect on the occurrence of the second event.

2. When the occurrence of one of the two events has no effect on the probability of the occurrence of the other event, the two events are called 'Independent Events'.

3. On the other hand, when the occurrence of first event has some effect on the probability of occurrence of second event, the second event is said to be 'Dependant' on the first event.

Conditional Probability:

If there are two events A and B such that the probability of event B depends on the occurrence or non-occurrence of the event A, then the probability of event B occurs when the event A occurs is called the 'conditions probability of event B given event A' and is written as:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \quad \text{(dependent event)}$$

and

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

If two events are independent then:

P (B/A) = P (B) (independent events)

Example:

A black card is drawn from an ordinary deck of 52 playing cards. What is the probability that it is of spade (♠) suit?

Solution:

Let B = Black card drawn

and S = Spade card drawn

Since the card drawn is black (event B has occurred), and there are 13 spade cards in 26 black cards, therefore:

$$P(S/B) = \frac{13}{26} = 0.5$$

Multiplication Law of Probability:

If there are two non-mutually exclusive events A and B such that event B is dependent on event A, then the probability of joint event (A and B) is given by:

$$P(A \text{ and } B) = P(A) \times P(B/A)$$

or

$$P(A \cap B) = P(A) \times P(B/A) \text{ or } P(A) + P(B) - P(AUB)$$

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) \text{ or } 1 - P(A \cup B)$$

or it may equivalently be written as:

$$P(A \cap B) = P(B) \times P(A/B)$$

When two events A and B are independent:

$$P(A \cap B) = P(A) \times P(B)$$

When two events A and B are mutually exclusive, their joint probability is zero:

$$P(A \cap B) = 0$$

Example:

In a graduate college, there are 500 male and female students learning B.Sc and B.Com. The break up is as follows:

	Male	Female	Total
B.Sc	70	150	220
B.Com	120	160	280

Total	190	310	500

What is the probability that a randomly selected student is (i) a female B.Sc student, and (ii) a male B.Com student?

Solution:

(i) a female B.Sc student:

Let A = student selected is learning B.Sc; $P(A) = \frac{220}{500}$

B = student selected is a female; $P(B) = \frac{310}{500}$

A∩B = female student learning B.Sc

P(A∩B) = P(A) × P(B/A), where $P(B/A) = \frac{150}{220}$

$$= \frac{220}{500} \times \frac{150}{220} = \frac{150}{500} = \frac{3}{10} = 0.3$$

(ii) a male B.Com student:

Let $\overline{A}$ = student selected is not learning B.Sc; $P(\overline{A}) = 1 - P(A) = 1 - \frac{220}{500} = \frac{280}{500}$

$\overline{B}$ = student selected is not a female; $P(\overline{B}) = 1 - P(B) = 1 - \frac{310}{500} = \frac{190}{500}$

$\overline{A} \cap \overline{B}$ = student selected is learning B.Com and a male

$P(\overline{A} \cap \overline{B}) = P(\overline{A}) \times P(\overline{B}/\overline{A})$, where $P(\overline{B}/\overline{A}) = \frac{120}{280}$

$$= \frac{280}{500} \times \frac{120}{280} = \frac{120}{500} = 0.24$$

Addition Law of Probability:

If A and B are two non-mutually exclusive events, then the probability of the composite event (A or B) is given by:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

or

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\overline{A} \cup \overline{B}) = 1 - P(A \cap B)$$

For two mutually exclusive events A and B:

$$P(A \cup B) = P(A) + P(B)$$

Example:

A card is drawn from a pack of 52 cards. What is the probability that:

(i) it is either Ace or King

(ii) it is either Queen or a Diamond (♦)

Solution:

(i) it is either Ace or King:

$$\left.\begin{array}{l} \text{Let A} = \text{Ace};\ P(A) = \dfrac{4}{52} \\ \text{B} = \text{King};\ P(B) = \dfrac{4}{52} \end{array}\right\}$$

A and B are mutually exclusive events, as the card cannot be an Ace and a King at the same time.

(AUB) = the card drawn is either an Ace or a King

PAUB) = P(A) + P(B)

$$= \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13} = 0.154$$

(ii) it is either Queen or a Diamond:

$$\left.\begin{array}{l}\text{Let C} = \text{Queen; } P(C) = \dfrac{4}{52} \\ \text{D} = \text{Diamond; } P(D) = \dfrac{13}{52}\end{array}\right\}$$ C and D are not mutually exclusive events

CUD = Card drawn is either a Queen or a Diamond

P(CUD) = P(C) + P(D) – P(C∩D); where P(C∩D) = P(C) × P(D/C) = $\frac{4}{52} \times \frac{1}{4} = \frac{1}{52}$

P(CUD) $= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} = 0.308$

Baye's Theorem:

1. The Baye's Theorem is based on conditions probabilities. It calculates probabilities of the causes that may have produced an observed event.
2. Given A_1, A_2, ……, A_i, ……., A_n mutually exclusive events, whose union is the entire possibility space S, and let B an arbitrary event in S, such that P(B) ≠ O, then:

$$P(A_i / B) = \frac{P(B \bigcap A_i)}{P(B)}$$
$$= \frac{P(B / A_i) \times P(A)}{\sum P(B / A_i) \times P(A_i)}$$

Where i = 1, 2, 3, 4, ……., n

3. P (A_i) are called 'prior probabilities' and P(A_i/B) are called 'posterior probabilities'. Thus Baye's Theorem is a process to revise the prior probabilities using additional sample information and get the posterior probabilities.

Random Variable and Its Probability Distribution

Random Numbers:

1. In our everyday life, we base many of our decisions on random outcomes, i.e., change occurrence. For e.g., captains of two cricket teams toss a coin to decide as to which team will play first, or lotteries are drawn by spinning wheel, etc.

2. Random numbers are the numbers obtained by some random process (manually or mechanically).
3. These numbers are assumed to be randomly and uniformly (equally) distributed. The basic random numbers are the 10 one-digit numbers, i.e., 0, 1, 2, ……. 9. Each of these numbers has an equal change $^1/_{10}$ of being selected.
4. Random numbers can be generated manually as well as mechanically. Random numbers can be generated manually by drawing cards from playing cards or rotating spinning wheel, etc. Mechanically generated random numbers are from calculators and computers.
5. The most common use of random numbers is for selection of samples.

Random Variables:

1. Experiments in which outcomes vary from trial to trial are called 'Random Experiments'.
2. A variable whose values are determined by the outcomes of a random experiment is called a random variable.
3. In other words, random variable is a rule which assigns numbers to the outcomes of the possibility space and is denoted by X.
4. For example, throwing of a die is a random experiment and its outcomes, i.e., the occurrence of 1, 2, 3, 3, 4, 5 and 6 is a random variable.
5. A random variable is also called a 'chance variable', 'stochastic variable' or simply a 'variable'. Capital letters of X or Y are used to denote a variable and lower case letters x or y are used to denote its values.
6. Many random variables may be defined for one and the same possibility space.
7. When any characteristics of the individuals of a population (or a sample) are measured or counted, the characteristic itself is a random variable.
8. The random variables are further bifurcated into:

    (a) Discrete Random Variable, and

    (b) Continuous Random Variable.

(a) Discrete Random Variable: A random variable which can assume only a finite number of values or a sequence of whole numbers is called a discrete random variable. For example, the number of spots on a die is a discrete random variable, number of persons enrolled for CSS examinations, number of students passed in 1st division in a particular class, number of defective items in a lot, etc. are discrete random variables, which could assume any of the possible values, i.e., 1, 2, 3…….

(b) Continuous Random Variable: A random variable which can assume all possible values on a continuous scale in a given interval is called a continuous random variable. For example, height, weight, temperature, distance, life periods, speed, etc. are continuous random variables.

Example:

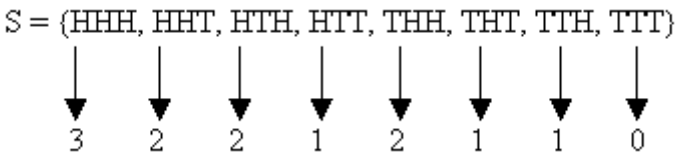
A coin is tossed three times. Find the possibility space and define two random variables for this possibility space.

Solution:

S = {HHH, HHT, HTH, THH, HTT, TTH, THT, TTT}

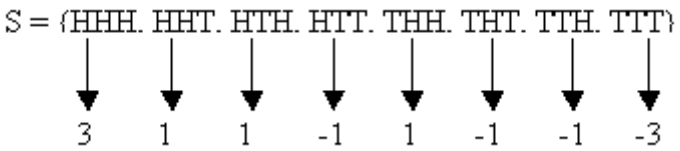
(i) Let a random variable (X) the number of heads:

X = no. of heads.



Note: The same value may be assigned to different outcomes of the possibility space.

(ii) Let a random variable (X) head as +1 and tail as –1:



Probability Distribution:

1. An arrangement of all possible values of a random variable along with their respective probabilities is called a 'probability distribution' or a 'probability function'.
2. Probability distribution can be further bifurcated into:

   (a) Discrete Probability Distribution, and

   (b) Continuous Probability Distribution.

(a) Discrete Probability Distribution: Let a discrete random variable X assume values x_1, x_2, x_3, ……., x_n with respective probabilities $P(x_1)$, $P(x_2)$, $P(x_3)$, ……., $P(x_n)$. Since the random variable takes a discrete set of values, it is also called a discrete probability distribution. A discrete probability distribution may take the form of a table, a graph or a mathematical equation.

A probability distribution is similar to a relative frequency distribution with probabilities replacing relative frequencies.

A discrete probability distribution must possess the following two properties:

(i) $0 \leq P(x_i) \leq 1$

(ii) $\sum P(x_i) = 1$, which means that the sum of probabilities is equal to one.

Example:

A coin is tossed three times. Find the probability distribution of the random variable number of heads.

Solution:

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

No. of Heads X	Probability of X P(X)
0	$^1/_8$
1	$^3/_8$
2	$^3/_8$
3	$^1/_8$
Total	1

Example:

Determine whether the function $P(X) = \frac{x+1}{14}$ for X = 1, 2, 3 and 4 can be a probability distribution.

Solution:

X	P(X)
1	$^2/_{14}$
2	$^3/_{14}$
3	$^4/_{14}$
4	$^5/_{14}$
Total	1

(b) Continuous Probability Distribution: As we know that a random variable which can assume all possible values within a given interval is called a continuous random variable. Within a

given interval, there are an infinite number of values. For example, there may be an infinite number of weights between 69.5 kgs and 70.5 kgs. In case of a continuous random variable, therefore, we compute probabilities for various intervals of continuous random variable, such as P ($a \le X \le b$) or P ($X \ge c$).

The probability distribution of a continuous random variable cannot be presented in tabular form. It can be represented by means of a formula or through a graph. The formula is necessarily in the form of a function of the numerical values of the continuous random variable X. For e.g., a continuous random variable can assume values between X = 2 and X = 4 and the function are given by:

$$f(x) = \frac{x+1}{8} \qquad 2 \le x \le 4$$

The continuous probability distribution is further discussed in detail later.

Mean and Variance of a Random Variable:

In a probability distribution of a random variable X, the mean, also referred to as 'Mathematical Expectation' or 'Expected Value', and the variance are defined as:

$$\mu = E(X) = \Sigma X \cdot P(X)$$

and $\sigma^2 = V(X) = \Sigma X^2 \cdot P(X) - [E(X)]^2$

Distribution Function:

A function showing probabilities that a random variable X has a value less than or equal to x is called the 'cumulative distribution function' or 'distribution function of x'.

Symbolically, the cumulative distribution function, denoted by f(x) is defined as:

$$f(x) = P(x \le x)$$

The cumulative distribution function has the following properties:

(i) $f(-\infty) = 0$ and $f(\infty) = 1$, which means that f(x) is an increasing function ranging from 0 to 1.

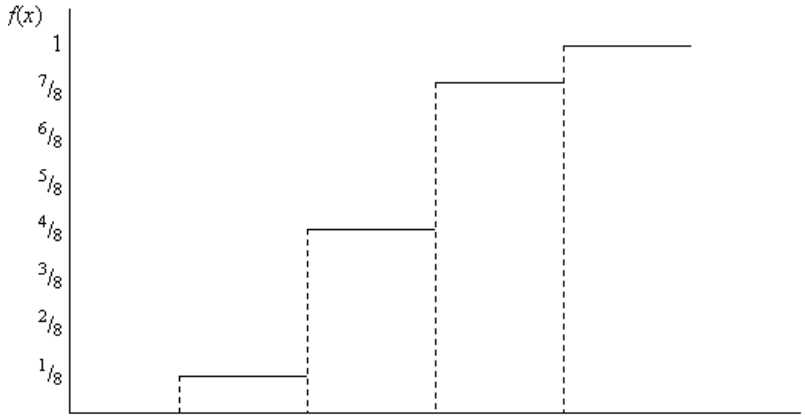
(ii) If a < b then f (a) < f (b) for any real numbers a and b.

For a discrete random variable, distribution function is obtained by cumulating probabilities just as we obtained cumulative distribution.

The distribution function for the probability distribution of the previous two examples is as below:

x	f(x)
x < 0	0
0 ≤ x < 1	$^1/_8$
1 ≤ x < 2	$^4/_8$
2 ≤ x < 3	$^7/_8$
x ≥ 3	1

x	f(x)
x < 1	0
1 ≤ x < 2	$^2/_{14}$
2 ≤ x < 3	$^5/_{14}$
3 ≤ x < 4	$^9/_{14}$
x ≥ 4	1



Example:

Calculate the mean and variance for the following probability distribution:

X	0	1	2	3	4	5	6	7
P(X)	0.11	0.23	0.34	0.16	0.10	0.06	0.04	0.01

Solution:

X	P(X)	XP(X)	$X^2P(X)$
0	0.11	0	0
1	0.23	0.23	0.23

2	0.34	0.68	1.36
3	0.16	0.48	1.44
4	0.10	0.40	1.60
5	0.06	0.30	1.50
6	0.04	0.24	1.44
7	0.01	0.07	0.49
Total	1	2.4	8.06

$\mu = E(X) = \Sigma X \cdot P(X) = 2.4$

$\sigma^2 = V(X) = \Sigma X^2 \cdot P(X) - [E(X)]^2 = 8.06 - (2.4)^2 = 2.3$

Binomial Probability Distribution:

1. Binomial probability is a mathematical formula to determine probabilities of the discrete values of a random variable called 'Binomial Random Variable'.
2. The following are the conditions of Binomial Probability:

   (i) If an experiment contains only two possible outcomes, i.e., success or failure.

   (ii) The probability of 'success' is denoted by 'p' and the probability of 'failure' is denoted by 'q' where q = 1 – p or p + q = 1.

   (iii) Such an experiment is repeated n times independently. In independent repetitions, the probability p remains constant.

3. The number of success in n experiments is the Binomial Random Variable and is denoted by X. The possible values of X are 0, 1, 2, 3, 4, ….., n. The probabilities of the values of X are calculated by the following formula:

$$P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}$$

   Where x = 1, 2, 3, 4, ……, n

   The above formula is 'Binomial Probability Distribution'. The two constant quantities p and n are called the parameters of a Binomial Distribution. The quantity q is not a separate parameter because q = 1 – p.

Mean and Variance of a Binomial Distribution:

The mean and variance of a binomial distribution are directly evaluated in terms of its parameters p and n.

$$E(X) = n \cdot p$$
$$V(X) = n \cdot p \cdot (1-p)$$
$$= n \cdot p \cdot q$$

Example:

A coin is tossed 3 times. 'Number of heads' in 3 tosses is the random variable X. Calculate probabilities of all possible values of X. Also calculate mean and variance.

Solution:

Experiment: A coin is tossed for 3 times.

Success: Head

p = P (success) = P (head) = ½

n = number of times the coin is tossed = 3

x = 0, 1, 2, 3.

Now applying the Binomial Formula:

$$P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}$$

P(x=0) = $P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x} = {}^{3}C_{0} \cdot \left(\frac{1}{2}\right)^{0} \cdot \left(\frac{1}{2}\right)^{3-0} = 1 \times 1 \times 0.125 = 0.125$

P(x=1) = $P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x} = {}^{3}C_{1} \cdot \left(\frac{1}{2}\right)^{1} \cdot \left(\frac{1}{2}\right)^{3-1} = 3 \times 0.5 \times 0.25 = 0.375$

P(x=2) = $P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x} = {}^{3}C_{2} \cdot \left(\frac{1}{2}\right)^{2} \cdot \left(\frac{1}{2}\right)^{3-2} = 3 \times 0.25 \times 0.5 = 0.375$

P(x=3) = $P(x)={}^{n}C_{x}\cdot p^{x}\cdot q^{n-x} = {}^{3}C_{3}\cdot\left(\frac{1}{2}\right)^{3}\cdot\left(\frac{1}{2}\right)^{3-3} = 1\times 0.125\times 1 = 0.125$

Mean and Variance:

X	P(X)	X.P(X)	X^2.P(X)
0	0.125	0	0
1	0.375	0.375	0.375
2	0.375	0.75	1.5
3	0.125	0.375	1.125
Total	1	1.5	3

$$E(X) = n\cdot p = 3\times\frac{1}{2} = 1.5$$

$$V(X) = n\cdot p\cdot(1-p)$$

$$= n\cdot p\cdot q = 3\times\frac{1}{2}\times\frac{1}{2} = 0.75$$

Hyper Geometric Probability Distribution:

1. It is a formula to determine the probabilities of the values for a random variable called 'Hyper Geometric Random Variable'.
2. Following are the conditions of hyper geometric random variable:

   (i) There are N items of which K are of first kind and the remaining (N – K) are of second kind,

   (ii) A sample of n items is randomly drawn without replacement from the N items.

   Number of items of first kind in the sample is the random variable X:

   Possible values of X are 0, 1, 2, ……, k when n ≥ K and

   0, 1, 2, ……., n when n < K

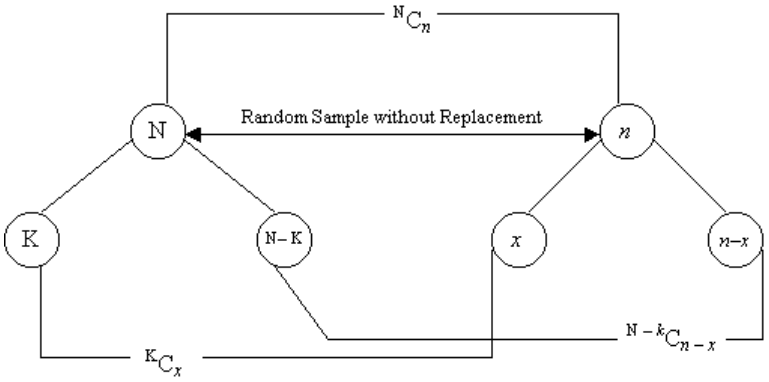
3. The probabilities of these values are calculated by the formula:

$$P(x) = \frac{{}^{k}C_{x}\cdot{}^{N-k}C_{n-x}}{{}^{N}C_{n}}$$

Where x = 0, 1, 2, 3, ……., k when n ≥ k

And x = 0, 1, 2, 3, ……., n when n< k

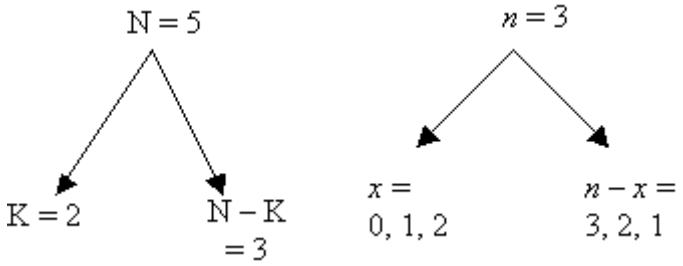
The above formula is called 'Hyper Geometric Probability Distribution'. A schematic explanation of this formula may be given as:



Example:

A committee of 3 persons is to be formed from among 3 men and 2 women. If the selection of the committee members is random, construct the probability distribution of the random variable 'Number of women in the committee'.

Solution:



Where x = 0, 1, 2, 3, ……, k; when n ≥ k

And x = 0, 1, 2, 3, ……., n; when n < k

$$P(x) = \frac{{}^{k}C_{x} \cdot {}^{N-k}C_{n-x}}{{}^{N}C_{n}}$$

$$P(x=0)=\frac{{}^{2}C_{0}\cdot{}^{3}C_{3}}{{}^{5}C_{3}}=\frac{1\times1}{10}=0.1$$

$$P(x=1)=\frac{{}^{2}C_{1}\cdot{}^{3}C_{2}}{{}^{5}C_{3}}=\frac{2\times3}{10}=0.6$$

$$P(x=2)=\frac{{}^{2}C_{2}\cdot{}^{3}C_{1}}{{}^{5}C_{3}}=\frac{1\times3}{10}=0.3$$

The Hyper Geometric Probability Distribution of RV 'No. of Women in the Committee' is as follows:

X	P(X)
0	0.1
1	0.6
2	0.3
Total	1

Poisson Probability Distribution:

1. A random variable created by counting the number of items or events in a unit of either time or space is called a 'Poisson Random Variable'.
2. Examples of Poisson random variable are the number of accidents per day on a highway, number of cars arriving at petrol pump in a five minute period of time, number of typing mistakes per page and number of defects in a painted surface, etc.
3. A Poisson probability distribution formula assigns probabilities to the values of the 'Poisson Random Variable':

$$P(x)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}$$

Where x = 0, 1, 2, 3, ……..

4. Where λ (lambda) is the only parameter of the distribution and e is the mathematical constant 2.71828……..:

(i) The number of events per unit of time or space remains stable for a long period of time. This is the parameter of the distribution denoted by λ.

(ii) The number of events in one time period is independent of the number of events in another time period.

Example:

In an industry, the average number of damaged output units per week is 10. What is the probability that there will be (i) no damaged unit in the next week, (ii) 5 damaged units in the next week, and (iii) 15 damaged units in the next week.

Solution:

(i) No damaged unit in the week:

X = number of damaged output units next week = 0

λ = average number of damaged units per week = 10

$$P(x=0)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}=\frac{e^{-10}\times10^{0}}{0!}=0.0000454$$

(ii) 5 damaged units in the next week:

X = number of damaged output units next week = 5

λ = average number of damaged units per week = 10

$$P(x=5)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}=\frac{e^{-10}\times10^{5}}{5!}=0.0378$$

(iii) 15 damaged units in the next week:

X = number of damaged output units next week = 15

λ = average number of damaged units per week = 10

$$P(x=5)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}=\frac{e^{-10}\times10^{15}}{15!}=0.0347$$

Poisson Approximation to Binomial Distribution:

The computation involved in the binomial distributions become quite tedious when n is large. In such cases the binomial distribution can be approximated to a Poisson distribution with $\lambda = n \cdot p$ under the following conditions:

(i) n is very large,

(ii) p is very small, and

(iii) $n \cdot p$ is finite.

A frequently used rule of thumb is that the approximation is appropriate when $p \leq 0.05$ and $n \geq 20$. However, the Poisson distribution sometimes provides close approximations even in cases where n is not large nor p is very small.

Example:

In a village, the local government approximated that 2% of the population are infected with seasonal flu due to absence of proper medication. What is the probability that the number of infected persons in a random sample of 50 will be 4?

Solution:

Using binomial distribution with:

n = 50, p = 0.02 and x = 4

$$P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x} = {}^{50}C_{4} \times 0.02^{4} \times 0.98^{50-4} = 0.0145$$

Using Poisson approximation to the binomial with:

$\lambda = n \cdot p = 50 \times 0.02 = 1$

$$P(x = 4) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!} = \frac{e^{-1} \times 1^{4}}{4!} = 0.0153$$

The Poisson probability is close to the binomial probability.

Mean and Variance of Poisson Distribution:

The mean of a Poisson Random Variable is the parameter of the Poisson distribution λ, that is:

$$E(X) = \lambda$$

The variance is also the parameter λ:

$$V(X) = \lambda$$

Thus mean and variance of Poisson distribution are equal to λ.

Continuous Probability Distribution (In Detail):

1. The concept of probability for continuous random variable is somewhat different with that of a discrete random variable.
2. The function or the formula of continuous probability distribution is generated and its curve is drawn on a graph paper such that:

   (i) The function is non-negative for all possible values of the random variable, and

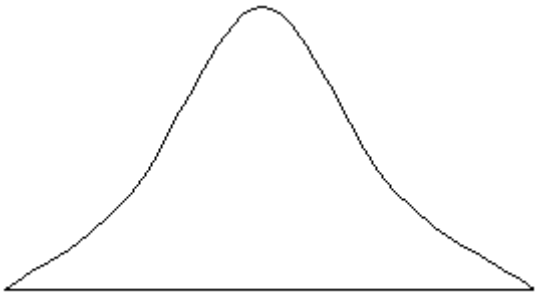
   (ii) The total area under the curve of the function is one.

   This function is called 'probability density function' and its curve a 'probability curve'.

3. The probability of an interval from a to b is defined as the area under the probability curve between the two vertical lines erected on the x-axis at the points a and b.
4. The probability of an individual value under the continuous probability distribution is considered zero.
5. Probabilities of continuous random variable are represented by areas under the probability curve.

Normal Probability Distribution (Properties of Normal Curve):

1. The most important and widely used probability density function is the 'Normal Distribution' where probability curve is a bell shaped symmetrical curve:



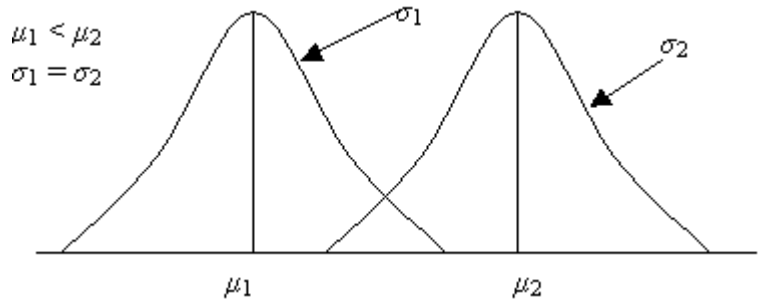
Normal Probability Curve

2. The most mathematical form of Normal Probability Density Function is:

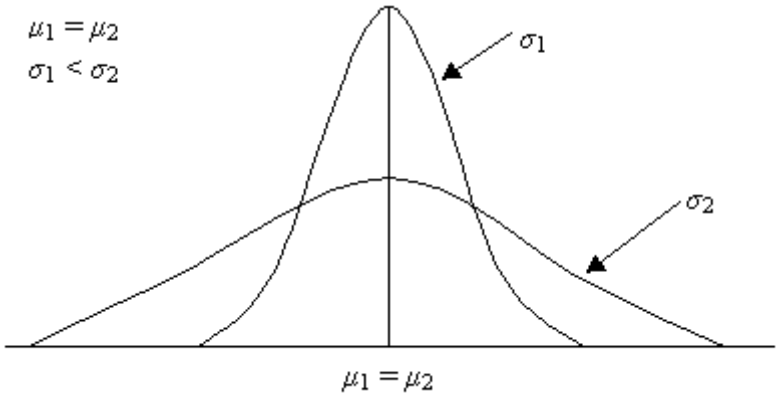
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where $-\infty \leq x \leq \infty$

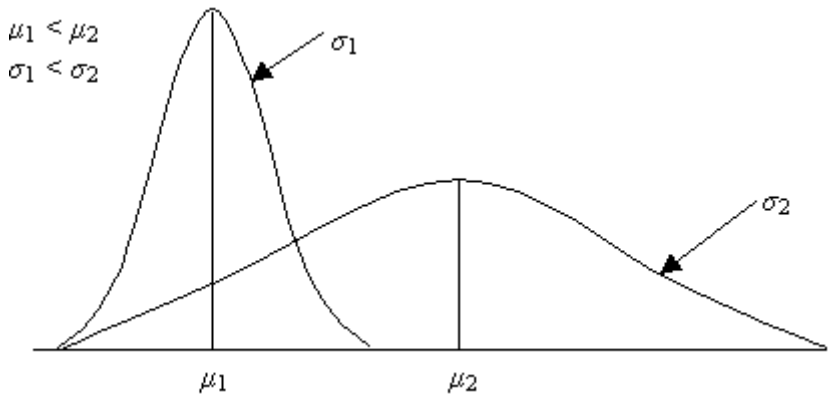
3. A normal probability distribution or its probability curve characterised by two quantities μ and σ called the parameters of the distribution.
4. Two normal curves with different means μ and equal standard deviations σ are as below:



5. The normal curves with different standard deviations σ and equal means μ:



6. Two normal curves with different means μ and different standard deviations σ:



Area under Normal Curve:

1. The area between two limits of an interval under a normal probability curve cannot be determined analytically.
2. Tables of areas evaluated numerically could have been constructed but it would be impossible for an infinite number of normal curves for all values of μ and σ.
3. This problem is overcome by 'Standard Normal Probability Distribution' whose mean is zero (μ = 0) and standard deviation is one (σ = 1). The standard normal variable is denoted by 'x':

$$z = \frac{x - \mu}{\sigma}$$

4. The table of areas under the standard normal curve is used to find area under normal probability curve:
5. Following steps are involved in determining the area or probability of a particular interval of a normal distribution with μ and σ:

   (i) Determine the z-values for each limit of interval,

   (ii) From the normal area table, determine the area for each z-value,

   (iii) Subtract the smaller area from the larger one.

6. Precisely, a value of random variable 'x' can be converted to value 'z' by:

$$z = \frac{x - \mu}{\sigma}$$

   Where μ and σ are the mean and standard deviation of the random variable z.

7. Conversely, the z-value can be converted into random variable x by:

$$x = \mu + \sigma \cdot z$$

8. 'z' is the number of standard deviations from or to the mean. All intervals containing the same number of standard deviations from mean will contain the same area under the curve for any normal distribution.
9. 'Normal Area Table' gives an idea under the curve to the left of a z-value. For example, for z = 1.51, the Area under Normal Curve (as shown in the Table) is 0.9345; for z = – 2.69, the Area under Normal Curve (from the Table) is 0.0036.
10. Some of the rules should be remembered:

    (i) Area to the left of z = 0 is 0.5000

(ii) Area to the left of z = -3.5 / -4.0 is 0

(iii) Area to the left of z = 3.5 / 4.0 is 1.000

Example:

A normal random variable x has mean μ = 24 and standard deviation σ = 1.8. Determine z values for x = 14, 15.9, 29.2 and 33. Also show these values on normal curve.

Solution:

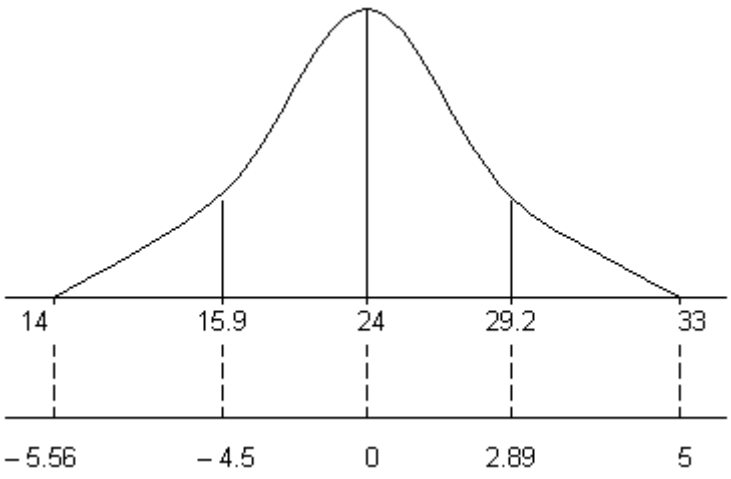
$$z = \frac{x - \mu}{\sigma}$$

For x = 14; $z = \frac{x - \mu}{\sigma} = \frac{14 - 24}{1.8} = -5.56$

For x = 15.9; $z = \frac{x - \mu}{\sigma} = \frac{15.9 - 24}{1.8} = -4.5$

For x = 29.2; $z = \frac{x - \mu}{\sigma} = \frac{29.2 - 24}{1.8} = 2.89$

For x = 33; $z = \frac{x - \mu}{\sigma} = \frac{33 - 24}{1.8} = 5$



Example:

A normal random variable x has mean μ = 36 and standard deviation 2.05, determine the values of x for z = – 3.36, – 1.8, 0.95 and 2.75.

Solution:

$x = \mu + \sigma \cdot z$

For z = – 3.36; x = 36 + 2.05 × (– 3.36) = 29.112 ≈ 29.11

For z = – 1.8; x = 36 + 2.05 × (– 1.8) = 32.31

For z = 0.95; x = 36 + 2.05 × 0.95 = 37.9475 ≈ 37.95

For z = 2.75; x = 36 + 2.05 × 2.75 = 41.6375 ≈ 41.64

Example:

The mean and SD of a normal random variable are 34.5 and 5.8 respectively. Find the following areas:

(i) to the left of 19.5

(ii) to the right of 40

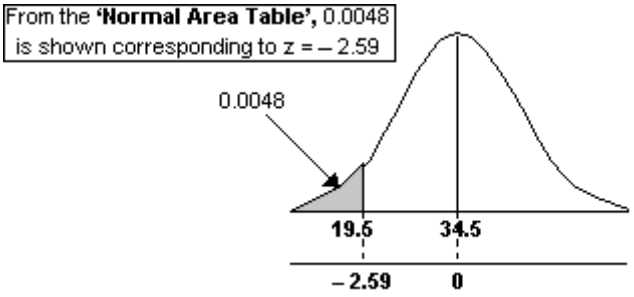
(iii) between 19.5 and 40

Solution:

(i) To the left of 19.5, i.e., P(x ≤ 19.5):

$P(-\infty \le x \le 19.5) = P(-\infty \le z \le -2.59) = 0.0048$

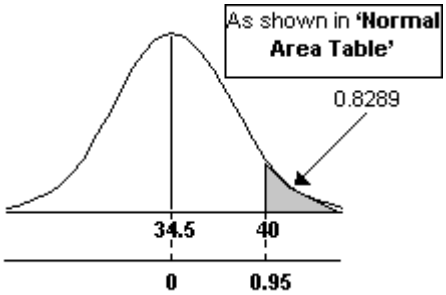
Where $z = \frac{19.5 - 34.5}{5.8} = -2.59$



(ii) To the right of 40, i.e., P (x ≥ 40):

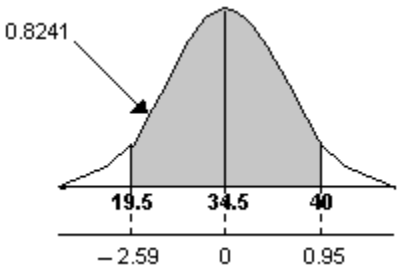
$P(40 \le x \le \infty) = P(0.95 \le z \le \infty) = 0.8289$

Where $z = \frac{40 - 34.5}{5.8} = 0.95$



(ii) Between 19.5 and 40, i.e., P(19.5 ≤ x ≤ 40):

$P(19.5 \le x \le 40) = P(-2.59 \le z \le 0.95) = 0.8289 - 0.0048 = 0.8241$



Continuity Correction:

1. A population with unknown mean and standard deviation can be assumed a normal population of the frequency distribution of a sample is symmetrical. The sample mean and sample standard deviation are used as estimates of population mean and population standard deviation respectively.

2. Observations or data are always discrete, recorded up to a certain degree of accuracy irrespective of whether the variable itself is discrete or continuous.
3. When the symmetrical distribution of any data is assumed to be normal, a continuity correction is applied to the observed values to make the data continuous.
4. If the data are recorded in whole numbers, data values are considered as mid-points of the intervals x ± 0.5, if the data are recorded up to one decimal place, data values are considered as mid points of the intervals x ± 0.05 and so on. It should be cleared that the 0.5 and 0.05 should be subtracted from lower limit and added to upper limit or at most limit.

Normal Approximation to Binomial Distribution:

A Binomial Distribution with large n and moderate p can be approximated to a Normal Distribution with mean μ = np and $\sigma = \sqrt{npq}$:

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

Example:

A pair of dice is rolled for 800 times. What is the probability that a total of 6 occur:

(i) at least 100 times, and

(ii) between 150 to 300 times.

Solution:

n = 800

p = $\frac{5}{36}$ E = {15, 24, 33, 42, 51}

q = $\frac{31}{36}$

$$\mu = n \times p = 800 \times \frac{5}{36} = 111.1$$

$$\sigma = \sqrt{n \times p \times q} = \sqrt{800 \times \frac{5}{36} \times \frac{31}{36}} = \sqrt{95.7} = 9.78$$

(i) **Probability of at least 100 times, i.e., P(100 ≤ x ≤ 800) or P(99.5 ≤ x ≤ 800.5):**

$$z_1 = \frac{99.5 - 111.1}{9.78} = -1.19$$
$$z_2 = \frac{800.5 - 111.1}{9.78} = 70.49$$

P (–1.19 ≤ z ≤ 70.49)

From **'Normal Area Table'** the Normal Area corresponding to – 1.19 is 0.1170

= 1 – 0.1170 = 0.8830

(ii) **Probability of between 150 and 300 times, i.e., P(130 ≤ x ≤ 300) or P(149.5 ≤ x ≤ 300.5):**

$$z_1 = \frac{129.5 - 111.1}{9.78} = 1.88$$
$$z_2 = \frac{300.5 - 111.1}{9.78} = 19.37$$

P (1.88 ≤ z ≤ 19.37)

From **'Normal Area Table'** the Normal Area corresponding to 1.88 is 0.9699

= 1 – 0.9699 = 0.0301

---

4. Hypothesis testing: Formulation, Rejection rule, one and two tailed tests, significance level, and degrees of freedom type I and type II errors, Standard Error. Different types of significance test for various purposes. Chi- square test, student's t- test.

Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students".* Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$
$$(or)\, H_1 : \mu > 100$$
$$(or)\, H_1 : \mu < 100$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e. 1) Type-I error

2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

α = P (Committing Type-I error)

= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$.

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called as two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}C_n$ possible samples. If we calculate some particular statistic from each of the ${}^{N}C_n$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

i.e. S.E (t) = $\sqrt{Var(t)}$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.
3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

i.e. test statistic $Z = \frac{t - E(t)}{S.E(t)}$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α.
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α.
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

Where, μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\frac{1}{n-1}\sum (x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1}^2 / n_1\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2}^2 / n_2\right)$

Therefore $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \sim N(0,1)$$ [since $\mu_1 - \mu_2$ =0 from H_0]

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: If ${\sigma_1}^2$ and ${\sigma_2}^2$ are unknown then we can consider ${S_1}^2$ and ${S_2}^2$ as the estimate value of ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2 , for large sample, sample standard deviations follows a normal distribution with mean σ and variance $\sigma^2/2n$ i.e. $s \sim N\left(\sigma, \sigma^2/2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$

against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$

$\Rightarrow Z = \frac{s - \sigma}{\sigma / \sqrt{2n}} \sim N(0,1)$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, {\sigma_1}^2 / 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, {\sigma_2}^2 / 2n_2\right)$

Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$

against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}} \sim N(0,1)$$ [since $\sigma_1 - \sigma_2$ =0 from H_0]

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$ and $S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Chi-Square Test

What is a Chi Square Test?

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

A chi-square goodness of fit test determines if sample data matches a population.

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.

A very large chi square test statistic means that the data does not fit very well. In other words, there isn't a relationship.

Uses

The chi-squared distribution has many uses in statistics, including:

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Sample variance study when the underlying distribution is normal.
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).

What is a Chi-Square Statistic?

The formula for the chi-square statistic used in the chi square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Steps for using and interpreting chi-square

To review, the chi-square method of hypothesis testing has seven basic steps.

1. State the null and research/alternative hypotheses.

2. Specify the decision rule and the level of statistical significance for the test, i.e., .05, .01, or .001. (A significance level of .01 would mean that the probability of the chi-square value must be .01 or less to reject the null hypothesis, a more stringent criterion than .05.)

3. Compute the expected values.

4. Compute the chi-square statistic.

5. Determine the degrees of freedom for the table. Then identify the critical value of chi-square at the specified level of significance and appropriate degrees of freedom.

6. Compare the computed chi-square statistic with the critical value of chi-square; reject the null hypothesis if the chi-square is equal to or larger than the critical value; accept the null hypothesis if the chi-square is less than the critical value.

7. State a substantive conclusion, i.e., describes the meaning and importance of the test results in terms of the historical problem under investigation.

Example question:

256 visual artists were surveyed to find out their zodiac sign. The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23). Test the hypothesis that zodiac signs are evenly distributed across visual artists.

Step 1: Make a table with columns for "Categories," "Observed," "Expected," "Residual (Obs-Exp)", "(Obs-Exp)2" and "Component (Obs-Exp)2 / Exp." Don't worry what these mean right now; We'll cover that in the following steps.

Step 2: Fill in your categories. Categories should be given to you in the question. There are 12 zodiac signs, so:

Step 3: Write your counts. Counts are the number of each items in each category in column 2. You're given the counts in the question:

Step 4: Calculate your expected value for column 3. In this question, we would expect the 12 zodiac signs to be evenly distributed for all 256 people, so 256/12=21.333. Write this in column 3.

Step 5: Subtract the expected value (Step 4) from the Observed value (Step 3) and place the result in the "Residual" column. For example, the first row is Aries: 29-21.333=7.667.

Step 6: Square your results from Step 5 and place the amounts in the (Obs-Exp)2 column.

Step 7: Divide the amounts in Step 6 by the expected value (Step 4) and place those results in the final column.

Step 8: Add up (sum) all the values in the last column.

Table: Calculation for Chi-square Test

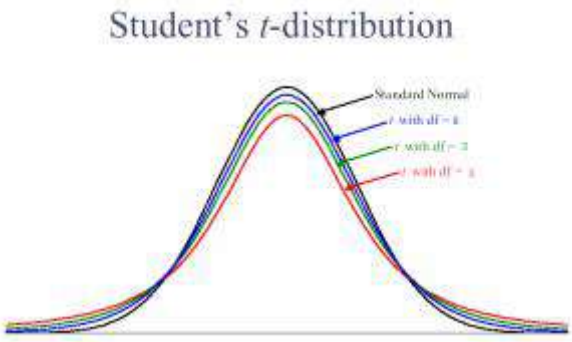
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

Student's t- Test

The t test tells you how significant the differences between groups are; In other words it lets you know if those differences (measured in means) could have happened by chance.

A very simple example: Let's say you have a cold and you try a naturopathic remedy. Your cold lasts a couple of days. The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week. You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you really want to know is, are these results repeatable? A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

Another example: Student's T-tests can be used in real life to compare averages. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill"). The control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.



The T Score.

The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different from each other as they are within each other. When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.

A large t-score tells you that the groups are different.

A small t-score tells you that the groups are similar.

T-Values and P-values

How big is "big enough"? Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05. Low p-values are good; they indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

Calculating the Statistic / Test Types

There are three main types of t-test:

An Independent Samples t-test compares the means for two groups.

A Paired sample t-test compares means from the same group at different times (say, one year apart).

A One sample t-test tests the mean of a single group against a known mean.

You probably don't want to calculate the test by hand (the math can get very messy, but if you insist you can find the steps for an independent samples t test here.

What is a Paired T Test (Paired Samples T Test / Dependent Samples T Test)?

A paired t test (also called a correlated pairs t-test, a paired samples t test or dependent samples t test) is where you run a t test on dependent samples. Dependent samples are essentially connected — they are tests on the same person or thing. For example:

Knee MRI costs at two different hospitals,

Two tests on the same person before and after training,

Two blood pressure measurements on the same person using different equipment.

When to Choose a Paired T Test / Paired Samples T Test / Dependent Samples T Test

Choose the paired t-test if you have two measurements on the same item, person or thing. You should also choose this test if you have two items that are being measured with a unique condition. For example, you might be measuring car safety performance in vehicle research and testing and subject the cars to a series of crash tests. Although the manufacturers are different, you might be subjecting them to the same conditions.

With a "regular" two sample t test, you're comparing the means for two different samples. For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills. If you take a random sample each group separately and they have different conditions, your samples are independent and you should run an independent samples t-test (also called between-samples and unpaired-samples).

The null hypothesis for the independent samples t-test is $\mu 1 = \mu 2$. In other words, it assumes the means are equal. With the paired t test, the null hypothesis is that the pairwise difference between the two tests is equal ($H0: \mu d = 0$). The difference between the two tests is very subtle; which one you choose is based on your data collection method.

Paired Samples T Test By hand

Example question:

Calculate a paired t test by hand for the following data:

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

Step 1: Subtract each Y score from each X score.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

Step 2: Add up all of the values from Step 1.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2
		SUM:	**-73**

Step 3: Square the differences from Step 1.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Step 4: Add up all of the squared differences from Step 3.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Step 5: Use the following formula to calculate the t-score:

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

ΣD: Sum of the differences (Sum of X-Y from Step 2)

ΣD2: Sum of the squared differences (from Step 4)

(ΣD)2: Sum of the differences (from Step 2), squared.

If you're unfamiliar with Σ you may want to read about summation notation first.

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -2.74$$

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this example problem, with df = 10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Note: You can ignore the minus sign when comparing the two t-values, as ± indicates the direction; the p-value remains the same for both directions.

---

5. Sampling techniques for geographical analysis.

What is sampling?

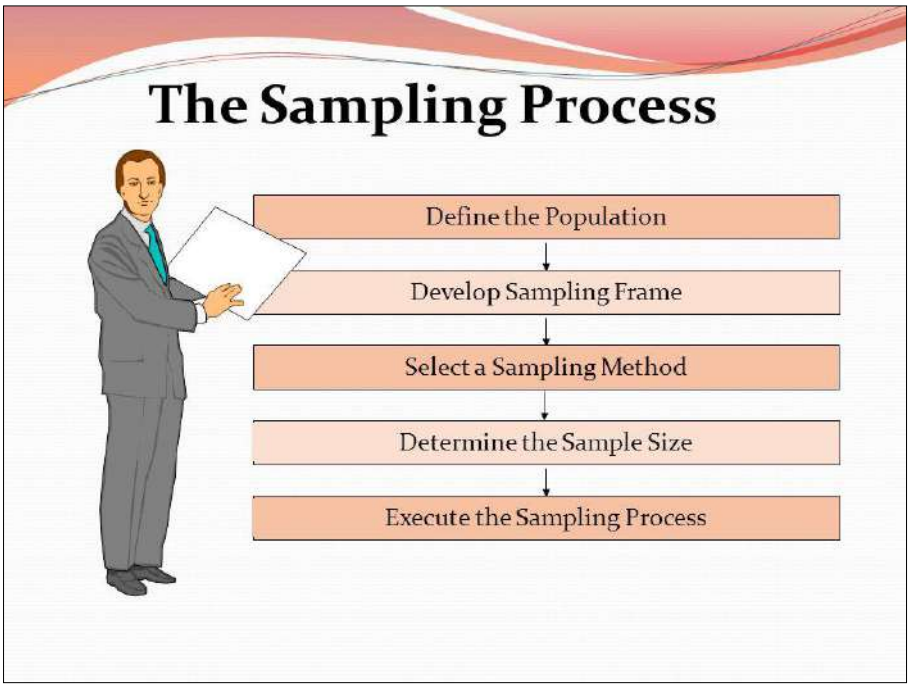
Sampling is the acquisition of information about a relatively small part of a larger group or population, usually with the aim of making inferential generalizations about the larger group.

- A shortcut method for investigating a whole population
- Data is gathered on a small part of the whole parent population or sampling frame, and used to inform what the whole picture is like

Why sample?

In reality there is simply not enough; time, energy, money, labour/man power, equipment, access to suitable sites to measure every single item or site within the parent population or whole sampling frame.

Therefore an appropriate sampling strategy is adopted to obtain a representative, and statistically valid sample of the whole.



Sampling considerations

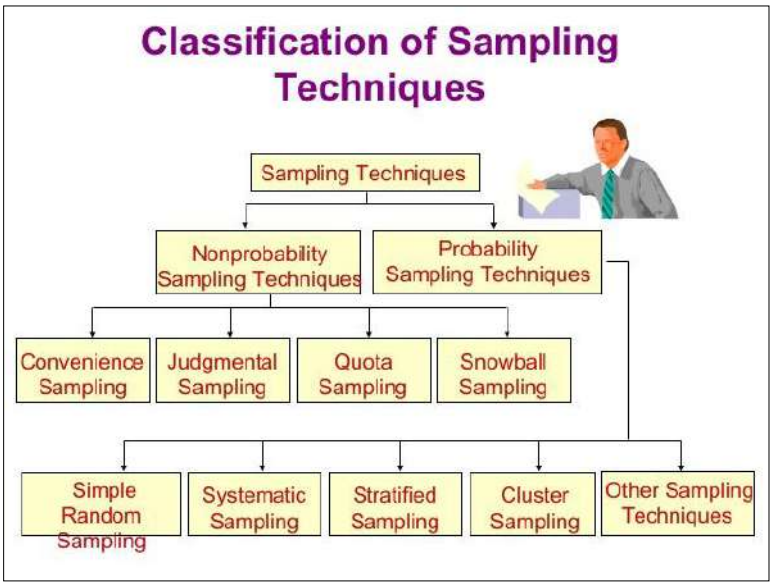
- Larger sample sizes are more accurate representations of the whole
- The sample size chosen is a balance between obtaining a statistically valid representation, and the time, energy, money, labour, equipment and access available
- A sampling strategy made with the minimum of bias is the most statistically valid
- Most approaches assume that the parent population has a normal distribution where most items or individuals clustered close to the mean, with few extremes
- A 95% probability or confidence level is usually assumed, for example 95% of items or individuals will be within plus or minus two standard deviations from the mean
- This also means that up to five per cent may lie outside of this - sampling, no matter how good can only ever be claimed to be a very close estimate

Sampling techniques

There are two types of sampling methods:

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.

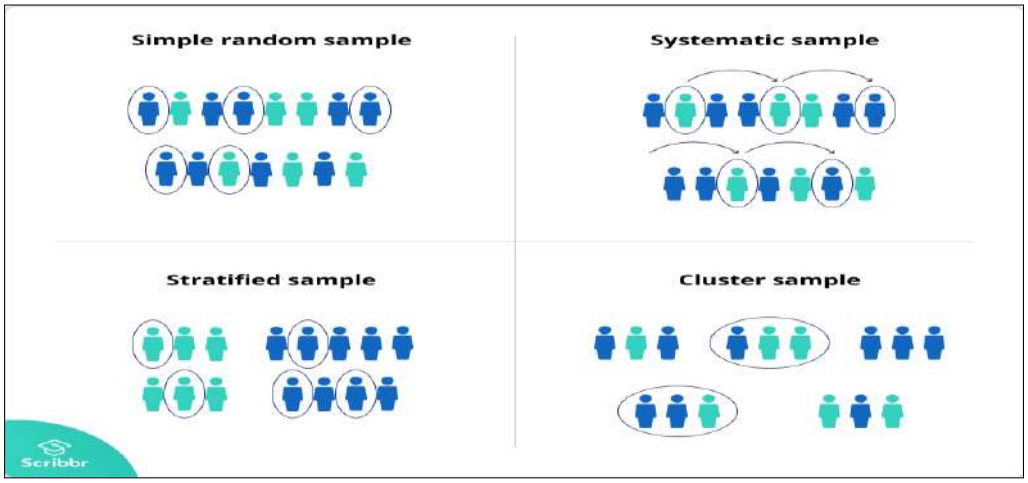
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.



Probability Sampling Techniques:

Three main types of probability sampling strategy are:

- Random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling



Within these types, you may then decide on a; point, line, area method.

Random sampling

- Least biased of all sampling techniques, there is no subjectivity - each member of the total population has an equal chance of being selected
- Can be obtained using random number tables
- Microsoft Excel has a function to produce random number

The function is simply:

- =RAND()

Type that into a cell and it will produce a random number in that cell. Copy the formula throughout a selection of cells and it will produce random numbers.

You can modify the formula to obtain whatever range you wish, for example if you wanted random numbers from one to 250, you could enter the following formula:

- =INT(250*RAND())+1

Where INT eliminates the digits after the decimal, 250* creates the range to be covered, and +1 sets the lowest number in the range.

Paired numbers could also be obtained using;

- =INT(9000*RAND())+1000

These can then be used as grid coordinates, metre and centimetre sampling stations along a transect, or in any feasible way.

Methodology

A. Random point sampling

- A grid is drawn over a map of the study area
- Random number tables are used to obtain coordinates/grid references for the points

- Sampling takes place as feasibly close to these points as possible

B. Random line sampling

- Pairs of coordinates or grid references are obtained using random number tables, and marked on a map of the study area
- These are joined to form lines to be sampled

C. Random area sampling

- Random number tables generate coordinates or grid references which are used to mark the bottom left (south west) corner of quadrats or grid squares to be sampled

Advantages and disadvantages of random sampling

Advantages:

- Can be used with large sample populations
- Avoids bias

Disadvantages:

- Can lead to poor representation of the overall parent population or area if large areas are not hit by the random numbers generated. This is made worse if the study area is very large
- There may be practical constraints in terms of time available and access to certain parts of the study area

Systematic sampling

Samples are chosen in a systematic or regular way.

- They are evenly/regularly distributed in a spatial context, for example every two metres along a transect line
- They can be at equal/regular intervals in a temporal context, for example every half hour or at set times of the day

- They can be regularly numbered, for example every 10th house or person

Methodology

A. Systematic point sampling

A grid can be used and the points can be at the intersections of the grid lines, or in the middle of each grid square. Sampling is done at the nearest feasible place. Along a transect line, sampling points for vegetation/pebble data collection could be identified systematically, for example every two metres or every 10th pebble

B. Systematic line sampling

The eastings or northings of the grid on a map can be used to identify transect lines. Alternatively, along a beach it could be decided that a transect up the beach will be conducted every 20 metres along the length of the beach

C. Systematic area sampling

A 'pattern' of grid squares to be sampled can be identified using a map of the study area, for example every second/third grid square down or across the area - the south west corner will then mark the corner of a quadrat. Patterns can be any shape or direction as long as they are regular.

Advantages and disadvantages of systematic sampling

Advantages:

- It is more straight-forward than random sampling
- A grid doesn't necessarily have to be used, sampling just has to be at uniform intervals
- A good coverage of the study area can be more easily achieved than using random sampling

Disadvantages:

- It is more biased, as not all members or points have an equal chance of being selected
- It may therefore lead to over or under representation of a particular pattern

Stratified sampling

This method is used when the parent population or sampling frame is made up of sub-sets of known size. These sub-sets make up different proportions of the total, and therefore sampling should be stratified to ensure that results are proportional and representative of the whole.

A. Stratified systematic sampling

The population can be divided into known groups, and each group sampled using a systematic approach. The number sampled in each group should be in proportion to its known size in the parent population.

For example: the make-up of different social groups in the population of a town can be obtained, and then the number of questionnaires carried out in different parts of the town can be stratified in line with this information. A systematic approach can still be used by asking every fifth person.

B. Stratified random sampling

A wide range of data and fieldwork situations can lend themselves to this approach - wherever there are two study areas being compared, for example two woodlands, river catchments, rock types or a population with sub-sets of known size, for example woodland with distinctly different habitats.

Random point, line or area techniques can be used as long as the number of measurements taken is in proportion to the size of the whole.

For example: if an area of woodland was the study site, there would likely be different types of habitat (sub-sets) within it. Random sampling may altogether ‘miss' one or more of these.

Stratified sampling would take into account the proportional area of each habitat type within the woodland and then each could be sampled accordingly; if 20 samples were to be taken in the woodland as a whole, and it was found that a shrubby clearing accounted for 10% of the total area, two samples would need to be taken within the clearing. The sample points could still be identified randomly or systematically within each separate area of woodland.

Advantages and disadvantages of stratified sampling

Advantages:

- It can be used with random or systematic sampling, and with point, line or area techniques
- If the proportions of the sub-sets are known, it can generate results which are more representative of the whole population
- It is very flexible and applicable to many geographical enquiries
- Correlations and comparisons can be made between sub-sets

Disadvantages:

- The proportions of the sub-sets must be known and accurate if it is to work properly
- It can be hard to stratify questionnaire data collection, accurate up to date population data may not be available and it may be hard to identify people's age or social background effectively

Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example

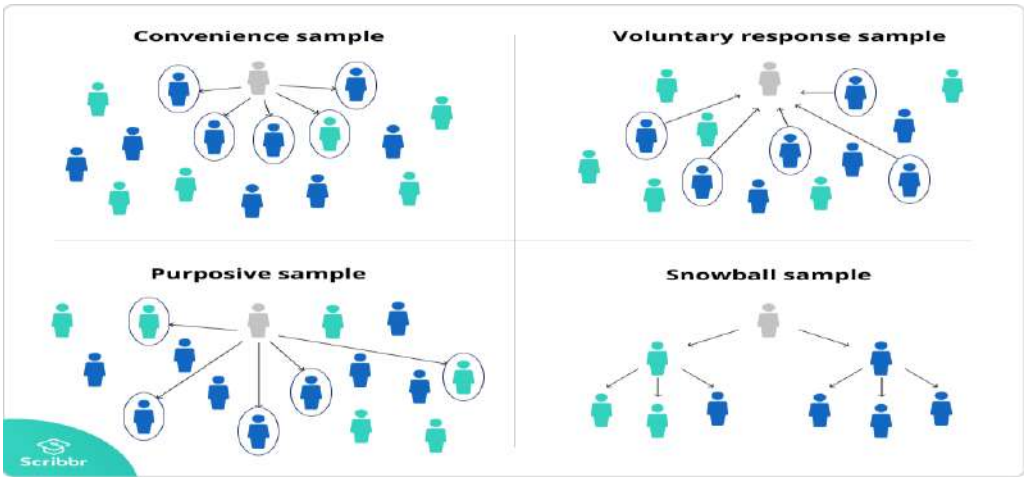
The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non Probability Sampling Techniques

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

Example

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

Example

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.

Example

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

Example

You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

****

GEO 295.2: ADVANCED QUANTITATIVE METHODS

1. Analysis of Variance: Objectives; One-way and Two-way ANOVA.

2. Fitting Second Degree Polynomial curves to bivariate geographical data and testing by ANOVA.

3. Multiple Regression: Linear multiple regression equation, Multiple and partial correlation coefficient.

4. Elementary multiple regression modeling techniques: Stepwise variable entry method, Path Analysis.

5. Model building techniques

1. Analysis of Variance: Objectives; One-way and Two-way ANOVA.

Analysis of Variance (ANOVA)

Professor R.A. Fisher was the first man to use the term 'Variance'* and, in fact, it was he who developed a very elaborate theory concerning ANOVA, explaining its usefulness in practical field. Later on Professor Snedecor and many others contributed to the development of this technique. ANOVA is essentially a procedure for testing the difference among different groups of data for homogeneity. "The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes."1 There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes. Hence, it is a method of analysing the variance to which a response is subject into its various components corresponding to various sources of variation. Through this technique one can explain whether various varieties of seeds or fertilizers or soils differ significantly so that a policy decision could be taken accordingly, concerning a particular variety in the context of agriculture researches. Similarly, the differences in

various types of feed prepared for a particular class of animal or various types of drugs manufactured for curing a specific disease may be studied and judged to be significant or not through the application of ANOVA technique. Likewise, a manager of a big concern can analyse the performance of various salesmen of his concern in order to know whether their performances differ significantly. Thus, through ANOVA technique one can, in general, investigate any number of factors which are hypothesized or said to influence the dependent variable. One may as well investigate the differences amongst various categories within each of these factors which may have a large number of possible values. If we take only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one-way ANOVA and in case we investigate two factors at the same time, then we use two-way ANOVA. In a two or more-way ANOVA, the interaction (i.e., inter-relation between two independent variables/factors), if any, between two independent variables affecting a dependent variable can as well be studied for better decisions.

ONE-WAY ANOVA:

Under the one-way ANOVA, we consider only one factor and then observe that the reason for said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor.

Illustration 1

Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	*Per acre production data*		
	Variety of wheat		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution

Through direct method:

First, we calculate the mean of each of these samples:

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

Mean of the sample means or

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$$

$$= \frac{6+5+4}{3} = 5$$

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + n_3\left(\bar{X}_3 - \bar{\bar{X}}\right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$
$$= 4 + 0 + 4$$
$$= 8$$

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \Sigma\left(X_{3i} - \bar{X}_3\right)^2, \qquad i = 1, 2, 3, 4$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$
$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$
$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$
$$= \{0+1+9+4\} + \{0+0+4+4\} + \{1+0+1+0\}$$
$$= 14 + 8 + 2$$
$$= 24$$

$$SS \text{ for total variance} = \Sigma\left(X_{ij} - \bar{\bar{X}}\right)^2 \qquad i = 1, 2, 3\ldots$$

$$j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2$$
$$+ (5-5)^2 + (5-5)^2 + (3-5)^2$$
$$+ (7-5)^2 + (5-5)^2 + (4-5)^2$$
$$+ (3-5)^2 + (4-5)^2$$
$$= 1+4+4+9+0+0+4+4+0+1+4+1$$
$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

$$= 8 + 24$$
$$= 32$$

We can now set up the ANOVA table for this problem:

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

The above table shows that the calculated value of *F* is 1.5 which is less than the table value of

4.26 at 5% level with d.f. being *v*1 = 2 and *v*2 = 9 and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

TWO-WAY ANOVA:

Two-way ANOVA technique is used when the data are classified on the basis of two factors. For example, the agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used. A business firm may have its sales data classified on the basis of different salesmen and also on the basis of sales in different regions. In a factory, the various units of a product produced during a certain period may be classified on the basis of different varieties of machines used and also on the basis of different grades of labour. Such a two-way design may have repeated measurements of each factor or may not have repeated values. The ANOVA technique is little different in case of repeated measurements where we also compute the interaction variation.

Illustration 2

Set up an analysis of variance table for the following two-way design results:

Per Acre Production Data of Wheat

(in metric tonnes)

Varieties of seeds	*A*	*B*	*C*
Varieties of fertilizers			
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Also state whether variety differences are significant at 5% level.

SOLUTION:

Step (i) $T = 60, n = 12, \therefore$ Correction factor $= \frac{(T)^2}{n} = \frac{60 \times 60}{12} = 300$

Step (ii) Total $SS = (36 + 25 + 25 + 49 + 25 + 16 + 9 + 9 + 9 + 64 + 49 + 16) - \left(\frac{60 \times 60}{12}\right)$

$= 332 - 300$

$= 32$

Step (iii) SS between columns treatment $= \left[\frac{24 \times 24}{4} + \frac{20 \times 20}{4} + \frac{16 \times 16}{4}\right] - \left[\frac{60 \times 60}{12}\right]$

$= 144 + 100 + 64 - 300$

$= 8$

Step (iv) SS between rows treatment $= \left[\frac{16 \times 16}{3} + \frac{16 \times 16}{3} + \frac{9 \times 9}{3} + \frac{19 \times 19}{3}\right] - \left[\frac{60 \times 60}{12}\right]$

$= 85.33 + 85.33 + 27.00 + 120.33 - 300$

$= 18$

Step (v) SS residual or error = Total SS – (SS between columns + SS between rows)

$= 32 - (8 + 18)$

$= 6$

THE ANOVA TABLE:

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (or the tables values)*
Between columns (i.e., between varieties of seeds)	8	$(3-1)=2$	$8/2=4$	$4/1=4$	$F(2,6)=5.14$
Between rows (i.e., between varieties of fertilizers)	18	$(4-1)=3$	$18/3=6$	$6/1=6$	$F(3,6)=4.76$
Residual or error	6	$(3-1)\times(4-1)=6$	$6/6=1$		
Total	32	$(3\times4)-1=11$			

From the said ANOVA table, we find that differences concerning varieties of seeds are insignificant at 5% level as the calculated *F*-ratio of 4 is less than the table value of 5.14, but the variety differences concerning fertilizers are significant as the calculated *F*-ratio of 6 is more than its table value of 4.76.

---

2. Fitting Second Degree Polynomial curves to bivariate geographical data and testing by ANOVA.

Fitting Second Degree Polynomial Curves

Fit a second-degree parabola for the following data:

x	0	1	2	3	4
y	1	3	4	5	6

Solution: Let$Y=a + bx+cx^2$be the second-degree parabola and we have to determine a, b and c. Normal equations for second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3, \text{ and}$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

To solve above normal equations, we need $\sum y$, $\sum x$, $\sum xy$, $\sum x^2 y$, $\sum x^2$,

$\sum x^3$and $\sum x^4$which are obtained from following table:

x	y	xy	X^2	X^2y	X^3	X^4
0	1	0	0	0	0	0
1	3	3	1	3	1	1

2	4	8	4	16	8	16
3	5	15	9	45	27	81
4	6	24	16	96	64	256
$\sum x=10$	$\sum y=19$	$\sum xy=50$	$\sum x^2=30$	$\sum x^2y=160$	$\sum x^3=100$	$\sum x^4=354$

Substituting the values of $\sum y$, $\sum x$, $\sum xy$, $\sum x^2y$, $\sum x^2$, $\sum x^3$and

$\sum x^4$in above normal equations, we have

$19 = 5a + 10b + 30c$ ……..(1)

$50 = 10a + 30b + 100c$ …… (2)

$160 = 30a + 100b + 354c$ ……(3)

Now, we solve equations (1), (2) and (3).

Multiplying equation (1) by 2, we get

$38 =10a + 20b + 60c$ ……. (4)

Subtracting equation (4) from equation (2)

$50 = 10a + 30b + 100c$

$38 =10a + 20b + 60c$

$12 = 10b + 40c$ …… (5)

Multiplying equation (2) by 3, we get

$150 = 30a + 90b + 300c$ …….. (6)

Subtracting equation (3) from equation (6), we get

$160 = 30a + 100b + 354c$

150 = 30a + 90b + 300c

---

10 = 10b + 54c ……. (7)

Now we solve equation (5) and (7)

Subtracting equation (7) from equation (5), we get

10 = 10b + 54c

12 =10b + 40c

−2 = 14 c

c = - 2/14

c = - 0.1429

Substituting the value of c in equation (7), we get

10 = 10 b + 54x (-0.1429)

10 = 10 b -7.7166

17.7166 = 10 b

b = 1.7717

Substituting the value of b and c in equation (1), we get

19 = 5a +10 x (1.7717) + (-0.1429x30)

19 = 5a + 17.717 − 4.287

a =1.114

Thus, the second degree of parabola of best fit is

$Y= 1.114 + 1.7717 X-0.1429X^2$

3. Multiple Regression: Linear multiple regression equation, Multiple and partial correlation coefficient.

Linear Multiple Regression Equation

When there are two or more than two independent variables, the analysis concerning relationship is known as multiple correlations and the equation describing such relationship as the multiple regression equation. We here explain multiple correlation and regression taking only two independent variables and one dependent variable (Convenient computer programs exist for dealing with a great number of variables).

In this situation the results are interpreted as shown below:

Multiple regression equation assumes the form:

$$Y = a + b_1X_1 + b_2X_2$$

Where X_1 and X_2 are two independent variables and Y being the dependent variable, and the constants

a, b_1 and b_2 can be solved by solving the following three normal equations:

$$\sum Y_i = na + b_1 \sum X_{1i} + b_2 \sum X_{2i}$$

$$\sum X_{1i}Y_i = a\sum X_{1i} + b_1\sum X_{1i^2} + b2\sum X1_iX2_i$$

$$\sum X_{2i}Y_i = a\sum X_{2i} + b_1\sum X_{1i}\, X_{2i} + b_2\sum X_{2i^2}$$

Example 1: A statistical analyst is analysing the vending machine routes in the distribution system. He/she is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The company manager responsible for the study has suggested that the two most important variables affecting the delivery time Y (in minutes), are (i) the number of cases (X1) and (ii) the distance travelled (in m) by the route driver (X2). The delivery time data collected by the statistical analyst is given below:

Time (Y)	No. of Cases (X1)	Distance (X2)
20	10	50
10	5	20
10	5	30
15	5	10
15	10	10
20	10	30
10	5	10
25	15	40
30	10	80
15	10	20
20	10	10
10	5	40

Check whether there is a linear relationship between Y (Time) and the two independent variables X1 (number of cases) and X2 (distance). Calculate the values of the regression coefficients and fit the regression equation.

Solution: To find the values of regression coefficients and fit the regression equation for the given data, we form the following table:

Time (Y)	No. of Cases (X_1)	Distance (X_2)	Y^2	$(X_1)^2$	$(X_2)^2$	X_1Y	X_2Y	X_1X_2
20	10	50	400	100	2500	200	1000	500
10	5	20	100	25	400	50	200	100
10	5	30	100	25	900	50	300	150
15	5	10	225	25	100	75	150	50
15	10	10	225	100	100	150	150	100
20	10	30	400	100	900	200	600	300
10	5	10	100	25	100	50	100	50
25	15	40	625	225	1600	375	1000	600
30	10	80	900	100	6400	300	2400	800
15	10	20	225	100	400	150	300	200
20	10	10	400	100	100	200	200	100
10	5	40	100	25	1600	50	400	200
$\sum Y_i$ =200	$\sum X_{1i}$ =100	$\sum X_{2i}$ =350	$\sum Y_i^2$ =3800	$\sum X_{1i}^2$ =950	$\sum X_{2i}^2$ =15100	$\sum X_{1i}Y_i$ =1850	$\sum X_{2i}Y_i$ =6800	$\sum X_{1i}X_{2i}$ =3150

On putting the values calculated in the table in the above equations, we get

12 Bˆ +100 Bˆ +350Bˆ =200 ..... (i)

100 Bˆ +950Bˆ +3150Bˆ =1850 ..... (ii)

350 Bˆ +3150Bˆ +15100Bˆ =6800 ..... (iii)

From equation (i), we have

$$\hat{B}_0 = \frac{\left(200 - 100\,\hat{B}_1 - 350\,\hat{B}_2\right)}{12}$$ ..... (iv)

On putting the value of B in equations (ii) and (iii) and simplifying,

We get

1400 Bˆ +2800Bˆ =2200 ..... (v)

2800Bˆ +58700Bˆ =11600 ..... (vi)

On solving equations (v) and (vi), we get

$B\hat{}_1$=1.3002 $B\hat{}_2$=0.1356

$$\hat{B}_0 = \frac{\left(200 - 100\left(1.3002\right) - 350\left(0.1356\right)\right)}{12} = 1.8765$$

Hence, the fitted equation is

$$Y = 1.8765 + 1.3002\, X_1 + 0.1356\, X_2$$

So we can conclude that there is a linear relationship between Y (time in seconds) and the two independent variables X1 (number of cases) and X2 (distance). As the regression coefficients for both variables are positive, these affect the delivery time. The numerical value of the regression coefficient ˆB_1associated with X1 is higher than the value of ˆB2associated with X2. It shows that the number of cases affects the delivery time more than the distance travelled.

Partial correlation, multiple regressions, and correlation

Partial correlation

Partial correlation measures the correlation between X and Y, controlling for Z. Comparing the bivariate (zero-order) correlation to the partial (first-order) correlation allows us to determine if the relationship between X and Y is direct, spurious, or intervening. Interaction cannot be determined with partial correlations.

Formula for partial correlation

Formula for partial correlation coefficient for X and Y, controlling for Z

$$r_{yx.z} = \frac{r_{yx} - \left(r_{yz}\right)\left(r_{xz}\right)}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

We must first calculate the zero-order coefficients between all possible pairs of variables (Y and X, Y and Z, X and Z) before solving this formula.

Example

• Husbands' hours of housework per week (Y)

• Number of children (X)

• Husbands' years of education (Z)

Scores on Three Variables for 12 Dual–Wage-Earner Families

Family	Husband's Housework (Y)	Number of Children (X)	Husband's Years of Education (Z)
A	1	1	12
B	2	1	14
C	3	1	16
D	5	1	16
E	3	2	18
F	1	2	16
G	5	3	12
H	0	3	12
I	6	4	10
J	3	4	12
K	7	5	10
L	4	5	16

Correlation matrix

The bivariate (zero-order) correlation between husbands' housework and number of children is +0.50. This indicates a positive relationship.

Zero-Order Correlations

↓	Husband's Housework (Y)	Number of Children (X)	Husband's Years of Education (Z)
Husband's Housework (Y)	1.00	0.50	−0.30
Number of Children (X)		1.00	−0.47
Husband's Years of Education (Z)			1.00

First-order correlation

Calculate the partial (first-order) correlation between husbands' housework (Y) and number of children (X), controlling for husbands' years of education (Z).

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

$$r_{yx.z} = \frac{(0.50) - (-0.30)(-0.47)}{\sqrt{1 - (-0.30)^2}\sqrt{1 - (-0.47)^2}}$$

$$r_{yx.z} = 0.43$$

Interpretation

Comparing the bivariate correlation (+0.50) to the partial correlation (+0.43) finds little change. The relationship between number of children and husbands' housework has not changed, controlling for husbands' education. Therefore, we have evidence of a direct relationship.

Bivariate & multiple regressions

Bivariate regression equation

$$Y = a + bX = \beta_0 + \beta_1 X$$

- $a = \beta_0$ = Y intercept
- $b = \beta_1$ = slope

Multivariate regression equation

$$Y = a + b_1X_1 + b_2X_2 = \beta_0 + \beta_1X_1 + \beta_2X_2$$

$b_1 = \beta_1$ = partial slope of the linear relationship between the first independent variable and Y

$b_2 = \beta_1$ = partial slope of the linear relationship between the second independent variable and Y

Multiple regression

$$Y = a + b_1X_1 + b_2X_2 = \beta_0 + \beta_1X_1 + \beta_2X_2$$

$a = \beta_0$ = the Y intercept, where the regression line crosses the Y axis

$b_1 = \beta_1$ = partial slope for X_1 on Y – β_1 indicates the change in Y for one unit change in X_1, controlling for X_2

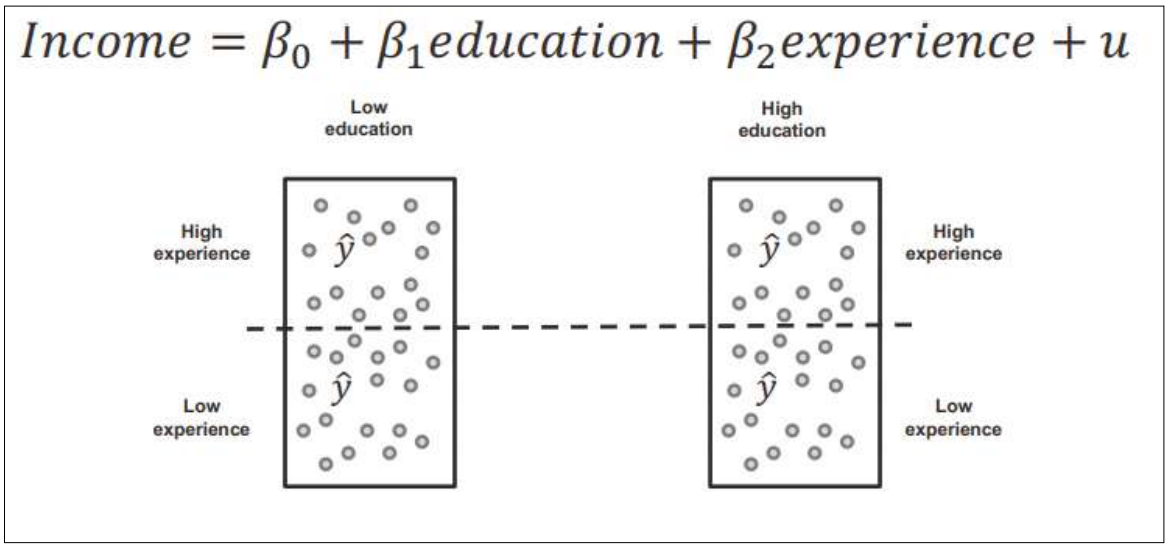
$b_2 = \beta_2$ = partial slope for X_2 on Y – β_2 indicates the change in Y for one unit change in X_2, controlling for X_1

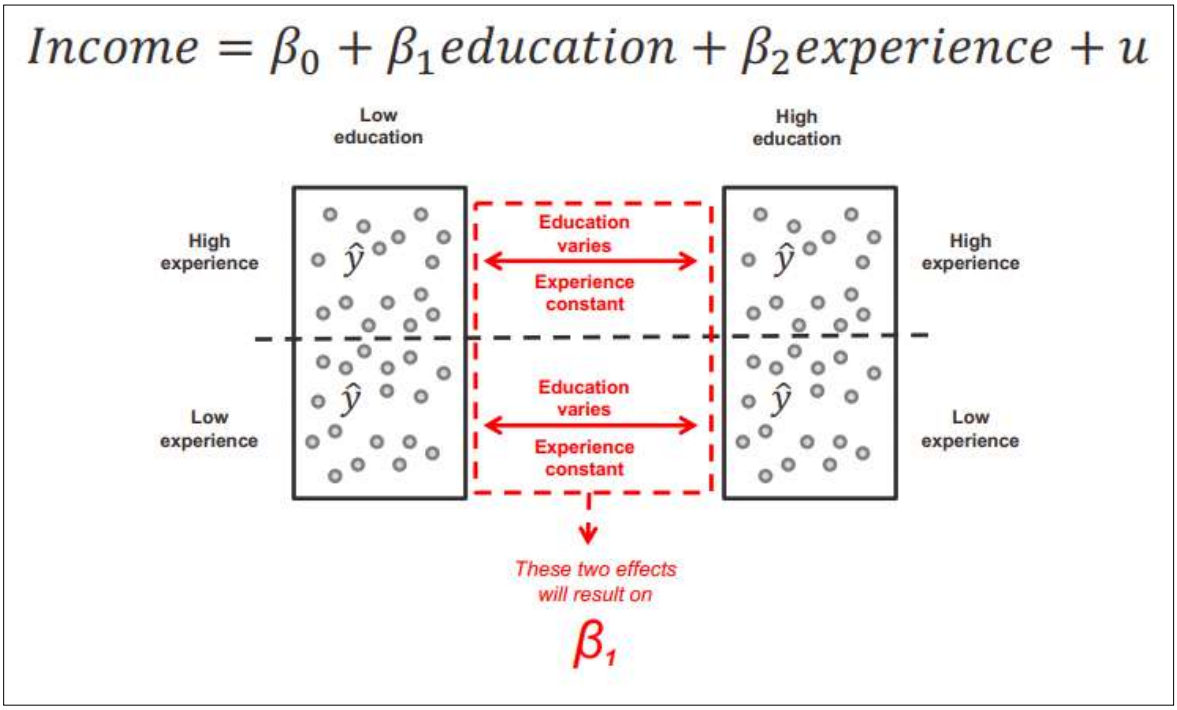
Partial slopes

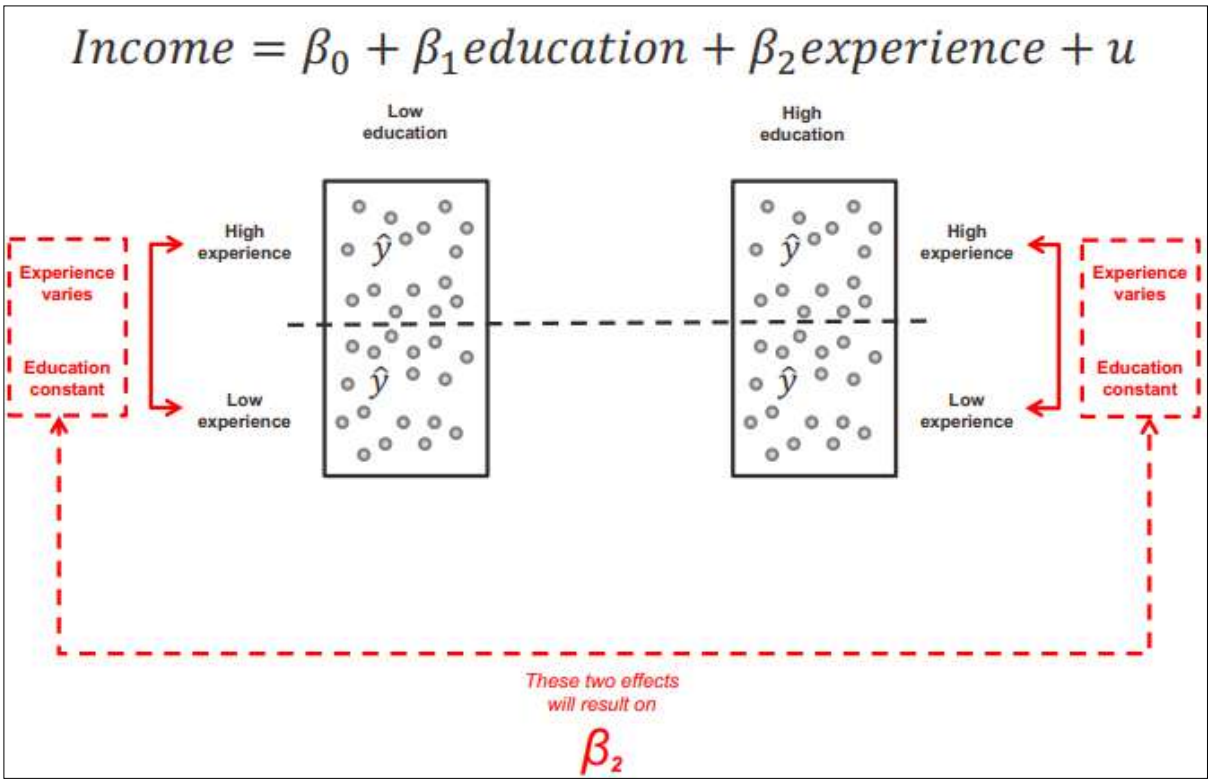
The partial slopes indicate the effect of each independent variable on Y, while controlling for the effect of the other independent variables. This control is called ceteris paribus

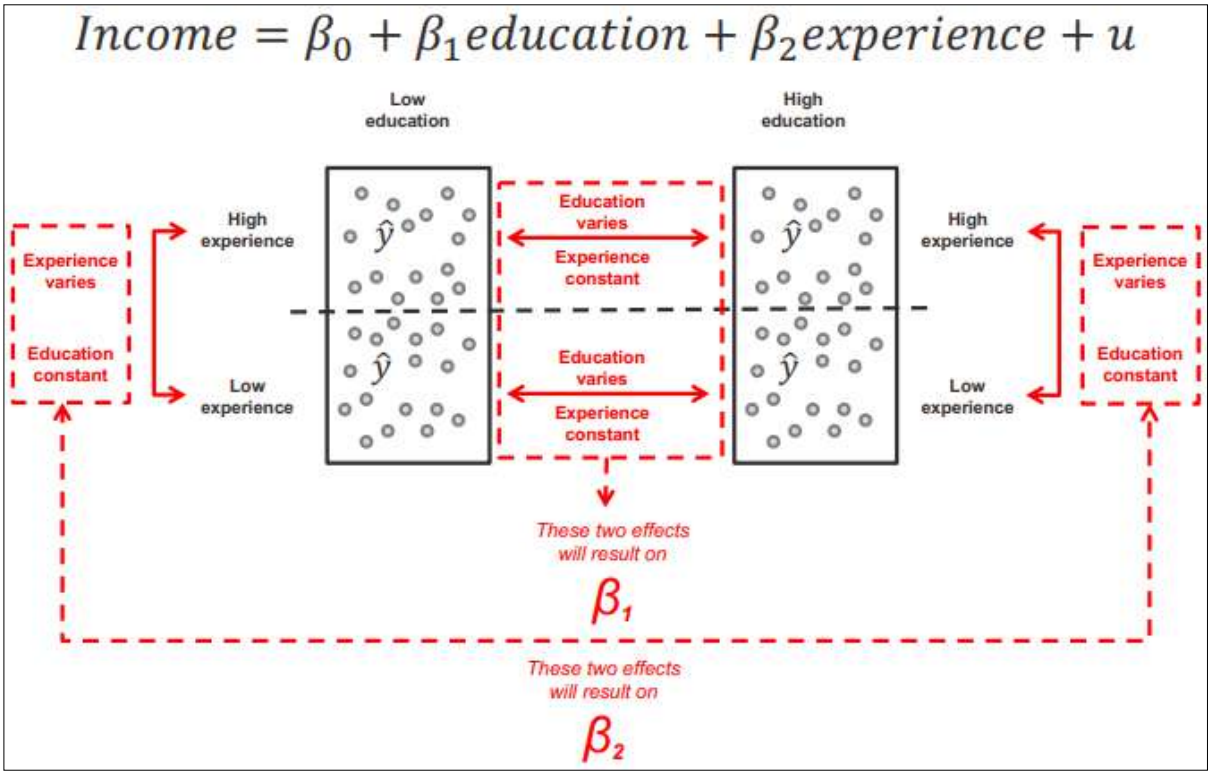
- Other things equal
- Other things held constant
- All other things being equal

Ceteris paribus









Interpretation of partial slopes

The partial slopes show the effects of the X's in their original units. These values can be used to predict scores on Y. Partial slopes must be computed before computing the Y intercept (β_0).

Formulas of partial slopes

$$b_1 = \beta_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$b_1 = \beta_1$ = partial slope of X_1 on Y
$b_2 = \beta_2$ = partial slope of X_2 on Y
s_y = standard deviation of Y
s_1 = standard deviation of the first independent variable (X_1)
s_2 = standard deviation of the second independent variable (X_2)
r_{y1} = bivariate correlation between Y and X_1
r_{y2} = bivariate correlation between Y and X_2
r_{12} = bivariate correlation between X_1 and X_2

Formula of constant

Once b_1 (β_1) and b_2 (β_2) have been calculated, use those values to calculate the Y intercept.

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2$$

Example

Husband's Housework	Number of Children	Husband's Education
$\bar{Y} = 3.3$	$\bar{X}_1 = 2.7$	$\bar{X}_2 = 13.7$
$s_y = 2.1$	$s_1 = 1.5$	$s_2 = 2.6$
Zero-Order Correlations		
$r_{y1} = 0.50$ $r_{y2} = -0.30$ $r_{12} = -0.47$		

Result and interpretation of b_1

$$b_1 = \beta_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_1 = \beta_1 = \left(\frac{2.1}{1.5}\right)\left(\frac{0.50 - (-0.30)(-0.47)}{1 - (-0.47)^2}\right) = 0.65$$

As the number of children in a dual-career household increases by one, the husband's hours of housework per week increases on average by 0.65 hours (about 39 minutes), controlling for husband's education

Result and interpretation of b_2

$$b_2 = \beta_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \left(\frac{2.1}{2.6}\right)\left(\frac{-0.30 - (0.50)(-0.47)}{1 - (-0.47)^2}\right) = -0.07$$

As the husband's years of education increases by one year, the number of hours of housework per week decreases on average by 0.07 (about 4 minutes), controlling for the number of children.

Result and interpretation of a

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2$$

$$a = \beta_0 = 3.3 - (0.65)(2.7) - (-0.07)13.7$$

$$a = \beta_0 = 2.5$$

With zero children in the family and a husband with zero years of education, that husband is predicted to complete 2.5 hours of housework per week on average.

Final regression equation

In this example, this is the final regression equation

$$Y = a + b_1X_1 + b_2X_2$$

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2$$

$$Y = 2.5 + (0.65)X_1 + (-0.07)X_2$$

$$Y = 2.5 + 0.65X_1 - 0.07X_2$$

Prediction

Use the regression equation to predict a husband's hours of housework per week when he has 11 years of schooling and the family has 4 children.

$$Y' = 2.5 + 0.65X_1 - 0.07X_2$$
$$Y' = 2.5 + (0.65)(4) + (-0.07)(11)$$
$$Y' = 4.3$$

Under these conditions, we would predict 4.3 hours of housework per week.

Standardized coefficients (b*)

Partial slopes ($b_1=\beta_1$; $b_2=\beta_2$) are in the original units of the independent variables

- This makes assessing relative effects of independent variables difficult when they have different units
- It is easier to compare if we standardize to a common unit by converting to Z scores

Compute beta-weights (b*) to compare relative effects of the independent variables

- Amount of change in the standardized scores of Y for a one-unit change in the standardized scores of each independent variable
    - While controlling for the effects of all other independent variables
- They show the amount of change in standard deviations in Y for a change of one standard deviation in each X

Formulas

Formulas for standardized coefficients

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right) = \beta_1^* = \beta_1\left(\frac{s_1}{s_y}\right)$$

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right) = \beta_2^* = \beta_2\left(\frac{s_2}{s_y}\right)$$

Example

Which independent variable, number of children (X1) or husband's education (X2), has the stronger effect on husband's housework in dual-career families?

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right) = (0.65)\left(\frac{1.5}{2.1}\right) = 0.46$$

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right) = (-0.07)\left(\frac{2.6}{2.1}\right) = -0.09$$

The standardized coefficient for number of children (0.46) is greater in absolute value than the standardized coefficient for husband's education (–0.09). Therefore, number of children has a stronger effect on husband's housework.

Standardized coefficients

Standardized regression equation

$$Z_y = a_z + b_1^* Z_1 + b_2^* Z_2$$

Where Z indicates that all scores have been standardized to the normal curve

The Y intercept will always equal zero once the equation is standardized

$$Z_y = b_1^* Z_1 + b_2^* Z_2$$

For the previous example

$$Z_y = (0.46)Z_1 + (-0.09)Z_2$$

Multiple correlation

The coefficient of multiple determination (R^2) measures how much of Y is explained by all of the X's combined. R^2 measures the percentage of the variation in Y that is explained by all of the independent variables combined. The coefficient of multiple determination is an indicator of the strength of the entire regression equation.

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$

- R^2 = coefficient of multiple determination
- r_{y1}^2 = zero-order correlation between Y and X_1
- $r_{y2.1}^2$ = partial correlation of Y and X_2, while controlling for X_1

Partial correlation of Y and X_2

Before estimating R2, we need to estimate the partial correlation of Y and X2 ($r_y^{2.1}$)

We need three correlations

- Between X_1 and Y: 0.50
- Between X_2 and Y: –0.30
- Between X_1 and X_2: –0.47

$$r_{y2.1} = \frac{r_{y2} - (r_{y1})(r_{12})}{\sqrt{1 - r_{y1}^2}\sqrt{1 - r_{12}^2}}$$

$$r_{y2.1} = \frac{(-0.30) - (0.50)(-0.47)}{\sqrt{1 - (0.50)^2}\sqrt{1 - (-0.47)^2}}$$

$$r_{y2.1} = -0.08$$

Result and interpretation

For this example, R^2 will tell us how much of husband's housework is explained by the combined effects of the number of children (X_1) and husband's education (X_2).

$$R^2 = r_{y1}^2 + r_{y2.1}^2(1 - r_{y1}^2)$$
$$R^2 = (0.50)^2 + (-0.08)^2(1 - 0.50^2)$$
$$R^2 = 0.255$$

Number of children and husband's education explain 25.5% of the variation in husband's housework.

Limitations

Multiple regression and correlation are among the most powerful techniques available to researchers. But powerful techniques have high demands. These techniques require:

– Every variable is measured at the interval-ratio level

– Each independent variable has a linear relationship with the dependent variable

– Independent variables do not interact with each other

– Independent variables are uncorrelated with each other

– When these requirements are violated (as they often are), these techniques will produce biased and/or inefficient estimates

– There are more advanced techniques available to researchers that can correct for violations of these requirements

---

4. Elementary multiple regression modeling techniques: Stepwise variable entry method, Path Analysis.

Stepwise Multiple Linear Regression Analysis

Stepwise regression analysis relates to evaluation of the relative efficiency of independent variables in explaining the dependent variable when the variables are added/deleted to the model, one by one, in several steps. It has two approaches: Forward and backward. In forward stepwise regression, we start with one most important variable followed by the second, third …….and the least important

variables. In backward approach we start with all the variables and keep on excluding variables one by one and proceed in the reverse direction until we reach the optimal position.

Need for Stepwise regression analysis

One of the important assumption of the tests of significance of OLS regression analysis is that independent or explanatory variables are independent of each other. This assumption is known as assumption of absence of multi-collinearity. Sometimes this assumption is violated and variables are found to be collinear where independent variables show significant inter-correlation among themselves. In such cases there is some overlap in the explanatory power of the two or more collinear variables. Higher the relationship between the variables higher will be the overlap. For example if a variable is explaining 40 % of the dependent variable and another variable related to it is explaining 30 %. When these two variables are taken together, they may explain 70% of the dependent variable provided both the independent variables are independent.

However, if they are collinear or correlated, they will explain less than 70%. Suppose these two variables together explain only 50 % of the dependent variable, it will be due to the fact that what second variable is contributing, 20 % of it has already been explained by the first variable. Second variable has now only 10 % contribution to explain in addition to first variable. In the absence of the first variable, however, second variable will have a higher explanatory power. In ordinary regression equation we will not have any idea about this complication arising due to the problem of multicollinearity.

Computer programmes have been developed to tackle this problem. Stepwise regression analysis is one such programme. In any regression model with some R2, if more variables are added the value of R2 will always increase, either the variable is positively or negatively related to the dependent variable. In stepwise approach of a regression analysis, independent variables are sequentially added to the model one by one, untilthe criterion of variable addition is not met. The sequence starts in such a manner that in first step it gives regression line with one independent variable choosing from all the independent variables the one which gives maximum R2. In the second step it adds a one more independent variable to the model which adds maximum value to the existing R2. Likewisein third step one more variable is added and so on. Every time the addition to R2 due to new variable will be less than the previous value and the value of F-ratio will also change. In stepwise regression analysis we can fix a criterion of adding a new variable. Generally it is done by choosing the probability level of the changed value of F due to the addition of new variable. In most of the cases, if the probability exceeds the fixed limit say 0.05, the variable is not added to the model. R2

(adjusted) is designed in such a way that with the increase of every new variable it will decrease unless the new variable causes a significant increase in the value of R2. In step wise regression analysis, we keep on allowing the addition of new variables until R2 (Adjusted) increases. After few steps though R2 will continue to rise, R2 (Adjusted) will start decreasing indicating the fact that addition to R2 is not big enough to be retained in the analysis.

Example

Declining sex ratio in India is a big concern of the society. There are a large number of factors behind it. In the following example, for the sake of simplicity in explanation, we have taken few of them and used a stepwise regression analysis to explain the variations in the "Sex Ratio" in the 50 districts across Madhya Pradesh for 2011, with the help of the following variables.

1. Sex Ratio (Female per thousand male) (V1).
2. Growth rate of population, 2001-11(in Percentage) (V2)
3. Levels of Literacy (in percentage) (V3)
4. Population Density per square kilo meter of area (V4)
5. Female work participation rate (in percentage) (V5)

Using SPSS when data (given in annexure) was subjected to the bivariate correlation and stepwise regression analysis, following results were obtained.

First, it gives the inter-correlation matrix of each variable with other variables given in Table1, given below. It also identifies the level of significance at which these coefficients of correlation are significant.

The table shows that, the dependent variable Sex Ratio (V1) has inter-significant negative correlation coefficient with the variables: growth rate of population, 2001-11 (V2) and Levels of Literacy (V3) and a significant negative correlation with the variable of Levels of Literacy (V3). It is found to be significant at 5% level of significance. The inter-correlation matrix also shows that the dependent variable has a strong positive relationship with the variable of female work participation rate (V5), significant at 1 % level of significance.

Note that:

1. Two tail test means that a value of coefficient of correlation could be ≠ 0 i.e. it could be greater than or less than 0.

2. One tail test will mean that coefficient of correlation could be either > 0 or <0 i.e. either greater than 0 or less than 0.

3. The diagonal elements of an inter-correlation matrix are always 1, indicating the correlation of a variable with itself is perfect and coefficient of correlation will be, therefore, 1.

Table 1

Inter correlation Matrix

		VAR00001	VAR00002	VAR00003	VAR00004	VAR00005
VAR00001	Pearson Correlation	1	-.141	-.354*	-.272	.840**
	Sig. (2-tailed)		.327	.012	.056	.000
	N	50	50	50	50	50
VAR00002	Pearson Correlation	-.141	1	-.335*	.456**	-.163
	Sig. (2-tailed)	.327		.017	.001	.258
	N	50	50	50	50	50
VAR00003	Pearson Correlation	-.354*	-.335*	1	.428**	-.538**
	Sig. (2-tailed)	.012	.017		.002	.000
	N	50	50	50	50	50
VAR00004	Pearson Correlation	-.272	.456**	.428**	1	-.516**
	Sig. (2-tailed)	.056	.001	.002		.000
	N	50	50	50	50	50
VAR00005	Pearson Correlation	.840**	-.163	-.538**	-.516**	1
	Sig. (2-tailed)	.000	.258	.000	.000	
	N	50	50	50	50	50

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

The inter-correlation matrix given above suggests reasonable justification for the choice of the explanatory or independent variables to explain the variations in the values of the dependent variable, Sex Ratio.

The above matrix of inter-correlations also suggests the overlap among the independent variables. The matrix shows inter-correlations among independent variables also. Fourth variable of the density of population (V4) has a strong positive relationship significant at 1% level of significance with the second variable of growth rate of population (V2) and third variable of literacy (V3) which has a strong significant positive relationship with the fifth variable of female work participation rate. Female work participation rate (V5) also has strong negative relationship with the density of population (V4).

The inter-correlation among independent variables suggests that there exist some multicollinearity among them and an ordinary regression analysis will not be the optimal regression equation. A stepwise regression is likely to give better results by excluding the redundant variables and retaining

only those which add a higher value to R2 as explained above. It will also give the order of the efficiency with which each of the independent variable explains the dependent variable of sex ratio.

Step wise regression analysis will give different models by adding independent variables one by one sequentially. The criterion to add the new variable is in terms of probability of its F value being less than 0.05 or 5% as is shown in Table 2 given below. We can also change the probability to 0.10 or 10% to allow more variables to enter into the analysis.

In the present example the result given in Table 2 given below shows that only two variables; female work participation rate (V5) and density of population (V4) are sufficient to be retained in the multiple regression analysis. Other variables; population growth rate (V2) and level of literacy (V3) are not found to explain much the variations in the sex ratio of the districts of Madhya Pradesh. Their part of explanation is already explained by first two variable; female work participation rate (V5).

Table 2
Criteria forChoosing the Independent Variables
Stepwise Regression Analysis

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	VAR00005	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	VAR00004	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: VAR00001

The summary results of the two steps of the regression model will follow in the computer output as given below in Table 3:

Table 3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.840[a]	.706	.700	23.09253
2	.861[b]	.742	.731	21.88012

a. Predictors: (Constant), VAR00005

b. Predictors: (Constant), VAR00005, VAR00004

Model 1 given by the first step shows that only one variable i.e. female work participation rate (V5) alone explains the sex ratio quite effectively. It explains 70.6 % variations of the sex ratio across 50 districts of Madhya Pradesh as per the data provided by the Census of India. The next variable which could be included in the model is the population density (V4) which could add to the explanatory power of the model only 3.6 % as the value of R2 could rise from 0.706 to 0.742 only. R2 (adjusted) could also rose from 0.700 to 0.731 only. Another two variables could not qualify the criterion of entering into the analysis due to multicollinearity.

Once the model is chosen, the main results follow. These include regression coefficient of the selected variables their standard errors, "t-statistics" and the level of their significance. These results are also given in Table 4 below.

Table 4
Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	811.164	12.030		67.430	.000
	VAR00005	3.712	.346	.840	10.744	.000
2	(Constant)	778.374	17.210		45.229	.000
	VAR00005	4.214	.382	.954	11.026	.000
	VAR00004	.063	.025	.220	2.543	.014

a. Dependent Variable: VAR00001

Table 4 given above show the regression coefficients in unstandardized form as well as in standardized form. An unstandardized coefficient relates to the data as provided in the computer input and also gives the value of the intercept as constant. Computer also converts the given data into their standard scores and gives corresponding regression coefficients as standardized coefficients. The purpose is to bring the data to a standard form of zero mean and unit standard deviations. In standardized form when the mean of all the variables is zero intercept is not given as it also become zero. Standard error and't- statistics' of the regression coefficient, however, in both the cases remain the same.

The results of the above table how that as it is a unit change in the employment to female will promote an increase of 4.214 rises in the sex ratio. Whereas, the density of population does not show much impact on the sex ratio. A change of one person per square km. will bring a change of only o.063 change in sex ratio.

It is important to note that female work participation rate varies with in a narrow range from 8.4 in Bhid to 52.9 in Dindori. Density of population has quite big range of variation from 855 in Bhopal to 94 in Annupur. These variations have been standardized by converting all the three variables into their standard scores. As a result the gap of 3.832 between the regression coefficients of unstandardized form (4.214 - 0.382 = 3.832) is reduced to 0.734 between the same in unstandardized form (0.954 – 0.220). The proportion of the two has also been reduce from 11.03 (= 4.214/ 0.382) to 4.33 (= 0.954/0.220).

Annexure I: Data for Stepwise Regression Analysis

S.N.	District	V1	V2	V3	V4	V5
1	Indore	928	32.88	80.87	841	20.9
2	Jabalpur	929	14.51	81.07	473	25.3
3	Sagar	893	17.63	76.46	232	28.9
4	Bhopal	918	28.62	80.37	855	19.6
5	Rewa	931	19.86	71.62	375	32.9
6	Satna	926	19.19	72.26	297	29.9
7	Dhar	964	25.6	59	268	40.2
8	Chhindwara	964	13.07	71.16	177	36.6
9	Gwalior	864	24.5	76.65	446	14.5
10	Ujjain	955	16.12	72.34	326	33.8
11	Morena	840	23.44	71.03	394	16.8
12	West Nimar	965	22.85	62.7	233	40.9
13	Chhattarpur	883	19.51	63.74	203	32.7
14	Shivpuri	877	22.76	62.55	171	34.5

15	Bhind	837	19.21	75.26	382	8.4
16	Balaghat	1021	13.6	77.09	184	47
17	Betul	971	12.92	68.9	157	42.9
18	Dewas	942	19.53	69.35	223	38.4
19	Rajgarh	956	23.26	61.21	251	41.8
20	Shajapur	938	17.2	69.09	244	39.1
21	Vidisha	896	20.09	70.53	286	21.6
22	Ratlam	971	19.72	66.78	255	39
23	Tikamgarh	901	20.13	61.43	157	36.9
24	Barwani	982	27.57	49.08	242	41.9
25	Seoni	982	18.22	72.12	157	42.4
26	Mandsaur	963	13.24	71.78	199	42.9
27	Raisen	901	18.35	72.98	178	23.4
28	Sehore	918	21.54	70.06	261	28.7
29	East Nimar	943	21.5	66.39	178	38.6
30	Katni	952	21.41	71.98	173	31
31	Damoh	910	16.63	69.73	185	34.3
32	Guna	912	26.97	63.23	208	30.3
33	Hoshangabad	914	14.49	75.29	232	22.4
34	Singrauli	920	28.05	60.41	213	34.6
35	Sidhi	957	23.72	64.43	172	32.2

36	Narsimhapur	920	14.01	75.69	182	29.5
37	Shahdol	974	17.39	66.67	285	38.1
38	Mandla	1008	17.97	66.87	142	49
39	Jhabua	990	30.7	43.3	285	48.9
40	Panna	905	18.67	64.79	142	32.4
41	Ashoknagar	904	22.66	66.42	271	20.8
42	Neemuch	954	13.77	70.8	221	42.9
43	Datia	873	18.46	72.63	200	26
44	Burhanpur	951	19.37	64.36	229	31.5
45	Anuppur	976	12.3	67.88	94	37.6
46	Alirajpur	1011	19.45	36.1	104	48.6
47	Dindori	1002	21.32	63.9	158	52.9
48	Sheopur	901	22.94	57.43	171	28.7
49	Umaria	950	24.96	65.89	158	36.3
50	Harda	935	20.25	72.5	171	26.9

Source: Census of India, 2011

1. Sex Ratio (Female per thousand male) (V1).
2. Growth rate of population, 2001-11(in Percentage) (V2)
3. Levels of Literacy (in percentage) (V3)
4. Population Density per square kilo meter of area (V4)
5. Female work participation rate (in percentage) (V5)

5. Model building techniques

Model building and generalization of geographical facts are important and significant issues which have been increasingly addressing with the use of 'deductive' explanation in the geographical studies. As geography deals with the spatial distribution of geographical phenomena, the construction of model is primarily concerned with three dimensions of geographical elements, called 3-Ds: *density, distance* and *division*. The detail elaboration of these elements in geographical context is given below.

(a) **The *density*** of geographical phenomena, which refers to the size- area ratio - the amount and intensity of an attribute, varies across area over time.

(b) **The *Distance*** (Networking and cost involvement): Analysis of location of a point/area becomes too important to make it more elaborative in its physical (natural resource base) and socio-economic (functional base of location of an attribute) contexts now-a-days when agglomeration/dispersion economies are 'location-based' and dominate the activities in a region.

(c) **The *Division*:** We are not indicating here the political division as country boundaries, but we are more concerned with the spatial organization of geographic entities and planning within a political unit (country). Space does not have plane surface; division of space is required to study features on account of spatial variation of economic activities, the geographical interpretation of attributes/variables follows its areal/regional characteristics. It involves a technique of division of space (physical or socio-economic).

Model Building:

- **Model building is a Procedure for the Development of a System**
- **Model is a Tool-box**
- **Models are used to Systematise Raw Data Set**

Aspects of Model-Building:

One must be careful at the time of construction of model when model are made for a specific purpose. It (model construction) requires quantified variables, controlled variables with relaxed ones (dependent-independent variables), aggregated view of variables, time concept also to be treated if

required, tools and techniques used, relevant data available and inferences of main traits for generalization and development of laws. Such steps of the procedure of model-construction are discussed below in detail:

Step-1: Purpose behind Model-Building

As per the research problems and issues taken up for solution, we have to determine the purpose of model-building to prepare an alternative strategy for the solution and to provide deductive explanation.

Step-2: Quantification of Variables

Since one has to deal with the research problem in quantitative manner, the field work and data collection of objects are major tasks to frame research design. Imposition of hypothesis and prove it are dimensions which compel us for selection of attributes and variables. Variables are to be in quantitative manner in the tabulated form for further operation. Scaling of selected variables is primary task for model builders to look into the problems.

Step-3: Controls on Variables

When model-builder is going to analyse the system, 'causality' is main principle keeping in mind that the answer of our research problem will be given logically. Cause-effect relationship is to establish by classifying variables into two categories, namely, the effect variables (dependent or base) and cause-variables (independent or factor). Say for example, crop yield is controlled by rainfall; we wish to establish relationship between crop yield and rainfall. Crop yield is dependent (variable) on rainfall (Independent).

Step-4: Time Concept

Geographical studies follow space-time concept. One must include time as factor, if required, in the model construction. For example, if the development of model is towards the analysis of spatial development of socio-economy of a region, a set of four processes of development; barrier, hierarchy, network and contiguity considering as elements of space are taken into account with their different stages of time (Fig. 1). Sometimes, we do not consider time dimension if our study is based on one-time scale and our data collection schedule do not include time dimension. The dimension of space in model building is used in a variety of ways. A modern quantitative way of understanding spatial pattern of an attribute has many steps like preparation of a map of the surveyed area and

showing location data of activities collected from samples (called spatial data). Such spatial data are depicted into geometric manner into zone system and particular zone is notated by number or algebraic symbols (Fig- 2).

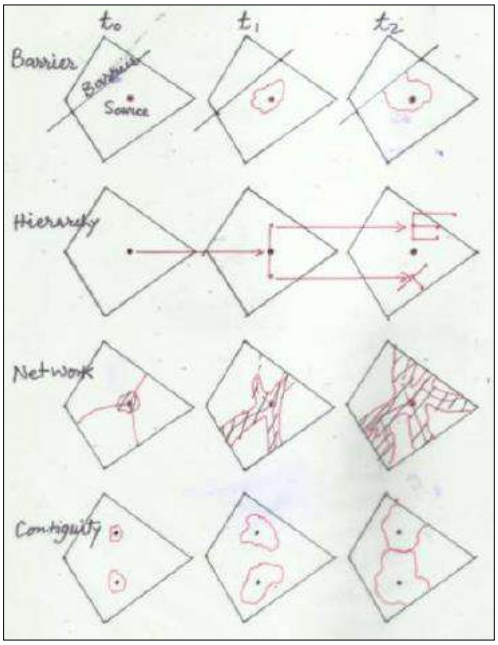


Fig.-1: Spatio-Temporal View of the Development of Economic Landscape

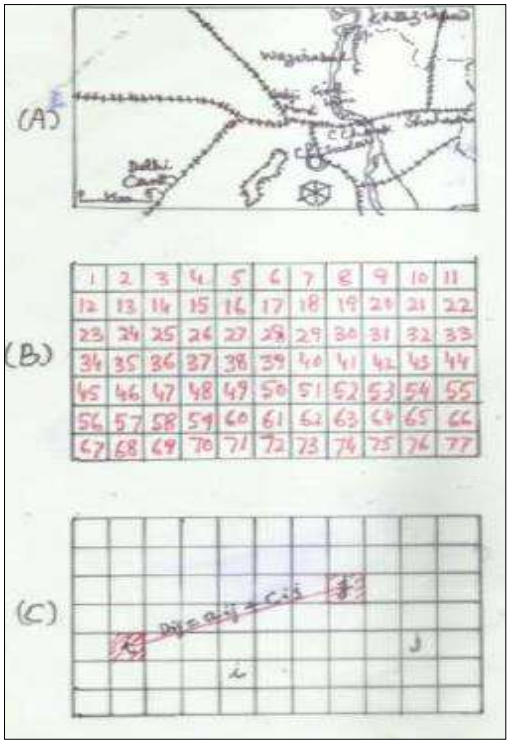


Fig-2: Generation of Spatial data Collected from the Study Area and its Depiction into Geometric Manner: (A) real Map of the part of Delhi, (B) its Geometric Grids, and (C) Algebraic notation by letters

Developing a good notation mathematically, a map-area is to show in coordinates following GIS for accurate and speedy calculation and depiction of spatial features of variables. Such type of data depiction of objects given on map into its geometric form provides a sound base to reach closer to reality. The data generated into a coordinate form can easily be used for the purpose through GIS tool.

Step-5: Techniques used for Establishing Cause-effect Relationship:

Choice of technique is based on the following three dimensions of its uses:

(a) ***Assumption*****–** It focuses direct attention towards the objective of research. When one constructs the model, some interrelated statements related to research problem are to assume based on speculation.

(b) ***Conditions*****-**That prepares a scale of the variability of variable. If we know the extreme conditions of variables, it must help in speculation of relationship between variables.

(c) ***Axioms*****-** That are well-established facts which are used sometimes somewhere for establishing new relationships of facts.

(d) ***Relationship prevalent in the system*****-** Relationship between two sets of variables acts as integral part of the system that we wish to develop. After assembling the parts (tools) of the system and knowing their relationships, we reach to a point from where we may obtain the 'output' of a model. If we are able to study the system which has parts and their assemblage, we are called 'mechanic' who knows about the system. But knowing simply about the outer part of the system, it indicates the operator of the system called 'driver'. Model builder is a good mechanic.

(e) ***Caliberation and Validation*****-** It is the step which is useful for simulation (prediction) of events which we need to analyse. Before it, validity testing of a model is needed. As per the use of model situation, there is a variety of 'deviation-based' validity testing methods to use the significance level of model error. What is deviation based testing? Models deviation from observed data is based on its error term, For example, a linear model which establishes relationship between Y (dependent) and X (independent) variables has error term, e, as:

$$Yo = a+bX+e \text{ and}$$
$$Yc = a+bX, \text{ so}$$
$$e = (Yo-Yc),$$

Where, Yo= observed and Yc= computed values of a variable, Y. It is expressed by substituting Yc function in Yo equation.

The error of a model is tested in many ways. Statistical testing of a model is particularly based on two types of variations: the explained variation which is expressed by degree of determinants, R2, and, secondly, the unexplained variations, that is expressed in its degree of significance measured by the student's t-test. However, for different purposes, the ratio of the 'Sum of Squared Deviation' with 'mean sum' is used.

Step-6: Simulation

When model is tested and ready to use, we can simulate the results of as event on the basis of generalisation that is the outcome of the model. Speculation of event and development of law based on generalization are major achievements of model-building. Note that this procedure of testing validity is imposed hypothesis through the use of model and explanation of real-world situation is primarily dependent on 'deductive' explanation.

There are two approaches of writing a report on research issues/problems; the 'pattern' approach when data of attribute(s) are collected, processed and classified to understand the facts and, second is the 'system' approach through which we wish to understand the process and the working of the internal assemblages of the parts of a system and to know about its output. Model building is the main aspect of the second approach through which a deductive explanation and generalization of facts is proceeded by using hypothesis/hunches/assumptions/specific conditions and so on. Such explanation leads us towards construction of law and theories.

****

PAPER- GEO 296: REMOTE SENSING AND COMPUTER APPLICATION

GEO 296.1: PRINCIPLES OF REMOTE SENSING AND AERIAL PHOTOGRAPHY

1. Physics of Remote Sensing: Electro Magnetic Radiation (EMR), Radiation laws (wavelength frequency- energy relationship of EMR numerical problems).

2. Satellite System: Keplers's Laws, Major-Semi-major axis, eccentricity, velocity (Numerical problems).

3. Satellite Sensors: Concept of IFOV, resolution and determination of pixel size, referencing scheme of satellite system (path/row calculation).

4. Basics of Aerial Photograph: Basics geometry of aerial photograph, determination of scale and height, Distortions, Image parallax, Relief displacement.

5. Stereoscopy and Aerial Photo Interpretation: Stereoscopy, Pseudoscopy, Mirror Stereoscope, mosaic, edge information, mapping of Physical and Cultural features with the Air photo interpretation keys: shape, size, pattern, tone, texture, shadow, site and associations.

1. Physics of Remote Sensing: Electro Magnetic Radiation (EMR), Radiation laws (wavelength frequency- energy relationship of EMR numerical problems).

Remote Sensing

"Remote sensing is the science (and to some extent, art) of acquiring information about the Earth's surface without actually being in contact with it. This is done by sensing and recording reflected or

emitted energy and processing, analyzing, and applying that information." In much of remote sensing, the process involves an interaction between incident radiation and the targets of interest. This is exemplified by the use of imaging systems where the following seven elements are involved. Note, however that remote sensing also involves the sensing of emitted energy and the use of non-imaging sensors.

Advantages and Limitations

After understanding the concept of remote sensing let us now look at its advantages and limitations. Remote sensing has several advantages as listed below:

• It provides a synoptic view.

• It is a cost effective means of data collection.

• It can provide data in wavelengths beyond the sensing capability of human eye.

• It can acquire data of inaccessible areas.

• It is an unobtrusive means of data collection which does not change characteristics of the object or phenomenon being observed.

• It provides historical data sets which is useful to know characteristics of an object in a given point of time in past.

Advantages of remote sensing have been oversold and despite having several advantages remote sensing has following limitations:

• It is sometimes found that appropriate data is not available or easily acquired particularly in the tropical regions where cloud cover obstructs acquisition of image because not all sensors can 'see' through cloud.

• Remote sensing equipments can become uncelebrated with time resulting into errors in data collected.

Stages of Remote Sensing

1**. Energy Source or Illumination (A) –** the first requirement for remote sensing is to have an energy source which illuminates or provides electromagnetic energy to the target of interest.

2. Radiation and the Atmosphere (B) – as the energy travels from its source to the target, it will come in contact with and interact with the atmosphere it passes through. This interaction may take place a second time as the energy travels from the target to the sensor.

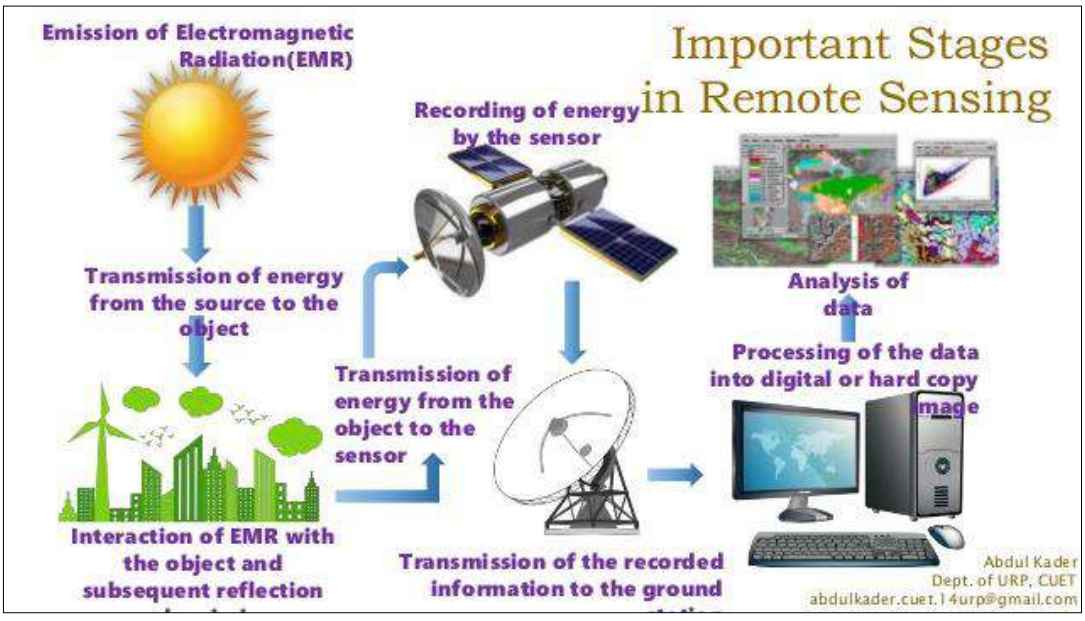
3. Interaction with the Target (C) - once the energy makes its way to the target through the atmosphere; it interacts with the target depending on the properties of both the target and the radiation.

4. Recording of Energy by the Sensor (D) - after the energy has been scattered by, or emitted from the target, we require a sensor (remote - not in contact with the target) to collect and record the electromagnetic radiation.

5. Transmission, Reception, and Processing (E) - the energy recorded by the sensor has to be transmitted, often in electronic form, to a receiving and processing station where the data are processed into an image (hardcopy and/or digital).

6. Interpretation and Analysis (F) - the processed image is interpreted, visually and/or digitally or electronically, to extract information about the target which was illuminated.

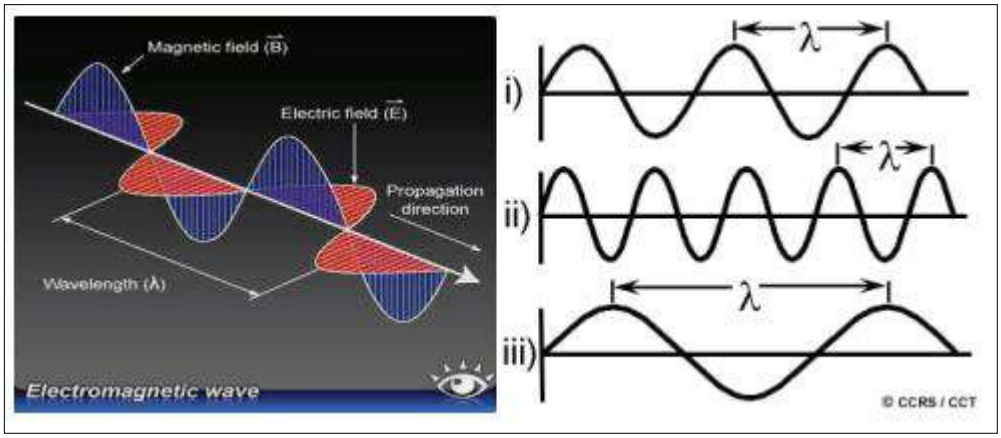
7. Application (G) - the final element of the remote sensing process is achieved when we apply the information we have been able to extract from the imagery about the target in order to better understand it, reveal some new information, or assist in solving a particular problem. These seven elements comprise the remote sensing process from beginning to end.



Electromagnetic Radiation

In physics, electromagnetic radiation refers to the waves of the electromagnetic field, propagating through space, carrying electromagnetic radiant energy. It includes radio waves, microwaves, infrared, light, ultraviolet, X-rays, and gamma rays. All of these waves form part of the electromagnetic spectrum.

All electromagnetic radiation has fundamental properties and behaves in predictable ways according to the basics of wave theory. Electromagnetic radiation consists of an electrical field (E) which varies in magnitude in a direction perpendicular to the direction in which the radiation is travelling, and a magnetic field (M) oriented at right angles to the electrical field. Both these fields travel at the speed of light (c). Two characteristics of electromagnetic radiation are particularly important for understanding remote sensing. These are the wavelength and frequency.



Wavelength and Frequency

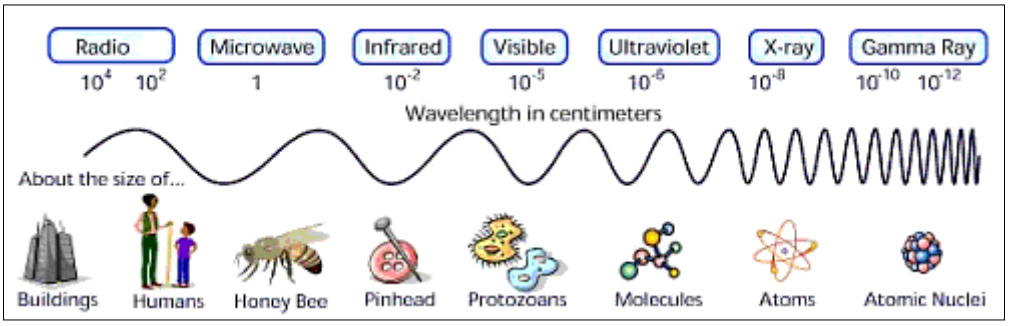
The wavelength is the length of one wave cycle, which can be measured as the distance between successive wave crests. Wavelength is usually represented by the Greek letter lambda (λ). Wavelength is measured in metres (m) or some factor of metres such as nanometres (nm, 10-9 metres), micrometres (µm, 10-6 metres) (µm, 10-6 metres) or centimetres (cm, 10-2 metres). Frequency refers to the number of cycles of a wave passing a fixed point per unit of time. Frequency is normally measured in hertz (Hz), equivalent to one cycle per second, and various multiples of hertz. Wavelength and frequency are related by the following formula:

$$c = \lambda \nu$$

where:

λ = wavelength (m)
ν = frequency (cycles per second, Hz)
c = speed of light (3×10^8 m/s)

Therefore, the two are inversely related to each other. The shorter the wavelength, the higher the frequency. The longer the wavelength, the lower the frequency. Understanding the characteristics of electromagnetic radiation in terms of their wavelength and frequency is crucial to understanding the information to be extracted from remote sensing data. Next we will be examining the way in which we categorize electromagnetic radiation for just that purpose.



Electromagnetic Spectrum

The electromagnetic spectrum ranges from the shorter wavelengths (including gamma and x-rays) to the longer wavelengths (including microwaves and broadcast radio waves). There are several regions of the electromagnetic spectrum which are useful for remote sensing.

For most purposes, the ultraviolet or UV portion of the spectrum has the shortest wavelengths which are practical for remote sensing. This radiation is just beyond the violet portion of the visible wavelengths, hence its name. Some Earth surface materials, primarily rocks and minerals, fluoresce or emit visible light when illuminated by UV radiation.

The light which our eyes - our "remote sensors" - can detect is part of the visible spectrum. It is important to recognize how small the visible portion is relative to the rest of the spectrum. There is a lot of radiation around us which is "invisible" to our eyes, but can be detected by other remote sensing instruments and used to our advantage. The visible wavelengths cover a range from approximately 0.4 to 0.7 µm. The longest visible wavelength is red and the shortest is violet. Common wavelengths of what we perceive as particular colours from the visible portion of the

spectrum are listed below. It is important to note that this is the only portion of the spectrum we can associate with the concept of colours.

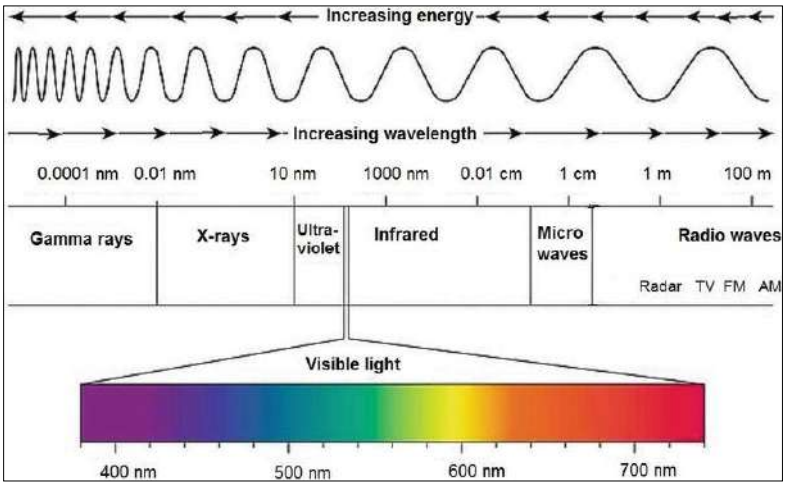
- Violet: 0.4 - 0.446 µm
- Blue: 0.446 - 0.500 µm
- Green: 0.500 - 0.578 µm
- Yellow: 0.578 - 0.592 µm
- Orange: 0.592 - 0.620 µm
- Red: 0.620 - 0.7 µm

Blue, green, and red are the primary colours or wavelengths of the visible spectrum. They are defined as such because no single primary colour can be created from the other two, but all other colours can be formed by combining blue, green, and red in various proportions. Although we see sunlight as a uniform or homogeneous colour, it is actually composed of various wavelengths of radiation in primarily the ultraviolet, visible and infrared portions of the spectrum. The visible portion of this radiation can be shown in its component colours when sunlight is passed through a prism, which bends the light in differing amounts according to wavelength.

The next portion of the spectrum of interest is the infrared (IR) region which covers the wavelength range from approximately 0.7 µm to 100 µm - more than 100 times as wide as the visible portion! The infrared region can be divided into two categories based on their radiation properties - the reflected IR, and the emitted or thermal IR.

Radiation in the reflected IR region is used for remote sensing purposes in ways very similar to radiation in the visible portion. The reflected IR covers wavelengths from approximately 0.7 µm to 3.0 µm. The thermal IR region is quite different than the visible and reflected IR portions, as this energy is essentially the radiation that is emitted from the Earth's surface in the form of heat. The thermal IR covers wavelengths from approximately 3.0 µm to 100 µm.

The portion of the spectrum of more recent interest to remote sensing is the microwave region from about 1 mm to 1 m. This covers the longest wavelengths used for remote sensing. The shorter wavelengths have properties similar to the thermal infrared region while the longer wavelengths approach the wavelengths used for radio broadcasts.



The electromagnetic spectrum

Radiation Laws

There are some important laws of radiation that describe the various characteristics of radiation.

1. Kirchoff's Law: Kirchhoff's Law states that "the absorptivity (a) of a substance for radiation of a specific wavelength is equal to its emissivity for the same wavelength"

2. Stefan-Boltzman's Law: The Stefan-Boltzman's law states that the intensity of radiation emitted by a radiating body is proportional to the fourth power of the absolute temperature of that body.

1. Kirchhoff's Law

Kirchhoff, a physicist, demonstrated in 1860 that a body acts as a perfect blackbody if,

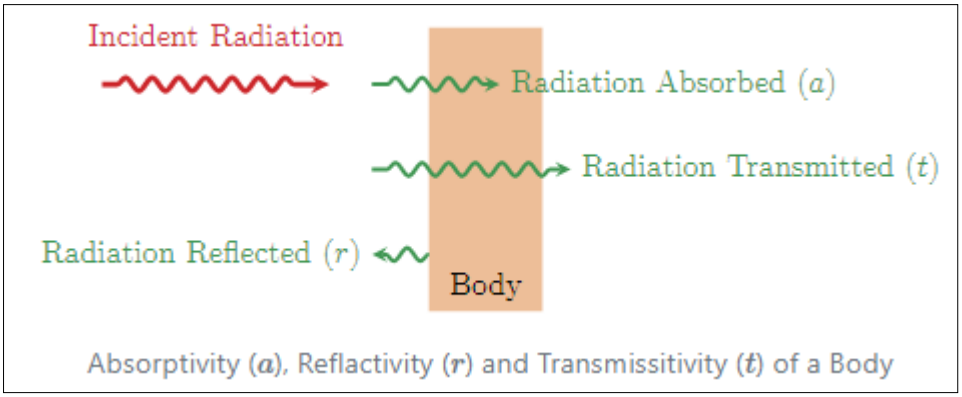
• The sides of the body are maintained at a constant absolute temperature (temperature of blackbody)

• A very small hole in comparison to the dimensions of the body is made in the body itself.

All bodies radiate energy in the form of photons. When these photons reach another surface, they may either be absorbed, reflected or transmitted. The behaviour of a surface with incident radiation is described by the following quantities:

- absorptivity (a) is the fraction of incident radiation absorbed
- reflectivity (r) is the fraction of incident radiation reflected
- transmissivity (t) is the fraction of incident radiation transmitted.

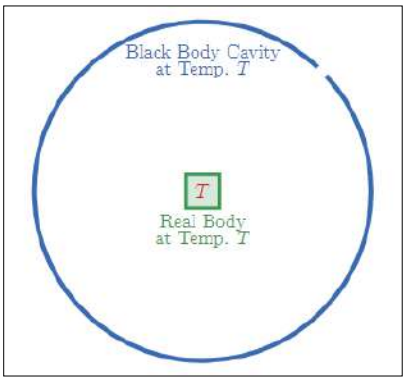
The energy conservation gives a + r + t = 1. For opaque objects, transmissivity t = 0and hence a + r = 1. A blackbody has absorptivity a = 1and absorbs all radiation incident on it. Its reflectivity is r = 0.



Real bodies radiate less effectively than black bodies. The measurement of this is the emissitivity (e) defined by

$$e = E/E_b$$

Where, E is radiated power per unit area from the real body at temperature T, and E_b is radiated power per unit area from a black-body at the same temperature T. The emissivity of the blackbody is 1 and that of a real body is between 0 and 1.



Consider a small real body in thermal equilibrium with its surrounding blackbody cavity i.e., $T_{body} = T_{cavity}$. The power emitted per unit area of the blackbody (cavity) is E_b. Thus, incident power per unit area of the real body is E_b. The power absorbed per unit area of the real body is aE_b. The power emitted per unit area of the real body is equal to $E = eE_b$. Since real body is in thermal equilibrium with its surrounding, energy balance gives $eE_b = aE_b$ i.e, a = e. The relation a = e is known as Kirchhoff's Law of radiation. It implies that good radiators are good absorbers. Note that Kirchhoff's law is valid only when body is in thermal equilibrium with its surrounding.

2. Stefan Boltzmann Law

According to Stefan Boltzmann law, the amount of radiation emitted per unit time from an area A of a black body at absolute temperature T is directly proportional to the fourth power of the temperature.

$$u = sAT^4 \ldots\ldots (1)$$

Where,

s is Stefan's constant = 5.67×10^{-8} $W/m^2 k^4$

A body which is not a black body absorbs and hence emit less radiation, given by equation (1)

$$\text{For such a body, } u = e\,\sigma\, AT^4 \ldots\ldots\ldots (2)$$

Where, e = emissivity (which is equal to absorptive power) which lies between 0 and 1.

With the surroundings of temperature T_0, net energy radiated by an area A per unit time.

$$\Delta u = u - u_o = e\sigma A\,[T^4 - T_0^4] \ldots\ldots (3)$$

Stefan Boltzmann Law relates the temperature of the blackbody to the amount of the power it emits per unit area. The law states that;

"The total energy emitted/radiated per unit surface area of a blackbody across all wavelengths per unit time is directly proportional to the fourth power of the black body's thermodynamic temperature."

$$\mathbf{\Rightarrow \varepsilon = \sigma T^4}$$

Derivation of Stefan Boltzmann Law

The total power radiated per unit area over all wavelengths of a black body can be obtained by integrating Plank's radiation formula. Thus, the radiated power per unit area as a function of wavelength is:

$$\frac{dP}{d\lambda}\frac{1}{A} = \frac{2\pi hc^2}{\lambda^5\left(e^{\frac{hc}{\lambda kT}}-1\right)}$$

Where,

- P is Power radiated.
- A is the surface area of a blackbody.
- λ is the wavelength of emitted radiation.
- h is Planck's constant
- c is the velocity of light
- k is Boltzmann's constant
- T is temperature.

On simplifying Stefan Boltzmann equation, we get:

$$\frac{d\left(\frac{P}{A}\right)}{d\lambda} = \frac{2\pi hc^2}{\lambda^5\left(e^{\frac{hc}{\lambda kT}}-1\right)}$$

On integrating both the sides with respect to λ and applying the limits we get;

$$\int_0^\infty \frac{d\left(\frac{P}{A}\right)}{d\lambda} = \int_0^\infty \left[\frac{2\pi hc^2}{\lambda^5\left(e^{\frac{hc}{\lambda kT}}-1\right)}\right] d\lambda$$

The integrated power after separating the constants is:

$$\frac{P}{A} = 2\pi hc^2 \int_0^\infty \left[\frac{d\lambda}{\lambda^5\left(e^{\frac{hc}{\lambda kT}}-1\right)}\right] \quad -(1)$$

This can be solves analytically by substituting:

$$x = \frac{hc}{\lambda kT}$$

$$\text{Therefore, } dx = -\frac{hc}{\lambda^2 kT} d\lambda$$

$$\Rightarrow h = \frac{x\lambda kT}{c}$$

$$\Rightarrow c = \frac{x\lambda kT}{h}$$

$$\Rightarrow d\lambda = -\frac{\lambda^2 kT}{hc} dx$$

As a result of substituting them in equation (1)

$$\Rightarrow \frac{P}{A} = 2\pi \left(\frac{x\lambda kT}{c}\right) \left(\frac{x\lambda kT}{h}\right)^2 \int_0^\infty \left[\frac{\left(-\frac{\lambda^2 kT}{hc}\right) dx}{e^x - 1}\right]$$

$$= 2\pi \left(\frac{x^3 \lambda^5 k^4 T^4}{h^3 c^2 \lambda^5}\right) \int_0^\infty \left[\frac{dx}{e^x - 1}\right]$$

$$= \frac{2\pi (kT)^4}{h^3 c^2} \int_0^\infty \left[\frac{x^3}{e^x - 1}\right] dx$$

The above equation can be comparable to the standard form of integral:

$$\int_0^\infty \left[\frac{x^3}{e^x - 1}\right] dx = \frac{\pi^4}{15}$$

Thus, substituting the above result we get,

$$\frac{P}{A} = \frac{2\pi (kT)^4}{h^3 c^2} \frac{\pi^4}{15} \Rightarrow \frac{P}{A} = \left(\frac{2k^4 \pi^5}{15 h^3 c^2}\right) T^4$$

On further simplifying we get,

$$\Rightarrow P/A = \sigma T^4$$

Thus, we arrive at a mathematical form of Stephen Boltzmann law:

$$\Rightarrow \varepsilon = \sigma T^4$$

Where,

ε = P/A

This quantum mechanical result could efficiently express the behaviour of gases at low temperature that classical mechanics could not predict!

$$\sigma = \left(\frac{2k^4\pi^5}{15h^3c^2}\right) = \left(5.670 \times 10^8 \ \frac{watts}{m^2K^4}\right)$$

Problems on Stefan Boltzmann Law

Example: A body of emissivity (e = 0.75), the surface area of 300 cm^2 and temperature 227 ºC are kept in a room at temperature 27 ºC. Using the Stephens Boltzmann law, calculate the initial value of net power emitted by the body.

Using equation (3);

$P = rsA (T^4 – T_0^4)$

$= (0.75) (5.67 \times 10^{-8} W/m^2 – k^4) (300 \times 10^{-4} m^2) \times [(500 K)^4 – (300 K)^4]$

= 69.4 Watts

Example 2: A hot black body emits the energy at the rate of 16 J m^{-2} s^{-1} and its most intense radiation corresponds to 20,000 Å. When the temperature of this body is further increased and it's most intense radiation corresponds to 10,000 Å, and then finds the value of energy radiated in Jm^{-2} s^{-1}.

Solution:

Wein's displacement law is, $l_m.T = b$

i.e. $T\mu [1/l_m]$

Here, l_m becomes half, the Temperature doubles.

Now from Stefan Boltzmann Law, $e = sT^4$

$e_1/e_2 = (T1/T2)^4$

$\Rightarrow e_2 = (T2/T1)^4 . e_1 = (2)^4 . 16$

$= 16.16 = 256$ J m^{-2} s^{-1}

The Wien and Stefan-Boltzmann Laws

The behaviour of blackbody radiation is described by the Planck's law. From the Planck's law, one can derive two other radiation laws i.e. the Stefan-Boltzmann law and the Wien's displacement law. These two laws illustrated below are very important in remote sensing to understand characteristics of EMR:

Stefan-Boltzmann law defines relationship between total emitted radiation (E) and temperature and is expressed as:

$$E = \sigma T^4$$

Where,

E = radiant energy per surface unit measured in Watts m-2 leaving a blackbody

ó = 5.6697 × 10–5 (Watts m–2 k–4 is the Stefan-Boltzmann constant, and

T = absolute temperature of the blackbody in Kelvin (K).

The Wien's displacement law defines the relationship between the wavelength of the radiation emitted and the temperature of the object and is expressed as:

$$ë_{max} =$$

Where,

$ë_{max}$ is the wavelength at which radiance is maximum (unit of the ë is in Angstroms), and

T is the absolute temperature in Kelvin (K).

The Wien's Displacement law gives the wavelength of the peak of the radiation distribution, while the Stefan-Boltzmann law gives the total energy being emitted at all wavelengths by the blackbody (which is the area under the Planck's law curve). Thus, the Wien's law explains the shift of the peak to shorter wavelengths as the temperature increases, while the Stefan-Boltzmann law explains the growth in the height of the curve as the temperature increases. Notice that this growth is very large, since it varies as the fourth power of the temperature.

---

2. Satellite System: Keplers's Laws, Major-Semi-major axis, eccentricity, velocity (Numerical problems).

Kepler's Law

Kepler's Three Laws

In the early 1600s, Johannes Kepler proposed three laws of planetary motion. Kepler was able to summarize the carefully collected data of his mentor - Tycho Brahe - with three statements that described the motion of planets in a sun-centered solar system. Kepler's efforts to explain the underlying reasons for such motions are no longer accepted; nonetheless, the actual laws themselves are still considered an accurate description of the motion of any planet and any satellite.

Kepler's three laws of planetary motion can be described as follows:

- The path of the planets about the sun is elliptical in shape, with the center of the sun being located at one focus. (The Law of Ellipses)
- An imaginary line drawn from the center of the sun to the center of the planet will sweep out equal areas in equal intervals of time. (The Law of Equal Areas)
- The ratio of the squares of the periods of any two planets is equal to the ratio of the cubes of their average distances from the sun. (The Law of Harmonies)

Motion is always relative. Based on the energy of the particle under motion, the motions are classified into two types:

- **Bounded Motion**
- **Unbounded Motion**

In bounded motion, the particle has negative total energy (E<0) and has two or more extreme points where the total energy is always equal to the potential energy of the particle i.e the kinetic energy of the particle becomes zero.

For eccentricity 0≤ e <1, E<0 implies the body has bounded motion. A circular orbit has eccentricity e = 0 and elliptical orbit has eccentricity e < 1.

In unbounded motion, the particle has positive total energy (E>0) and has a single extreme point where the total energy is always equal to the potential energy of the particle i.e the kinetic energy of the particle becomes zero.

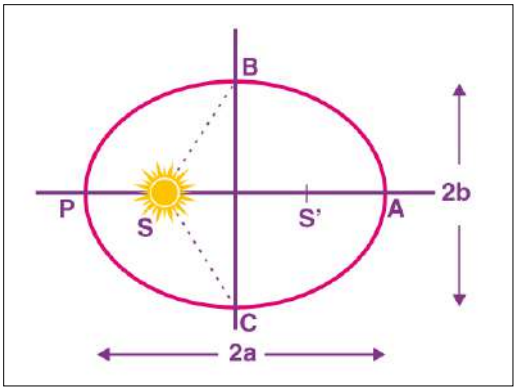
For eccentricity e ≥ 1, E > 0 implies the body has unbounded motion. Parabolic orbit has eccentricity e = 1 and Hyperbolic path has eccentricity e>1.

Kepler's laws of planetary motion can be stated as follows:

Kepler First law – The Law of Orbits

According to Kepler's first law," All the planets revolve around the sun in elliptical orbits having the sun at one of the foci". The point at which the planet is close to the sun is known as perihelion and the point at which the planet is farther from the sun is known as aphelion.

It is the characteristics of an ellipse that the sum of the distances of any planet from two foci is constant. The elliptical orbit of a planet is responsible for the occurrence of seasons.



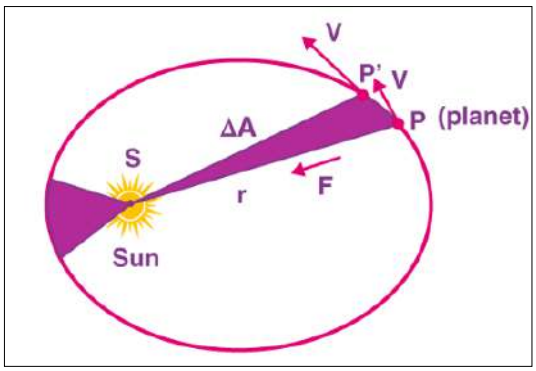
Kepler First law – The Law of Orbits

Kepler's Second Law – The Law of Equal Areas

Kepler's second law states" The radius vector drawn from the sun to the planet sweeps out equal areas in equal intervals of time"

As the orbit is not circular, the planet's kinetic energy is not constant in its path. It has more kinetic energy near perihelion and less kinetic energy near aphelion implies more speed at perihelion and less speed (v_{min}) at aphelion. If r is the distance of planet from sun, at perihelion (r_{min}) and at aphelion (r_{max}), then,

$$r_{min} + r_{max} = 2a \times \text{(length of major axis of an ellipse)}$$



Kepler's Second Law – The law of Equal Areas

Using the law of conservation of angular momentum the law can be verified. At any point of time, the angular momentum can be given as, $L = mr^2\omega$.

Now consider a small area ΔA described in a small time interval Δt and the covered angle is Δθ. Let the radius of curvature of the path be r, then the length of the arc covered = r Δθ.

$$\Delta A = 1/2[r.\ (r.\Delta\theta)] = 1/2r^2\Delta\theta$$

Therefore, $\Delta A/\Delta t = [1/2r^2]\Delta\theta/dt$

$$\lim_{\Delta t\to 0} \frac{\Delta A}{\Delta t} = \frac{1}{2}r^2 \frac{\Delta\theta}{\Delta t}$$, taking limits both side as, Δt→0

$$\Rightarrow \frac{dA}{dt} = \frac{1}{2}r^2\omega \quad \frac{dA}{dt} = \frac{L}{2m}$$

Now, by conservation of angular momentum, L is a constant

Thus, dA/dt = constant

The area swept in equal interval of time is a constant.

Kepler's second law can also be stated as "The areal velocity of a planet revolving around the sun in elliptical orbit remains constant which implies the angular momentum of a planet remains constant". As the angular momentum is constant all planetary motions are planar motions, which is a direct consequence of central force.

Kepler's Third Law – The Law of Periods

According to Kepler's law of periods," The square of the time period of revolution of a planet around the sun in an elliptical orbit is directly proportional to the cube of its semi-major axis".

$$T^2 \propto a^3$$

Shorter the orbit of the planet around the sun, shorter the time taken to complete one revolution. Using the equations of Newton's law of gravitation and laws of motion, Kepler's third law takes a more general form:

$$P^2 = 4\pi^2 / [G\ (M_1 + M_2)] \times a^3$$

Where, M_1 and M_2 are the masses of the two orbiting objects in solar masses.

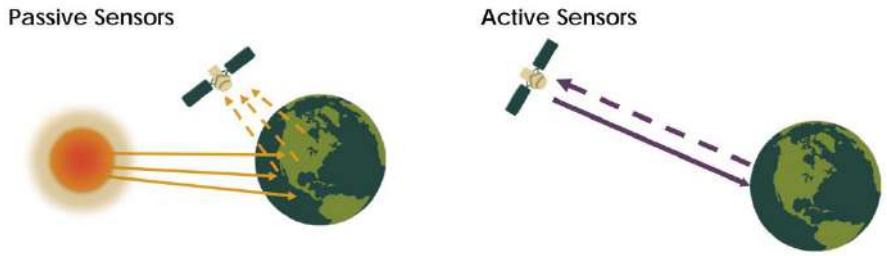
3. Satellite Sensors: Concept of IFOV, resolution and determination of pixel size, referencing scheme of satellite system (path/row calculation).

Satellite sensors

Satellite sensors that measure infrared radiation infer the amount of ***heat*** emitted from an object at the Earth's surface. Objects with an average earth temperature (roughly − 50° to 50°C, or − 58° to 122°F) emit most of their energy in the infrared region. Infrared sensors can easily detect sea ice, because its temperature is generally much colder than the surrounding ocean waters.

Remote sensing sensors are of two primary types:

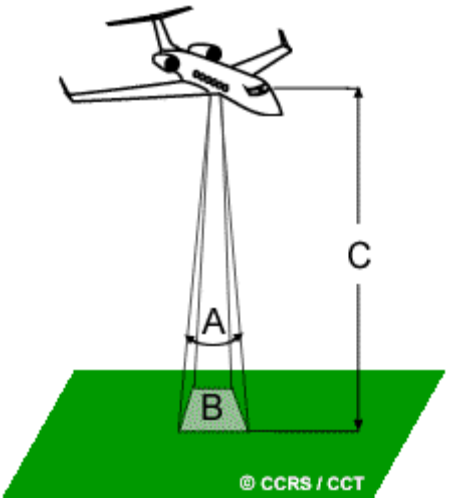
- **Active sensors** provide their own source of energy to illuminate the objects they observe. An active sensor emits radiation in the direction of the target to be investigated. The sensor then detects and measures the radiation that is reflected or backscattered from the target.
- **Passive sensors**, on the other hand, detect natural energy (radiation) that is emitted or reflected by the object or scene being observed. Reflected sunlight is the most common source of radiation measured by passive sensors.



Concept of IFOV

The IFOV is the angular cone of visibility of the sensor (A) and determines the area on the Earth's surface which is "seen" from a given altitude at one particular moment in time (B). The size of the area viewed is determined by multiplying the IFOV by the distance from the ground to the sensor

(C). This area on the ground is called the **resolution cell** and determines a sensor's maximum spatial resolution. For a homogeneous feature to be detected, its size generally has to be equal to or larger than the resolution cell. If the feature is smaller than this, it may not be detectable as the average brightness of all features in that resolution cell will be recorded. However, smaller features may sometimes be detectable if their reflectance dominates within an articular resolution cell allowing sub-pixel or resolution cell detection.



Resolutions

The resolution of remote sensed raster data can be characterized in several different ways. There are four primary types of "resolution" for rasters:

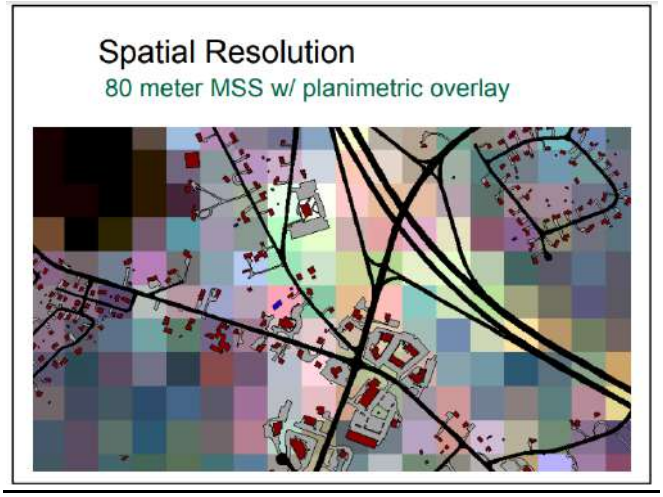
1. Spatial
2. Spectral
3. Radiometric
4. Temporal

Spatial Resolution

It is nearly impossible to acquire imagery that has high spatial, spectral, radiometric and temporal resolution. This is known as Resolution Trade-off, as it is difficult and expensive to obtain imagery with extremely high resolution.

Spatial Resolution describes how much detail in a photographic image is visible to the human eye. The ability to "resolve or separate", small details is one way of describing what we call spatial resolution.

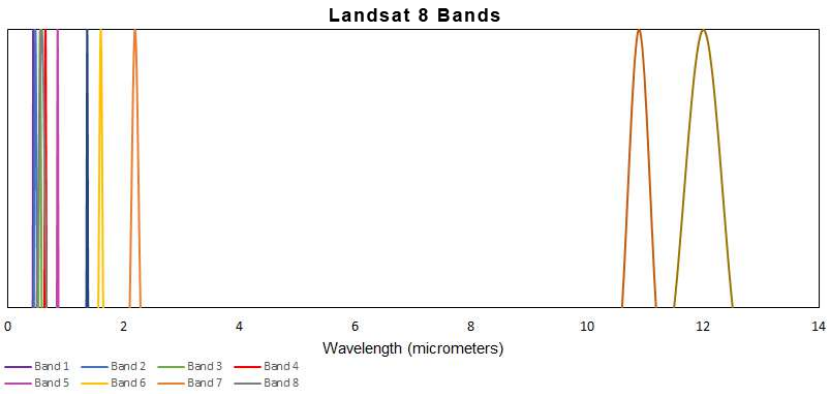
Spatial resolution of images acquired by satellite sensor systems is usually expressed in meters. For example, we often speak of Landsat as having “30- meter" resolution, which means that two objects, thirty meters long or wide, sitting side by side, can be separated (resolved) on a Landsat image. Other sensors have lower or higher spatial resolutions.



Spectral Resolution

Spectral resolution describes the ability of a sensor to define fine wavelength intervals. The finer the spectral resolution, the narrower the wavelength ranges for a particular channel or band. Black and white film records wavelengths extending over much, or the entire visible portion of the electromagnetic spectrum. Its spectral resolution is fairly coarse, as the various wavelengths of the visible spectrum are not individually distinguished and the overall reflectance in the entire visible portion is recorded. Colour film is also sensitive to the reflected energy over the visible portion of the spectrum, but has higher spectral resolution, as it is individually sensitive to the reflected energy at the blue, green, and red wavelengths of the spectrum. Thus, it can represent features of various colours based on their reflectance in each of these distinct wavelength ranges.

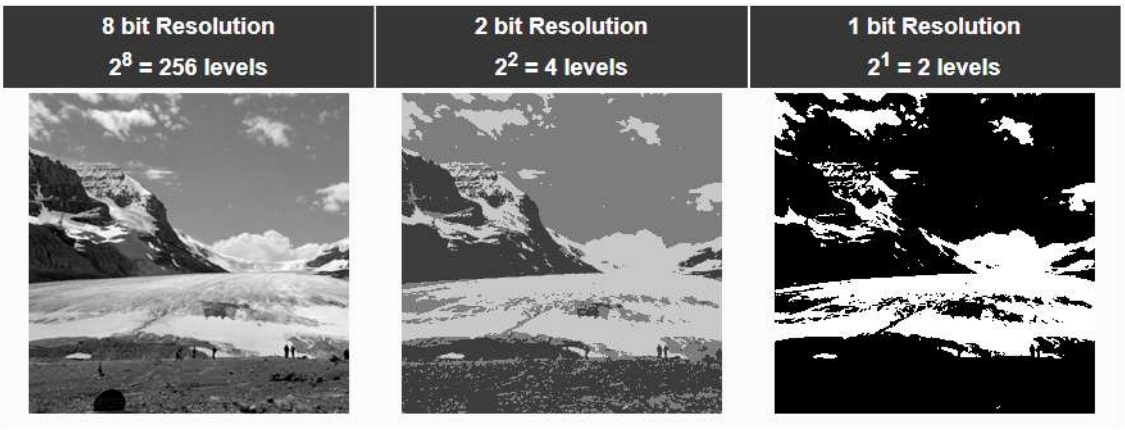
Many remote sensing systems record energy over several separate wavelength ranges at various spectral resolutions. These are referred to as multi-spectral sensors and will be described in some detail in following sections. Advanced multi-spectral sensors called hyper-spectral sensors, detect hundreds of very narrow spectral bands throughout the visible, near-infrared, and mid-infrared portions of the electromagnetic spectrum. Their very high spectral resolution facilitates fine discrimination between different targets based on their spectral response in each of the narrow bands.



Radiometric Resolution

While the arrangement of pixels describes the spatial structure of an image, the radiometric characteristics describe the actual information content in an image. Every time an image is acquired on film or by a sensor, its sensitivity to the magnitude of the electromagnetic energy determines the radiometric resolution. The radiometric resolution of an imaging system describes its ability to discriminate very slight differences in energy. The finer the radiometric resolution of a sensor, the more sensitive it is to detecting small differences in reflected or emitted energy.

Imagery data are represented by positive digital numbers which vary from 0 to (one less than) a selected power of 2. This range corresponds to the number of bits used for coding numbers in binary format. Each bit records an exponent of power 2 (e.g. 1 bit=2 1=2). The maximum number of brightness levels available depends on the number of bits used in representing the energy recorded. Thus, if a sensor used 8 bits to record the data, there would be 28=256 digital values available, ranging from 0 to 255. However, if only 4 bits were used, then only 24=16 values ranging from 0 to 15 would be available. Thus, the radiometric resolution would be much less. Image data are generally displayed in a range of grey tones, with black representing a digital number of 0 and white representing the maximum value (for example, 255 in 8-bit data). By comparing a 2-bit image with an 8-bit image, we can see that there is a large difference in the level of detail discernible depending on their radiometric resolutions.



Temporal Resolution

In addition to spatial, spectral, and radiometric resolution, the concept of temporal resolution is also important to consider in a remote sensing system. Temporal resolution refers to the length of time it takes for a satellite to complete one entire orbit cycle. The revisit period of a satellite sensor is usually several days. Therefore the absolute temporal resolution of a remote sensing system to image the exact same area at the same viewing angle a second time is equal to this period. However, because of some degree of overlap in the imaging swaths of adjacent orbits for most satellites and the increase in this overlap with increasing latitude, some areas of the Earth tend to be re-imaged more frequently. Also, some satellite systems are able to point their sensors to image the same area between different satellite passes separated by periods from one to five days.

Thus, the actual temporal resolution of a sensor depends on a variety of factors, including the satellite/sensor capabilities, the swath overlap, and latitude. The ability to collect imagery of the same area of the Earth's surface at different periods of time is one of the most important elements for applying remote sensing data.

Spectral characteristics of features may change over time and these changes can be detected by collecting and comparing multi-temporal imagery. For example, during the growing season, most species of vegetation are in a continual state of change and our ability to monitor those subtle changes using remote sensing is dependent on when and how frequently we collect imagery. By imaging on a continuing basis at different times we are able to monitor the changes that take place on the Earth's surface, whether they are naturally occurring (such as changes in natural vegetation cover or flooding) or induced by humans (such as urban development or deforestation).

The time factor in imaging is important when: persistent clouds offer limited clear views of the Earth's surface (often in the tropics) short-lived phenomena (floods, oil slicks, etc.) need to be

imaged multi-temporal comparisons are required (e.g. the spread of a forest disease from one year to the next) the changing appearance of a feature over time can be used to distinguish it from near similar features (wheat / maize).

Comparison of Landsat Sensors

	Thematic Mapper (TM) Landsat 4 and 5	**Enhanced Thematic Mapper Plus (ETM+)** Landsat 7	**Multispectral Scanner (MSS)** Landsat 1-5
Spectral Resolution (μm)	1. 0.45-0.52 (B) 2. 0.52-0.60 (G) 3. 0.63-0.69 (R) 4. 0.76-0.90 (NIR) 5. 1.55-1.75 (MIR) 6. 2.08-2.35 (MIR) 7. 10.4-12.5 (TIR)	1. 0.45-0.52 2. 0.53-0.61 3. 0.63-0.69 4. 0.78-0.90 5. 1.55-1.75 6. 2.09-2.35 7. 10.4-12.5 8. 0.52-0.90 (Pan)	0.5-0.6 (green) 0.6-0.7 (red) 0.7-0.8 (NIR) 0.8-1.1 (NIR)
Spatial Resolution (meter)	30 x 30 120 x 120 (TIR)	15 x 15 (Pan) 30 x 30 60 x 60 (TIR)	79 x 79
Temporal Resolution (revisit in days)	16	16	18
Spatial coverage (km)	185 x 185	183 x 170	185 x 185
Altitude (km)	705	705	915 (Landsat 1,2,3)

Referencing scheme of satellite system (path/row calculation)

Referencing scheme, which is unique for each satellite mission, is a means of conveniently identifying the geographic location of points on the earth. This scheme is designated by Paths and Rows. The Path-Row concept is based on the nominal orbital characteristics.

- **Path:** An orbit is the course of motion taken by the satellite, in space and the descending ground trace of the orbit is called a 'Path'.
- **Row:** Along a path, the continuous stream of data is segmented into a number of scenes of convenient size. The uniformly separated scene centres are, such that, same rows of different paths fall at the same latitude. The lines joining the corresponding scene centres of different paths are parallel to the equator and are called Rows.

Use of referencing scheme

The Path-Row referencing scheme eliminates the usage of latitude and longitudes and facilitates convenient and unique identification of a geographic location. It is useful in preparing accession and product catalogues and reduces the complexity of data products generation. Using the referencing scheme, the user can arrive at the number of scenes that covers his area of interest.

However, due to orbit and attitude variations during operation, the actual scene may be displaced slightly from the nominal scene defined in the referencing scheme. Hence, if the user's area of interest lies in the border region of any scene, the user may have to order the overlapping scenes in addition to the nominal scene.

Determination of observation dates

For the chosen path, the ground track repeats every 24 days after 341 orbits. Therefore, the coverage pattern is almost constant. The deviations of orbit and attitude parameters are controlled within limits such that the coverage pattern remains almost constant throughout the mission. Therefore, on any given day, it is possible to determine the orbit which will trace a designated path. Once the path is known, with the help of referencing scheme, it is possible to find out the region covered by that path. Therefore, an orbital calendar, giving the details of paths, covered on different days is helpful to users to plan their procurement of satellite data products.

Considering a typical path calendar (see Path Calendar Table, Table 1), assuming that path number 1 is covered on January 11, if data over a geographic area covered by path 60 is required, it is seen that this path is covered on days, 18th of January, 11th of February, 06th of March and so on. Thus, it is possible to know on which day the required data has been collected or is going to be collected.

Path	167	172	177	182	187	168	173	178	183	188	169	174	179	184	189	170	175	180	185	190	171	176	181	186
	143	148	153	158	163	144	149	154	159	164	145	150	155	160	165	146	151	156	161	166	147	152	157	162
	119	124	129	134	139	120	125	130	135	140	121	126	131	136	141	122	127	132	137	142	123	128	133	138
	95	100	105	110	115	96	101	106	111	116	97	102	107	112	117	98	103	108	113	118	99	104	109	114
	71	76	81	86	91	72	77	82	87	92	73	78	83	88	93	74	79	84	89	94	75	80	85	90
	47	52	57	62	67	48	53	58	63	68	49	54	59	64	69	50	55	60	65	70	51	56	61	66
	23	28	33	38	43	24	29	34	39	44	25	30	35	40	45	26	31	36	41	46	27	32	37	42
	340	4	9	14	19	341	5	10	15	20	1	6	11	16	21	2	7	12	17	22	3	8	13	18
	316	321	326	331	336	317	322	327	332	337	318	323	328	333	338	319	324	329	334	339	320	325	330	335
	292	297	302	307	312	293	298	303	308	313	294	299	304	309	314	295	300	305	310	315	296	301	306	311
	268	273	278	283	288	269	274	279	284	289	270	275	280	285	290	271	276	281	286	291	272	277	282	287
	244	249	254	259	264	245	250	255	260	265	246	251	256	261	266	247	252	257	262	267	248	253	258	263
	220	225	230	235	240	221	226	231	236	241	222	227	232	237	242	223	228	233	238	243	224	229	234	239
	196	201	206	211	216	197	202	207	212	217	198	203	208	213	218	199	204	209	214	219	200	205	210	215
					192					193					194					195				191
Jan	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	25	26	27	28	29	30	31																	
Feb								1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	18	19	20	21	22	23	24	25	26	27	28	29												
Mar													1	2	3	4	5	6	7	8	9	10	11	12
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31					
Apr																				1	2	3	4	5
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
	30																							
May		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
	24	25	26	27	28	29	30	31																
Jun									1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	17	18	19	20	21	22	23	24	25	26	27	28	29	30										
Jul															1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31			
Aug																						1	2	3
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
	28	29	30	31																				
Sep					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	21	22	23	24	25	26	27	28	29	30														
Oct											1	2	3	4	5	6	7	8	9	10	11	12	13	14
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31							
Nov																		1	2	3	4	5	6	7
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Dec																								1
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	26	27	28	29	30	31																		

Table 1: Path Calendar Table

Estimation of path and row, local clock time and other details for any point on the Indian sub-continent:

The procedure outlined below may be used to determine the path and row, Greenwich Meridian Time (GMT) and the local clock time when the satellite passes over any point in the Indian subcontinent.

i. Define the latitude and longitude of the point of interest over Indian region.

ii. Determine the approximate descending node as follows:

1. Locate the latitude of the point of interest in Table 2. Table 2 gives the longitudinal difference from the given longitude to the descending node longitude as a function of latitude.

2. Read the value of longitude from this table. If the latitude falls within two values, then, interpolate and get required longitude.

3. Add this value to the longitude of the point of interest, to get rough estimate of descending node longitude.

iii. The actual descending node details are obtained as follows:

1. Table 3 gives the descending node longitude of all paths over the Indian region. Find the path nearest to the longitude computed in step ii. This gives the path number and descending node longitude of the path.

2. Table 4 gives the descending node time (GMT) expected for each path over the Indian sub-continent.

iv. GMT at the point of interest is found as follows:

1. Given a latitude, using the nominal inclination of the orbit, the time of descending node can be calculated

2. Add the time to the GMT of the descending node as obtained in step iii, by carefully noting the algebraic sign.

v. The Indian Standard Time (IST) is obtained by adding five and a half hours to the time (GMT) obtained in step iv.

vi. Table 2 gives the row numbers versus latitude. Find the nearest row latitude from this table and assign the same row number. Thus, with the above procedure, the path and row numbers and other details of the point of interest can be obtained.

Latitude	Row No.	Longitude	Latitude	Row No.	Longitude
81.30	1	-88.78	42.79	39	-11.12
81.06	2	-81.05	41.61	40	-10.70
80.68	3	-73.82	40.43	41	-10.30
80.18	4	-67.23	39.25	42	-9.91
79.56	5	-61.34	38.07	43	-9.53
78.84	6	-56.15	36.89	44	-9.16
78.05	7	-51.58	35.71	45	-8.80
77.20	8	-47.59	34.53	46	-8.44
76.30	9	-44.08	33.34	47	-8.10
75.36	10	-41.00	32.16	48	-7.76
74.38	11	-38.29	30.97	49	-7 .42
73.37	12	-35.87	29.79	50	-7.10
72.35	13	-33.72	28.60	51	-6.77
71.30	14	-31.79	27.41	52	-6.46
70.23	15	-30.05	26.22	53	-6.15
69.16	16	-28.48	25.04	54	-5.84
68.07	17	-27.04	23.85	55	-5.54
66.97	18	-25.73	22.66	56	-5.24
65.86	19	-24.53	21.47	57	-4.94
64.74	20	-23.41	20.28	58	-4.65
63.62	21	-22.38	19.09	59	-4.36
62.49	22	-21.42	17.89	60	-4.08
61.36	23	-20.53	16.70	61	-3.79
60.22	24	-19.69	15.51	62	-3.51
59.07	25	-18.90	14.32	63	-3.23
57.93	26	-18.16	13.13	64	-2.96
56.78	27	-17.45	11.93	65	-2.68
55.62	28	-16.79	10.74	66	-2.41
54.47	29	-16.16	9.55	67	-2.14
53.31	30	-15.55	8.35	68	-1.87
52.15	31	-14.98	7.16	69	-1.60
50.98	32	-14.43	5.97	70	-1.33
49.82	33	-13.90	4.77	71	-1.06
48.65	34	-13.39	3.58	72	-0.80
47.48	35	-12.91	2.39	73	-0.53
46.31	36	-12.43	1.19	74	-0.27
45.13	37	-11.98	0.0	75	0.00
43.96	38	-11.54			

The difference in longitude of a given row latitude and descending time

Table: 2

Path	Longitude	Path	Longitude
65	37.866	101	75 .72
66	38.922	102	76.928
67	39.977	103	77.983
68	41.033	104	79.039
69	42.089	105	80.095
70	43.145	106	81.150
71	44.200	107	82.206
72	45.256	108	83.262
73	46.312	109	84.318
74	47.367	110	85.373
75	48.423	111	86.429
76	49.479	112	87.485
77	50.535	113	88.540
78	51.590	114	89.596
79	52.646	115	90.652
80	53.702	116	91.708
81	54.757	117	92.763
82	55.813	118	93.819
83	56.869	119	94.875
84	57..925	120	95.930
85	58.980	121	96.986
86	60.036	122	98.042
87	61.092	123	99.098
88	62.148	124	100.153
89	63.203	125	101.209
90	64.259	126	102.265
91	65.315	127	103.321
92	66.370	128	104.376
93	67.426	129	105.432
94	68.482	130	106.488
95	69.538	131	107.543
96	70.593	132	108.599
97	71.649	133	109.655
98	72.705	134	110.711
99	73.760	135	111.766
100	74.816		

Equatorial crossing longitude for paths over Indian region

Path	GMT	Path	GMT
65	7:59	101	5:27
66	7:54	102	5:22
67	7:50	103	5:18
68	7:46	104	5:14
69	7:42	105	5:10
70	7:37	106	5:05
71	7:33	107	5:01
72	7:29	108	4:57
73	7:25	109	4:53
74	7:21	110	4:49
75	7:16	111	4:44
76	7:12	112	4:40
77	7:08	113	4:36
78	7:04	114	4:32
79	6:59	115	4:27
80	6:55	116	4:23
81	6:51	117	4:19
82	6:47	118	4:15
83	6:43	119	4:11
84	6:38	120	4:06
85	6:34	121	4:02
86	6:30	122	3:58
87	6:26	123	3:54
88	6:21	124	3:49
89	6:17	125	3:45
90	6:13	126	3:41
91	6:09	127	3:37
92	6:05	128	3:32
93	6:00	129	3:28
94	5:56	130	3:24
95	5:52	131	3:20
96	5:48	132	3:16
97	5:43	133	3:11
98	5:39	134	3:07
99	5:35	135	3:03
100	5:31		

Equatorial crossing time (GMT) for paths over Indian region (Local time at descending node 10:30 hrs)

Table: 3 and 4

4. Basics of Aerial Photograph: Basics geometry of aerial photograph, determination of scale and height, Distortions, Image parallax, Relief displacement.

Aerial Photography

Aerial photography is defined as the science of obtaining photographs from the air using various platforms, mostly aircraft, for studying the surface of the earth. The sun provides the source of energy (electromagnetic radiation or EMR) and the photo-sensitive film acts as a sensor to record the images. Variations in the grey tones of the various images in a photograph indicate different amounts of energy reflected from the objects as recorded on the film.

The earth's atmosphere, which contain various particles and molecules of gases and water vapor, attenuates the incoming as well the outgoing energy/radiation (scattering) after interaction (reflectance, transmittance and absorption) with the object and thus reduces the contrast between different images formed on the photographic film. Therefore, the quality of aerial photography largely depends upon the atmospheric conditions prevailing at that time. Different filter/lens combinations can, however, be used to eliminate some of the atmospheric effect in black and white photography by making use of a yellow (minus blue) filter to reduce the effects of haze. The problem becomes more complex in the case of colour photography.

Applications of Aerial Photography

Mapping: The application of aerial photography in photogrammetric mapping is an established procedure all over the world. It has been found to be fast, accurate, and indispensable in inaccessible areas and cost effective in the long run, as initially the establishment of a photogrammetric survey/mapping unit involves capital expenditure due to the cost of photogrammetric instruments and other ancillary equipment.

Interpretation: Photo interpretation has revolutionalized the methods of data collection in various disciplines. It greatly reduces the field work and thereby the cost. The information is reliable and

acceptance for most studies such as in the fields of geology, water resources, geomorphology, hydrogeology, forestry and ecology, soil surveys and urban and regional planning.

Map: Substitute In a situation where there are no adequate large scale maps available, aerial photographs can serve as map substitutes in the form of photomaps. In the case of relatively flat terrain, these photomaps can be produced by rectification to remove the effects of tilt distortion and scale correction. This method has been found to be three to four times faster than conventional mapping by photogrammetric methods. In the case of hilly terrain, such photomaps (orthophoto maps) can be produced by the orthophoto technique, which has also proved to be faster than conventional mapping. In some urgent situations, simple mosaics prepared from aerial photographs can substitute for maps.

Types of Photographs

Photographs which are used for mapping and photo-interpretation can be divided into the following main classes according to the direction of the camera axis:

a) Vertical photographs b) Horizontal or terrestrial photographs and c) Oblique photographs.

The terms 'vertical' and 'horizontal' refer to the direction in which the camera axis was pointing at the time of exposure.

- **Vertical Air Photographs**

These are taken with the axis of the aerial camera vertical or nearly vertical. A vertical photograph closely resembles a map and is particularly suitable for obtaining uniform coverage. As these photographs can be obtained with reasonably low tilt (tilt is deviation of the camera axis from the vertical) they are generally used for mapping and photo-interpretation work and also extension of control.

- **Terrestrial Photographs**

These are taken with photo-theodolites from camera stations on the ground with the axis of the camera horizontal and they present the more familiar elevation view. These types of photographs are used for survey of structures and monuments of architectural or archaeological value. Terrestrial photographs taken with normal good cameras can also be of considerable use in supplementing photo-interpretation of vertical aerial photographs particularly so in geology and forestry, where study of a profile may be needed.

- **Oblique Photographs**

Aerial photographs taken with the optical axis of the aerial camera tilted from the vertical are known as oblique photographs. These photographs cover large areas of ground but clarity of details diminishes towards the far end of the photographs. Aerial photographs on which the horizon does not appear are known as low oblique and are, sometimes, used to compile reconnaissance maps of inaccessible areas. High oblique photographs, which are tilted sufficiently to contain the horizon, were previously used for extension of planimetric and height control, when the available ground control was insufficient to provide necessary accuracy. These have very limited use at present.

There are combinations of above types of photography taken with two or more cameras in a single camera unit in the photographic air plane.

i) Convergent Photographs

These are low oblique photographs taken with two cameras exposed simultaneously at successive exposure stations, with their axes tilted at a fixed inclination from the vertical in opposite directions in the direction of the flight line so that the forward exposure of the first station forms stereo-pair with the backward exposure of the next station. Special plotting instruments are required for compiling topographical maps from convergent photographs.

ii) Trimetrogon

Photography Another type of photography which is a combination of a vertical and two oblique photographs is time trogon photography, in which the central photograph is vertical and the side ones are oblique. This photography can be used for rapid production of reconnaissance maps on small scales.

Classification According to Angle of Coverage

Another classification of aerial photography is based on angle of coverage. The angle of coverage is defined as the angle, diagonal of the negative format subtends at the rear node of the lens or the apex angle of the cone of rays passing through the front nodal point of the lens.

We distinguish

1. Standard or normal – angle photography

The angle of coverage is of the order of 60 degrees

Format size

(I) 18 cm x 18 cm, f = 21 cm and

(II) 23 cm x 23 cm, f = 30 cm

2. Wide – angle photography

The angle of coverage is of the order of 90 degrees

Format size

(I) 18 cm x 18 cm, f = 11.5 cm and

(II) 23 cm x 23 cm, f = 15 cm

3. Super wide or ultra-wide angle photography

The angle of coverage is of the order of 120 degrees Format size

(i)18 cm x 18 cm, f = 70 mm and

(ii)23 cm x 23 cm, f = 88 mm

4. Narrow – angle photography

The angle of coverage is less than 50 degrees

Information recorded on Aerial Photographs

The following information appears on all aerial photographs:

a) Fiducial marks or collimating marks for the determination of the principal point

b) Altimeter recorded for determining the flying height of the aircraft at the moment of exposure above M.S.L. (mean sea level)

c) Watch recording gives the time of exposure

d) Level bubble indicates the tilt of the camera axis at the moment of exposure (not very accurate)

e) Principal distance for determining the scale of the photograph

f) Number of the photograph, the strip number and the specification number for easy handling and indexing of photographs

g) Number of the camera, so that the camera calibration report can be obtained, if required

h) Date of photography

Geometry of Aerial Photographs

1. Projection

In order to understand the geometric qualities of a photograph it is necessary to understand what projection means in terms of geometry. In the examples given the triangle ABC and the line LL' on which the projection is made are in the same plane.

a) **Parallel Projection:** In this projection, the projecting rays are parallel. The triangle ABC is projected on the LL'. The projection of the triangle is 'abc'. The projection rays Aa, Bb, Cc, are all parallel in this case.

b) **Orthogonal Projection:** In this case the projecting rays are all perpendicular to the line LL'. This is a special case of parallel projection. Maps are an orthogonal projection of the ground on a certain scale. The advantage of this projection is that the distances, angles and areas in the plane are independent of the elevation differences of the objects.

c) **Central Projection:** The projecting rays Aa, Bb, Cc, pass through one point 0, called the Projection Centre or Perspective Centre. The image projected by a lens system is treated as a central projection, (though strongly it is not, as the lens is not a single point).

2. TILT

It is the angle between the optical axis of the camera and the plumb line. It is also the angle between the ground plane and the photo plane. Tilt can be resolved into two components, one in the direction of flight (the X-axis) and the other perpendicular to it (the Y-axis). i) The component about the Y-axis, i.e. in the direction of X is called Longitudinal Tilt or X-tilt or Fore and Aft Tilt or Tip. It is denoted by letter (Phi). ii) The component about the X-axis, i.e. in the direction of Y is called Lateral Tilt or Y-Tilt or simple Tilt. It is donated by letter (Omega).

The vertical ON through the perspective center meets the photo plane at point 'n' called the Photo nadir point and the ground plane at point N called the Ground nadir point. These points are also called Plumb Point. The foot of the perpendicular (p) from O on the photo plane is called Principle Point. The length of this perpendicular (op) is called Principle Distance.

The approximate position of the principal point of a photograph is determined by joining the opposite fiducial marks (or collimating marks). Line joining opposite fiducial marks is known as fiducial axis. The point of intersection of the fiducial axes is called fiducial centre (0 and is, for practical purposes, coincident with the principal point (p) in a well adjusted camera.

Reasons for Photo Tilt

i) Atmospheric conditions (air pockets or currents)

ii) Human error of the pilot fails to maintain a steady flight

iii) Imperfections in the camera mounting, etc.

3. SWING

Swing is the angle measured in the plane of the photograph between the fiducial axis in the direction of flight and the actual flight line.

Scale of Photographs

Scale is the relationship between distance on a map or photo and the actual ground distance.

Scale is represented in two ways:

a) Equating different units of measurement on map and ground, i.e. 1 inch = 1 Mile.

64 inches = 1 Mile

b) As R.F. (representative fraction) in which the numerator is unity, e.g. 1:10,000 or 1/10,000 which means 1 unit on the map or photo represents 10,000 units on the ground.

Methods of scale determination

In decreasing order of accuracy these are:

i) By establishing the relation of photo to ground

If the distance between the same two points on the photo as well as on the ground can be measured, R.F. can be set up:

$$\text{R.F.} = \frac{\text{Photo distance}}{\text{Ground distance}}$$

ii) By establishing the relation of photo to ground with the help of a map. If the distance between two points on a photo which can be located on the map as well is measured, the horizontal measurements of these distances form a ratio, which when multiplied by the R.F. of the map gives the R.F. of the photo. If 'g' be the ground distance between two points, 'm' the map distance and 'p' the photo distance then R.F. of map is m/g and R.F. of photo scale is p/g.

$$\frac{\text{R.F. of photo}}{\text{R.F. of map}} = \frac{p/g}{m/g} = \frac{p}{ni}$$

$$\text{Hence R.F. of photo} = \frac{p}{m} \times \text{RF. of map}$$

iii) By establishing the relation between focal length of the camera and the flying altitude

In a true vertical photograph of fiat terrain the scale of photograph is the ratio f/H. Distance 'AB' is imaged as 'ab' on the photo.

$$\text{Scale of the photo} = \frac{\text{Photo distance}}{\text{Ground distance}}$$

$$= \frac{ab}{AB}$$

$$= \frac{f}{H} \text{ (from similar triangles OAB \& Oab)}$$

If the terrain is not fiat, the scale of the photograph is not uniform. Hmi is the flying height above the average height of the terrain photographed. Then the average scale of the photograph = f/Hm. The scale of photo for a point A which is at a height of h meters/ft above the average ground level.

$$= \frac{f}{Hm - h} \text{ (the units of focal length and the heights being in the same terms)}$$

Similarly, the scale for another point B which is at a vertical distance 'd' metres/ft below the average terrain level.

$$= \frac{f}{Hm + d}$$

Thus the scale of photograph is not uniform if there is irregular terrain. We can determine either the average scale of the photograph as a whole or the scale of the photograph at a particular point or elevation. Higher areas will be on a larger scale than that of lower vellies. In tilted photograph the scale is not constant. it is constant along any plate parallel (if the ground is fiat). The scale along isometric parallel (discussed earlier) is true, i.e., equal to f/H. The scale increases continuously on the nadir point side of the iso-centre and decreases continuously towards the principal point side of it. Thus we arrive at an important result:

The scale of aerial photograph changes irregularly due to height difference in the terrain but continuously due to inclination of the camera axis.

Resolution

Resolution of aerial photograph is expressed in lines pair per millimeter, i.e., numbers of lines and equal size gap can be resolved. We can get approximately 20 lines pair per mm on the scale of the original negative. For example if the original scale of negative is 1:10,000, then the ground resolution will be –

$$\frac{1}{(20+20)} \times 10,000 \text{ mm} = \frac{1}{40} \times 10,000 \text{ mm} = 25 \text{ cm on the ground.}$$

Image Displacement

On a planimetric map all features/details are shown in their correct horizontal position on a certain scale. This is not so in the case of aerial photographs due to image, displacement or distortion. A disturbance of the principle of geometry is called displacement/distortion. There are three major sources of displacement/distortion which arc due to: optical or photographic deficiencies i.e. lens distortion and aberration; relief variation of the object photographed and tilt of the camera axis at the moment of exposure.

(a) Lens distortion Object point O is imaged at I' instead of its correct position I on the image plane. d is the image displacement in this case. In the modern aerial camera lens this type of distortion is negligible.

b) Image displacement due to relief Relief is the most significant source of image displacement. O is the camera station. NA' is a fiat plain on which stands a tower AB with its base at B. The image of B on the truly vertical positive photographic plane is b. This is the correct planimetric position (orthogonal) of the image of the tower AB. Top A is imaged at 'a'. The image of A is thus displaced from its correct planimetric position b, as 'A' is vertically above 'B', on the photograph. This shift of 'a' from 'b' represented by distance ba is called relief displacement. Let h be the height of the tower, H the flying height above the datum plane, n and N be the photo and ground nadir points.

$$\text{The scale of photo along na is } \frac{na}{NA'}$$

$$\text{The scale of photo along ab is } \frac{ab}{BA'}$$

Since the photograph is truly vertical and datum plane is truly horizontal the scale will be constant.

$$\text{Hence } \frac{ab}{BA'} = \frac{na}{NA'}$$

$$\text{or } \frac{ab}{na} = \frac{BA'}{NA'} \qquad \text{------ (1)}$$

From similar triangles ONA' and ABA'

$$\frac{BA'}{NA'} = \frac{h}{H}$$

$$\text{Equation (1) becomes } \frac{ab}{na} = \frac{h}{H}$$

$$\text{Hence } ab = na \frac{h}{H}$$

If we denote 'ab', the displacement by r' and na, the distance between the nadir point and the image of top of the object by r then we can write the above equation as

From this relation we conclude:

$$r' = r \frac{h}{H}$$

i) Relief displacement increases with increasing value of 'r' i.e., it is zero at plumb point and maximum at the edges of the photograph.

ii) Smaller the height of the object, smaller is the displacement and vice versa. If h = o, i.e. for objects in the datum plane there is no displacement.

iii) With increasing value of 'H' i.e. with high flying heights the displacement decreases. The satellite pictures can thus be considered having very low relief displacement.

It can also be proved that the relief displacement is radial from the plumb point.

While relief displacement constitutes a source of error in the measurement of horizontal distances on aerial photographs, it is the characteristics that make it possible to study overlapping photographs stereoscopically and in the determination of height differences between objects photographed.

c) Image displacement due to tilt

I. <u>Flat Terrain</u> - Let O be the perspective center I and II be the positive planes for a truly vertical and tilted photographs respectively. The figure shows a cross- section in the principal plane. For a point

'A' which appears at a' in I and at a in II, the displacement is equal to ia' - ia. It can be shown that it is equal to:

$$\frac{ia^2 \sin\theta}{f - ia\sin\theta}$$

and is radial from the isocentre. For a point b in plane II (Fig. 10) which does not lie in the principal plane and is that an angle with principal line at the isocentre 'i', the tilt displacement which is still radial from the isocentre can be shown to be equal to

$$ib' - ib = \frac{ib^2 \sin\theta \cos 2\theta}{f - ib\sin\theta\cos\theta}$$

The displacement due to tilt is outward from the isocentre when the point is nadir point side of the isometric parallel and inward when on the principal point side. If the tilt is small n and i will be closer to p and, therefore, for near vertical photographs we assume that the relief displacement is radial from the principal point. (Principal point coinciding with the nadir point and the iso-centres for all practical purposes) and the displacement due to tilt is negligible. This assumption is valid for all graphical methods of plotting, mean height of relief being less than 10% of the flying height. The only mark easily available on the photograph is the principal point which can be easily plotted and is convenient to use. The isocentres or the plumb point, though easy to define are difficult to locate on the photograph.

II. Accidental Terrain - We know that displacement due to relief is radial from the plumb point and that displacement due to tilt, in case of flat terrain, is radial from the iso-centres. There is, however, no such point on the photograph where angles are true to the corresponding angles on the ground in the case of accidental terrain i.e. terrain in which there is elevation differences.

Parallax for Height Measurement using Aerial Photography

Parallax Concept

Photogrammetry is capable of measuring ***elevation*** of earth surface. Aerial photographs/stereo pair satellite images can be used to measure elevation differences through the use of parallax method.

New launched satellite is providing stereo pair ***satellite images*** images such worldview-2 etc.

Parallax can be defined as the apparent displacement of a point due to a change in view of the point.

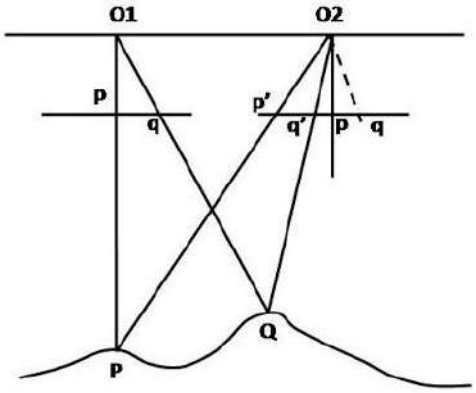
One of the parallax examples for human eye can be explained by a simple exercise of childhood.

i.e. hold finger in front of eyes and look at finger, where it is relative to a wall in the background with your right eye. Then look at it with your left eye and its appearance relative to the wall has changed.

This relative change in appearance is due to parallax. It took place due to change in view point for the finger.

Parallax of Aerial Photographs

Parallax can be described by below figure



In this figure point **P** & **Q** on surface are captured on two aerial photographs as **p** & **q** respectively.

- *Consider that at one instant the airplane is at **O1**, vertically above a point **P**. The image of **P** will appear at **p** on the image plane.*
- *After sometime when the plane is at **O2, P** will appear as **p**' on this image plane. The ground point **P** appear as **p** on first image plane and **p**' on the second image plane so this shift of **pp**' in the position of the image of **P** on the image plane is the parallax of **P**.*
- *Similarly for any other point **Q**, one instant the airplane is at **O1**, vertically above a point **P**. The image of **Q** will appear at q on the image plane.*

- *After sometime when the plane is at **O2, Q** will appear as **q'** on this image plane. The ground point **Q** appear as **q** on first image plane and **q'** on the second image plane so this shift of **qq'** in the position of the image of **Q** on the image plane is the parallax of **Q**.*

Types of Parallax for Aerial Photographs

There are two **types of Parallax;** Absolute Parallax and Differential Parallax

Absolute Parallax (X-Parallax/Horizontal Parallax): This parallax is in the X direction, It is the algebraic difference of the distances of the two images from their respective photograph nadirs, measured in horizontal plane and parallel to the airbase.

Differential Parallax(Y-Parallax): This parallax in Y direction, the difference between the perpendiculars distances of two the images of a point from the vertical plane containing the airbase.

Y-parallax is an indication of tilt in either or both photographs, or a difference in a flying height.

Parallax used for height determination by below method

$$\Delta h = \frac{\Delta p \, H'}{pc}$$

$$\Delta h = ha - hc$$

Where,

Δh = change in elevation between two points a and c

Δp = parallax of point a subtracted from parallax of point c

H' = flying height of airplane

pc = parallax of point c

While measuring height using Parallax, below are required

- Any point being measured has to appear on both aerial photos that overlap
- If elevation (benchmark) of point c has known and then its parallax can be measured

- Any point a's elevation can be calculated relative to point c's known elevation
- Parallax measurement facilitates computation of an elevation model of ground using aerial photographs.

Numerical Example

Question: Benchmark c has an elevation of 1545.32 ft., x coordinate on left photo of +74.12 mm and on right photo of -18.41 mm. Unknown point 'a' has x coordinate on left photo of +65.78 mm and on right photo of -24.38 mm. If flying height above average ground is 3000 ft. what is the elevation of point a?

Solution: Parallax is the change in x coordinates defined as parallel to the flight line.

Parallax of point c = 74.12 – (-18.41) = 92.53 mm

Parallax of point a = 65.78 – (-24.38) = 90.16 mm

Height measured using below formula

$$\Delta h = ha - hc = \frac{\Delta p \, H'}{pc}$$

Here all values calculated below

Δp = pa – pc = 90.16 – 92.53 = -2.37 mm

Δh = (-2.37 mm) * (3000 ft.) / 92.53 mm

Δh = -76.84 ft.

By putting Δh value in below equation

ha = hc + Δh = 1545.32 + (-76.84)

ha= 1468.48 ft.

It can be assumed from question that if point 'a' has less parallax so it should have lower elevation than point c.

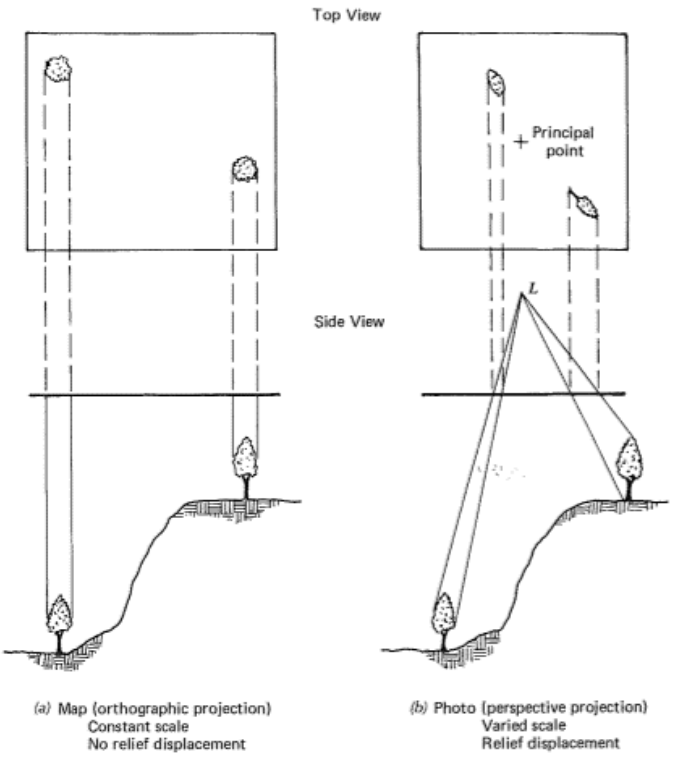
This Parallax concept is important in aerial photography as this can be used to measure the ***height of ground***.

Distortion in Aerial Photos

All aerial imagery, whether it is acquired by a sensor on a satellite, or by an aircraft, will have some amount of geometric distortion. This is an issue in remotes sensing as we want to be able to accurately represent a three-dimensional surface on a two-dimensional image. The errors can be due to a number of factors including:

- Camera/Sensor tilt
- Camera Lens distortion
- Terrain/Relief

Most aerial photographs are taken with specialized cameras to minimize lens distortion, but some distortion is still present.

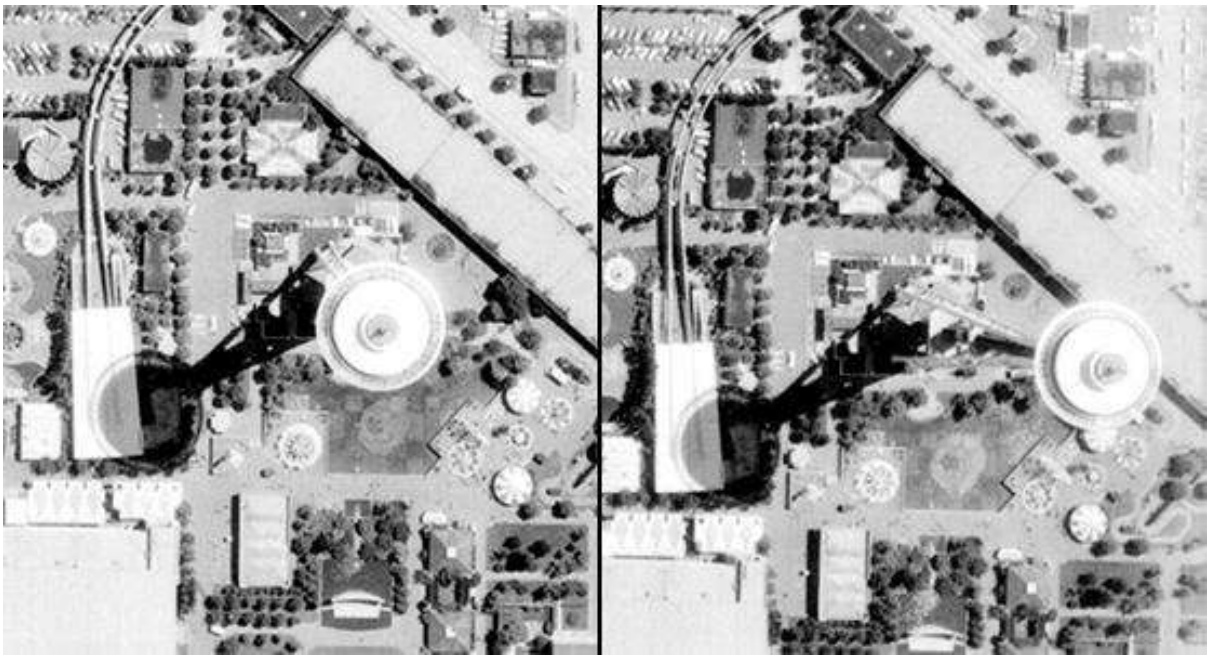


Terrain/Relief Displacement

The scale of ordinary aerial images varies across the image, primarily due to the changing elevation of the terrain surface. Scale distortion is zero at the principal point (center) and increases radially

from the center of the image. Therefore when scale is given for an un-rectified aerial photograph, it is actually an average scale based on the average terrain elevation for the image. On a photograph, areas of terrain (or buildings) at higher elevation lie closer to the camera at the time of exposure. Therefore they appear larger than areas at lower elevations.

The tops of tall objects are always displaced from their bases. This is known as **relief displacement** and causes any objects above the ground to lean away from the principal point of the photograph. The greater the distance the object is away from the principal point, the greater the displacement. The displacement also increases as the height of the object increases. Relief displacement is caused by differences in elevation. If the elevation of the terrain surface is known throughout a scene, the geometric distortion can be rectified.



In the left image the Space Needle in Seattle was located close to the principal point so there is minimal displacement. In the right image the Space Needle is located further away from the principal point causing it the "lean" away from the center.

---

5. Stereoscopy and Aerial Photo Interpretation: Stereoscopy, Pseudoscopy, Mirror Stereoscope, mosaic, edge information, mapping of Physical and Cultural features with the Air photo

interpretation keys: shape, size, pattern, tone, texture, shadow, site and associations.

Stereoscopic Vision and Stereoscopes

For deriving maximum benefit from photographs they are normally studies stereoscopically. A pair of photographs taken from two camera stations but covering some common area constitutes a stereoscopic pair which when viewed in a certain manner gives an impression as if a three dimensional model of the common area is being seen. The basis of this subjective impression is dealt in the end of this lesson.

Depth Perception

Human beings can distinguish depth instinctively. However, there arc many aids to depth perception, for instance, closer objects partly cover distant objects or distant objects appear smaller than similar objects nearby. These aids apply to monocular vision. For short distances binocular vision is more important and is of interest to Photogrammetrists, for it is binocular vision which enables us to obtain a spatial impression of a MODEL formed by two photographs of an object (or objects) taken from different viewpoints.

Normally, our eyes give us two slightly different views, which arc fused physiologically by the brain, and result in a sensation of seeing a model having three dimensions. This three-dimensional effect, due to binocular vision, is very limited however, decreasing rapidly beyond a viewing distance of one metre. Thus it may be concluded that binocular vision is primarily an aid in controlling and directing the movements of one's limbs.

A small percentage of the people do not have the facility of binocular vision and no amount of training will give it to them. Unfortunately, there is no known Physical aid to provide stereoscopic sight to such people who do not possess it naturally, but training can help those having weak fusion.

Requirements of Stereoscopic

Photographs If, instead of looking at the original scene, we observe photos of that scene taken from two different viewpoints, we can, under suitable conditions, obtain a three dimensional impression

from the two dimensional photos. This impression may he very similar to the impression given by the original scene, hut in practice this is rarely so.

In order to produce a spatial model, the two photographs of a scene must fulfill certain conditions:

a) The camera (spatial) axes should be approximately in one plane, though the eyes can accommodate the difference to a limited degree.

b) The ratio B/H, in which B is the distance between the exposure stations and H is the distance between an object point and the line joining the two stations, must have an appropriate value. In aerial photogrammetry this ratio is called the base- height ratio. If this ratio is too small, say smaller than 0.02, we can obtain a fusion of the two pictures, hut the depth impression will not be stronger than if only one photograph was used.

The ideal value of B/H is not known, but is probably not far from 0.25. In photogrammetry, values upto 2 are used, although depending on the object, sometimes much greater values may be appropriated.

c) The scale of the two photographs should be approximately the same. Difference upto 15% may, however, be successfully accommodated. For continuous observation and measurements, differences greater than 5% may be disadvantageous.

d) Each photograph of the pair should be viewed by one eye only, i.e., each eye should have a different view of the common overlay area.

The brightness of both the photographs should be similar.

Such a pair of photograph is known as stereoscopic pair or stereogram.

Stereoscopic vertical photography is the most commonly used one in aerial survey. The terrain is covered with strips of photographs. Overlap between two photographs in the same strip varies from 55 to 90%. Overlap of adjacent strips varies from 5 to 55%. The most usual overlaps are, in the strip, 60% and between two adjacent strips, 15%.

Mathematical Consideration

In order to study how the most natural impression of a stereoscopic pair of photographs is obtained, we will give some definitions. Two convergent photographs in positive positions are shown. The

perspective centres O1 and O2 are on one side and two points P and Q of a scene are behind the positive planes. The line joining the perspective centres O1 and O2 (the base) is called the epipolar axis.

The points of intersection of the epipolar axis with the positive planes are called epipoles (K1 and K2). Plane through the epipolar axis and an object point (plane O1 O2 or plane O1 O2 Q) is called an epipolar plane. Lines of intersection of epipolar planes with the positive planes are called epipolar lines (rays) (K1P1, K2P2). The condition in which the corresponding images of a streopair of photographs lie in the same epipolar plane is called correspondence. Departure from this condition is called want of correspondence.

When we focus our eyes on some object, the eye base and the object are in the same plane. If we press slightly on one of the eye balls upward, we disturb the plane (i.e., disturb the correspondence condition) and we get a double image of the object. The vertical difference between the two images has the character of a vertical parallax or y- parallax. Obviously, a squint-eyed person who constantly has such a defect in his eyes will not be able to see stereoscopically. In precise stereo plotting, instruments provision is generally made to correct for squint.

In natural stereoscopic vision we observe always in an epipolar plane, determined by the two eyes and the object. In artificial stereoscopic vision we have to create this situation. This can be achieved:

a) By giving the photographs relatively the same position as they had during exposure and placing the eyes in the perspective centres.

b) If the photographs arc observed while they arc in one plane (e.g. on that table), by positioning them in such a way that corresponding epipolar lines are collinear and in one plane with the two eyes. Thus, rotation of the photographs is necessary during scanning, if epipolar lines are not parallel. This is the case in convergent photography.

If the optical axes are vertical and the flying height the same, epipolar axis will be parallel to the negative planes. Then the epipolar lines arc all parallel to the epipolar axis (because the epipole is at infinity). Thus vertical photographs need no rotation during scanning.

Binocular observation of stereoscopic photographs

Accommodation and convergence

If we have a pair of stereoscopic photographs in front of us, on paper, glass plates or projected with projectors and they are oriented in such a way that epipolar lines are situated in the way described before we can observe them in different ways. In order to evaluate the different ways of observation, we have to use the terms accommodation and convergence. Accommodation refers to focusing of eye-lens to see objects sharply at different distances. An un-accommodated eye is considered to be focused at infinity.

Convergence refers to the directing of lines of sight (i.e., the optical axes) of the two eyes to the same point. The optical axis of the eye can be changed in direction by rotating the eye in its socket. The angle the eye base subtends at the point is called angle of convergence or parallactic angle.

Normal reading distance is 250 mm, i.e., while reading we accommodate and converge the eyes at this distance. As the eye-base is one an average about 65 mm (2.5 inches) for human eye, the angle of convergence them is approximately 16 degrees. (The line joining the nodes of the eyes is called eye-base or the interocular or interpupillary distance. The relation between the accommodation distance (d) and angle of convergence (in radians) is given by

$$\gamma = \frac{E}{D}$$

E, being the interpupillary distance

Normally accommodation and convergence are automatically linked up. If we look at a point at a certain distance, accommodation and convergence are set for that distance. We can disconnect this link but not without much strain on eyes. A lot of practice is required fur accommodation at a distance other than the distance of convergence.

There are three ways of observation of stereoscopic photographs:

a) Observation with Crossed eye axes

This involves looking with the right eye at the left photograph and with the left eye at the right photograph. The convergence and accommodation are at two different distances, and this type of observation is, therefore, very tiring. Large photographs can be used conveniently by this method, but due to strain on the eye, this method is not used in practice.

b) Observation with parallel eye axes

This method is possible without any optical aids, but is tiring as well as the eyes are converged on infinity, yet accommodating at approximately 250 mm (Fig. 13(b). It is less tiresome if positive lenses are placed between the eyes and the photographs so that the photos are placed at the focal length of the lenses. The accommodation then corresponds with the convergence and the eyes are viewing naturally. The 'pocket- stereoscope' was developed on this principle.

c) Observation with convergent eye-axes

When the accommodation and convergence arc at the same distance the viewing is least tiring and this is the normal method of viewing. But in order to view the photos stereoscopically they must be superimposed, such that the point A and the corresponding point A' on the other photo lie at the point of convergence.

The images have to be separated so that left eye sees only the left hand photographs and the right eye only the right hand photograph. The resulting stereoscopic perception is similar to that of normal three dimensional perceptions. The separation may he achieved by colour filters or by polarized filters.

There is an interesting phenomenon in Stereoscopy. In viewing terrain in aerial photography a reversal of the relief is sometimes obtained by the eyes. Such a phenomenon is known as pseudoscopic illusion or Pseudo copy. Such an impression can be obtained by viewing the photos with crossed eye axes. Sometimes, viewing with the shadows in case of excessive relief (e.g. hills) away from the observer can also result in pseudoscopy. So, in the initial stages, to avoid pseudoscopic view, it is desirable to view the photographs with shadows of objects falling towards the observer.

Separation by colour filters

The photos are either projected or printed in two different colours. By placing a filter of the same colour over each eye corresponding picture is observed by one eye only. In practice this problem is difficult to solve completely. The human eye is sensitive for light with wavelength from 400 to 720 milli-microns (mu). The vertex lies at about 560 mu. A possibility for separation of the two superimposed images would be to use filters of which one cuts off all wave length over 560 mu (its colour would be blue-green) and

b) Observation with parallel eye axes

This method is possible without any optical aids, but is tiring as well as the eyes are converged on infinity, yet accommodating at approximately 250 mm.It is less tiresome if positive lenses are placed between the eyes and the photographs so that the photos are placed at the focal length of the lenses. The accommodation then corresponds with the convergence and the eyes are viewing naturally. The 'pocket- stereoscope' was developed on this principle.

c) Observation with convergent eye-axes

When the accommodation and convergence arc at the same distance the viewing is least tiring and this is the normal method of viewing. But in order to view the photos stereoscopically they must be superimposed, such that the point A and the corresponding point A' on the other photo lie at the point of convergence.

The images have to be separated so that left eye sees only the left hand photographs and the right eye only the right hand photograph. The resulting stereoscopic perception is similar to that of normal three dimensional perceptions. The separation may he achieved by colour filters or by polarized filters.

There is an interesting phenomenon in Stereoscopy. In viewing terrain in aerial photography a reversal of the relief is sometimes obtained by the eyes. Such a phenomenon is known as pseudoscopic illusion or Pseudo copy. Such an impression can be obtained by viewing the photos with crossed eye axes. Sometimes, viewing with the shadows in case of excessive relief (e.g. hills) away from the observer can also result in pseudoscopy. So, in the initial stages, to avoid pseudoscopic view, it is desirable to view the photographs with shadows of objects falling towards the observer.

Separation by polarized filters

Light has the characteristics of a wave motion in which the waves vibrate in all possible planes perpendicular to the direction of prorogation. These are called transverse waves. It is possible to analyse the transverse waves into separate components along two axes perpendicular to each other and to the direction of propagation by means of filters.

For stereoscopic vision the filters are placed so that polarized light rays forming the left image are at right angles to the light rays forming the right image. There are several advantages in using polarized light:

- Light loss is about 50% only in both projections,

- There is not colour contrast between the two picture, and

- It is possible to use colour photography on this principle

However, there is one big disadvantage in using the method, which has so far prevented its use in photogrammetry. With the type of plotting instrument which uses this system, it is important that the screen on which the image is projected be diffuse, so that it can be viewed equally well from all directions but a diffuse surface acts as a depolarizer and so no stereoscopic image would be apparent.

Stereoscopes

The function of a stereoscope is to deflect normally converging lines-of-sight so that each eye views a different photographic image.

Stereoscopes are grouped into 2 basic types:

i) Lens stereoscopes

ii) Mirror prism stereoscopes

Pocket Stereoscope

By far the most popular is the lens stereoscope commonly known as pocket stereoscope. The pocket stereoscope usually has plane-convex lens, upper side flat with a focal length of 100 mm. The rays entering the eyes are now parallel and converge at infinity and have been accommodated (focused) at 100 mm distance.

Since the normal viewing distance is 250 mm, a closer view, i.e. at 100 mm results in a magnification. The magnification is then 250/100 = 2.5. More expensive types have a changeable eye-base. Such a refinement is not necessary for operators with an average eye-base range of 60 to 68 mm. The pocket stereoscope is cheap, transportable, and has a large field of view. It has two big disadvantages:

a) Limited magnification. Pocket stereoscopes with more than three times magnification cannot be equipped with simple plane-convex lenses, due to the too large and increase in lens aberrations. In

addition the distance between the head and the photos becomes too small for adequate illumination without undue complications.

b) The distance between corresponding points on the photos must be equal to or smaller than the eye-base. With normal size photographs this becomes difficult or impossible without bending or folding the photos. It should not be forgotten, however, that due to the simple optical system the image quality of the pocket stereoscope is very good.

Mirror Stereoscope

The two above mentioned drawbacks have led to the development of the mirror stereoscope. The normal size photos (23 cm x 23 cm) can be separated and seen under the stereoscope without folding them. The path of the bundle of rays has been diverted and brought to the eyes at 65 mm separation. This is achieved by reflecting mirrors. Normally the distance between corresponding points is kept at 240 mm so that photographs are placed separately, i.e., it effectively increases the eye base from 65 mm to 240 mm. As in pocket stereoscope the picture must be at the focal plane of the lenses in order to have convergence at infinity. The mirrors M 1 are placed in such a way that the picture distance via the small mirrors M2 (generally prisms) become equal to the focal length of the lens, usually 300 mm. This gives approximately 250/300 = 0.8 x magnification, or rather reduction the picture observed, to magnify the image additional oculars of magnification 3x to 8x can be used over the prisms or a lens placed before each prism giving a magnification of about 1.8x.

Classification of aerial photography

There are different criteria to classify aerial photographs depending upon the scale, tilt, coverage, film and spectral coverage/response. These classifications can be defined as follows:

Scale

- Large scale: between 1:5,000 and 1:20,000
- Medium scale: between 1:20,000 and 1:50,000
- Small scale: smaller than 1:50,000 (Scale classification may differ from country¬ to country)

TILT

- Vertical: when the tilt is within n 3 (nearly vertical)
- Oblique: Low oblique (horizon does not appear but tilt is more than 3^{o}), high oblique (horizon appears)

- Horizontal or terrestrial: camera axis is kept horizontal.
- Angular Coverage
- Narrow angle: angle of coverage less than 50 degrees
- Normal angle: angle of coverage of 60 degrees.
- Wide angle: angle of coverage of 90 degrees
- Super-wide angle: angle of coverage of 120 degrees

Film

- Black and white panchromatic
- Black and white infrared

Colour

- Colour infra-red/false colour
- Obtaining aerial photography

As per the existing policy of the Government of India, all types of aerial photographs are classified documents (secret or restricted), depending upon the location and its strategic importance. The Surveyor General of India coordinates all activities relating to the execution of aerial photographic tasks for all civilian needs. The coordinating authority performs the following functions:

Design and issue of the specifications for photographic tasks

1. Layout and priorities, clearance from various agencies and distribution of tasks among the three flying agencies
2. Flight planning and evaluation for suitability of the executed tasks
3. Distribution of photographs to the indenter
4. Accounting for the above

Flying agencies

As the coordinating agency does not have its own flying facilities, the flying operations for aerial photography are earned out by the Indian Air Force; the Air Survey Company, Dum Dum, Calcutta and the National Remote Sensing Agency (NRSA), Hyderabad.

Specifications of aerial photography

For planning fresh photography, the purpose of the photography and scale are the main considerations. However, while defining these specifications, the following factors should be kept in view. Unless otherwise specified, the overlaps should be kept 60 per cent in the forward direction and 25 per cent in the lateral direction. For special tasks and terrains, the overlaps can be increased to 80 percent in the forward direction and 50 to 60 per cent in the lateral direction, especially in steep hilly areas and in city centers with high-rise buildings.

- **Camera lens:** depending on the type of photography required
- Film/filter combination: depending on the type of photography required
- **Shutter speed:** depending on the scale, type of aircraft, its speed and film speed/aperture (between 1/100 to 1/1,000 seconds)
- **Image motion:** to be kept within tolerable limits (i.e. 20 µm on the negative scale) by the proper combination of shutter speed/aperture and speed of aircraft
- **Camera frame:** stable mounts
- **Platforms:** ceiling height, stability in flying and speed limits
- **Auxiliary data:** as required
- **Processing:** depending on the film type and the requirements of the data products.

Mapping of Physical and Cultural features with the Air photo interpretation keys: shape, size, pattern, tone, texture, shadow, site and associations

Aerial photo interpretation is a process of examining and extracting useful information from aerial photographs. During this process, some features may be easily identifiable while others may not, depending upon your own perceptions and experience. The reliability of information collected from aerial photographs depends on the quality of aerial photographs, instruments used for interpretation, working conditions and personal experience with photo interpretation techniques. In addition, preliminary knowledge of the area of interest which comprises of its geographic location, past and present climate conditions, vegetation and published literature are always useful for accurate identification of features.

Aerial photography has many applications such as cartography, urban and rural planning, environmental impact studies, civil law cases, real estate evaluations and can even be used as wall art. Following are the advantages of aerial photography:

• **Synoptic Viewpoint:** Aerial photographs give a bird's eye view of large areas enabling us to see surface features in their spatial context.

• **Permanent Recording:** They are virtually permanent records of the existing conditions on the Earth's surface at one point in time and are used as an historical document.

• **Capability to Stop Action:** They provide a view of dynamic conditions and are useful in studying phenomena e.g., flooding, wildlife, oil spills, etc.

• **Three Dimensional Perspectives:** It provides a stereoscopic view of the Earth's surface and makes it possible to take measurements horizontally and vertically.

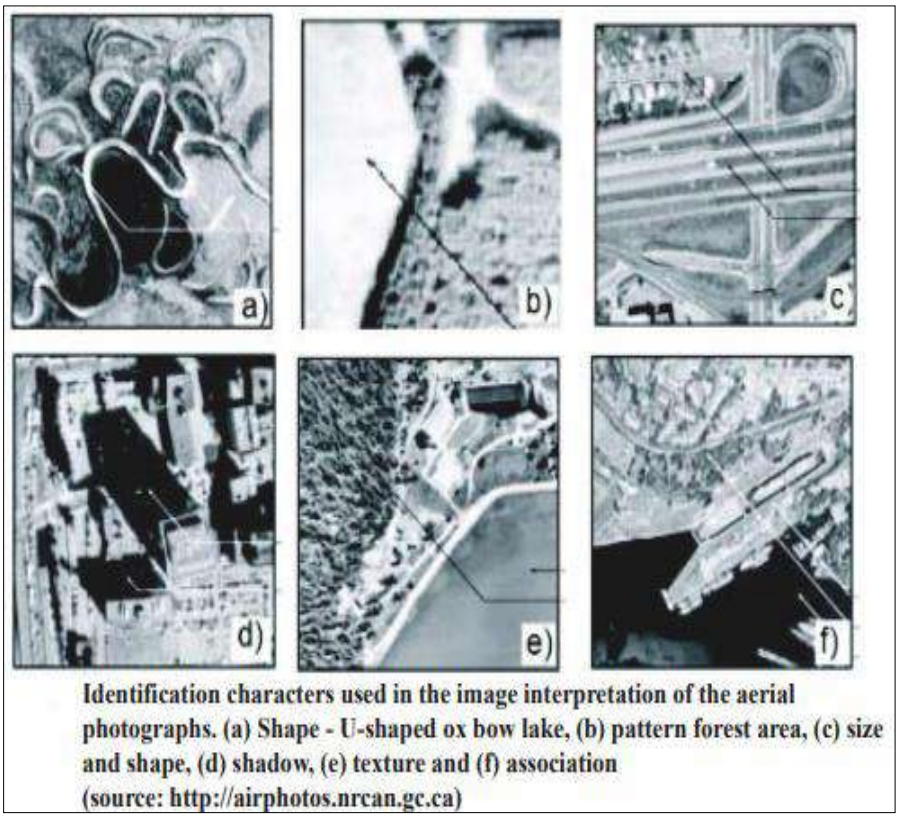
• **Spectral and Spatial Resolution:** Aerial photographs are sensitive to radiation in wavelengths that are outside of spectral sensitivity of the human eye. They also have better spatial resolution than many ground based remote sensing methods.

• **Availability:** They are readily available at a range of scales for much of the world.

• **Economy:** They are much cheaper than field surveys and are often cheaper and more accurate than maps. The aerial image differs from everyday photograph in the following aspects:

- Overhead perspective
- Beyond visible light spectrum and
- Unfamiliar scales and orientation.

Tone is the most basic of the interpretive element and refers to the relative brightness or colour of elements on an aerial photograph. Size of objects must be considered in the context of the scale of a photograph. The scale will help you to determine if a water body is a pond or lake or sea. Shape refers to the general outline of objects and regular geometric shapes. Texture is the impression of smoothness or roughness of image features and is caused by the frequency of change of tone in photographs. Pattern or spatial arrangement is formed by objects in a photo which can be diagnostic. Shadows aid interpreters in determining height of objects in aerial photographs. Site refers to topographic or geographic location. Association refers to position of the objects of interest in relation with the other objects.



Identification characters used in the image interpretation of the aerial photographs. (a) Shape - U-shaped ox bow lake, (b) pattern forest area, (c) size and shape, (d) shadow, (e) texture and (f) association
(source: http://airphotos.nrcan.gc.ca)

To interpret aerial photographs, a number of sophisticated instruments such as Visual Image Interpretation pocket stereoscope, mirror stereoscope, or plotter are used for measuring area, height and slopes of different parts of the Earth.

Stereoscope is used for viewing the area in 3-dimension and is important for determining topographical relief of an area, as well as the height of objects such as trees and building.

Stereoscopic imagery is the result of overlap (generally 60%), which is the amount by which one photograph includes an area covered by a neighbouring photograph.

For mapping, inventory and vegetation studies, for example, a survey is flown in a series of to-and-fro parallel strips with side overlaps between strips over the entire area.

For non-stereoscopic coverage, used in crop sampling or pollution detection, the photographer may choose a 20% forward overlap.

Image Interpretation Tasks

The image interpretation procedure is a complex task and requires several tasks to be conducted in a methodical manner which include:

- Classification
- Enumeration
- mensuration and
- Delineation.

1. Classification is the assignment of object, features, or area to the classes based on their appearance on the images. Often the distinctions are made between three levels of confidence and precision namely- detection, recognition and identification. Detection is the determination of presence or absence of the feature. Recognition implies a higher level of knowledge about a feature or an object such that the object can be assigned identity. And, identification means that the identity of an object or feature can be specified with enough confidence and detail to place it in a specific class.
2. Enumeration is the task of listing or counting discrete items visible on an image.
3. Mensuration or measurement is an important function in many image interpretation problems. Two kinds of measurements are important, first, is the measurement of distance and height, and by extension, volumes and areas as well. A second form of measurement is quantitative assessment of image brightness.
4. Finally, the interpreter must delineate, or outline, regions as observed on remotely sensed images. The interpreter must be able to separate distinct aerial units that are characterised by specific tones and textures and to identify edges or boundaries. The image analyst may simultaneously apply several of these skills in examining an image. Recognition, delineation and mensuration may all be required as the interpreter examines an image.

Prerequisites for Image Interpretation

Now you know that following are the requirements for image interpretation:

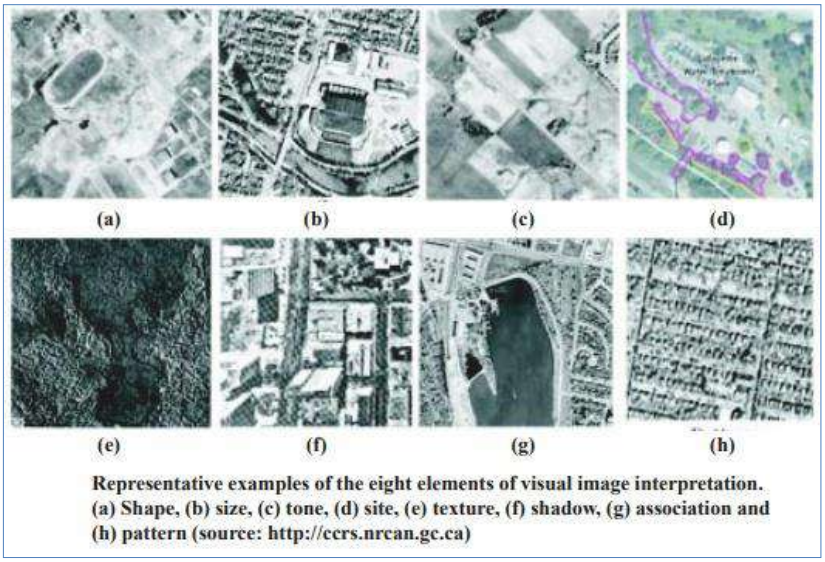
- Remote sensing system
- Knowledge of image and sensor characteristics
- Proficiency based on knowledge of the subject and
- Adequate familiarity of the geographic region and locality

Elements of Visual Image Interpretation

You have read in the previous section that visual interpretation of aerial photographs involves the study of various basic characteristics of an object. In case of interpretation of satellite images, these characteristics of objects are studied with reference to a single or multiple spectral bands because

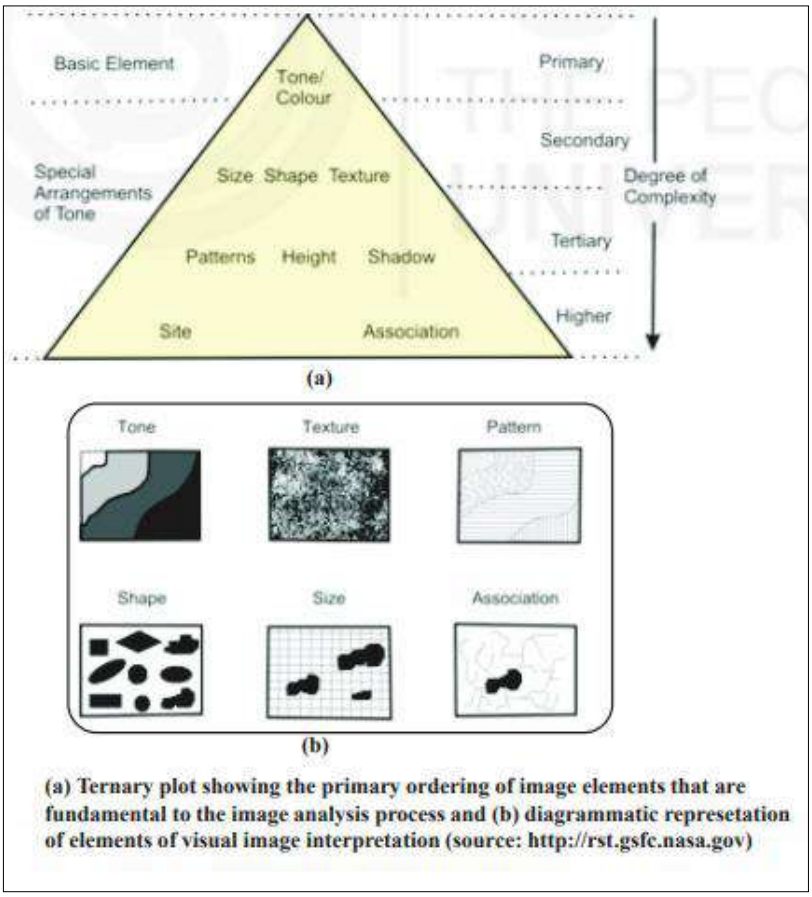
there are generally more than one images acquired in different spectral regions of electromagnetic spectrum. However, the basic elements are tone, texture, shape, size, pattern, shadow, location and association, similar to those used in aerial photo interpretation. Image interpretations employ combination of the following eight elements:

- tone
- size
- shape
- texture
- association
- shadow
- site and
- pattern



Representative examples of the eight elements of visual image interpretation. (a) Shape, (b) size, (c) tone, (d) site, (e) texture, (f) shadow, (g) association and (h) pattern (source: http://ccrs.nrcan.gc.ca)

A systematic study and visual interpretation of satellite images usually involves consideration of two basic elements, namely image elements and terrain elements. Out of the eight elements listed above, the first seven elements comprise image elements and the 8th element; pattern is the terrain element such as drainage, landform, erosion, soil, vegetation and land-use patterns. These elements are shown in the order of their complexity in following figure.



(a) Ternary plot showing the primary ordering of image elements that are fundamental to the image analysis process and (b) diagrammatic represetation of elements of visual image interpretation (source: http://rst.gsfc.nasa.gov)

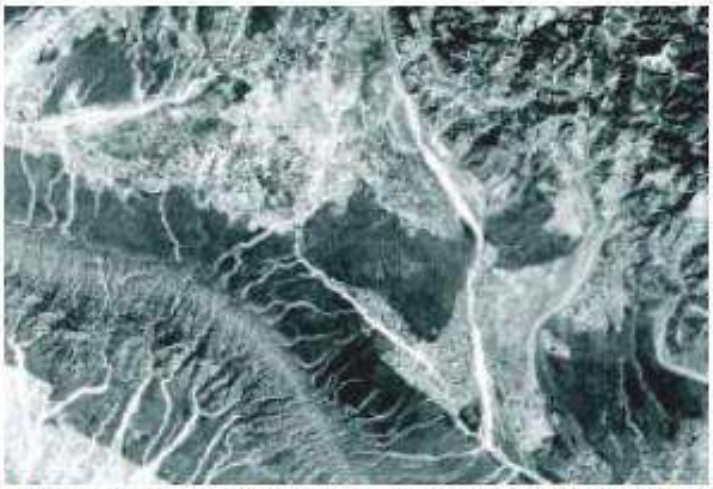
We shall now discuss the elements of image interpretation.

Tone

Tone refers to the colour or relative brightness of an object in colour image and the relative and quantitative shades of gray in black and white image. Tone is one of the most basic elements because it is difficult to discern other elements without tonal differences. Tone in aerial photographs is influenced by the following factors:

- Light reflectivity of the object
- Angle of reflected light
- The geographic latitude
- Type of photography and film sensitivity
- Light transmission of filters and
- Photographic processing.

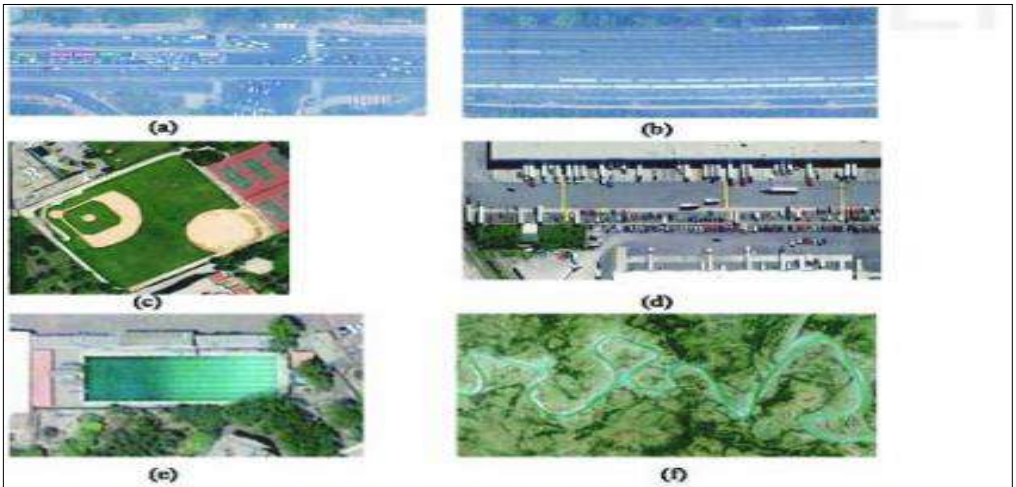
Strong tonal contrasts on satellite imageries are always desirable for better image interpretation. You can observe different tones for different features.



Satellite image showing Doon valley and surroundings. The drainage patterns and lithological differences can be clearly observed (source: Rao, 2002)

Size

Objects can be misinterpreted if their sizes are not evaluated properly. Size of objects in an image is a function of scale hence, the size of objects must be considered in the context of the scale of a photograph/image. Size of an object can be important tool for its identification. The size of an object or feature is relative in relation to other objects on the image. This is probably the most direct and important function of size, as it provides the interpreter with an intuitive notion of the scale and resolution of an image even though no measurements or calculations may have been made. This role is achieved by recognition of familiar objects like dwellings, highways and rivers. You should remember that size of an object in an image depends on the scale and resolution of the image.



Variation in size and shapes in the images provides clue for different objects. (a) Automobiles, (b) railway track, (c) baseball court, (d) trailer, (e) swimming pool and (f) a meandering river (source: www.earth.google.com)

Shape

Shape relates to the general form, configuration or outline of an individual object. Shape is one of the most important single factors for recognising objects from images. Regular geometric shapes are usually indicators of human presence and use. Similarly, irregular shapes are usually indicators of natural objects. Some objects can be identified almost solely on the basis of their shapes. For example, a railway line is usually readily distinguished from a highway or an un-metalled road because its shape consists of long straight tangents and gentle curves as opposed to the shape of highway. Features in nature often have such distinctive shapes that shape alone might be sufficient to provide clear identification e.g., beach, ponds, lakes and rivers occur in specific shapes unlike others found in nature.

Texture

Texture is an expression of roughness or smoothness as exhibited by the images. It is the rate of change of tonal values (frequency of tonal changes). Texture signifies the frequency of change and arrangement of tones in an image and is produced by an aggregate of unit features too small to be clearly recognised individually on an image. Texture can be expressed qualitatively as coarse, moderate, fine, very fine, smooth, rough, rippled and mottled. It is rather easier to distinguish various textural classes visually than in the digital oriented techniques. Texture is, thus, dependent upon tone, shape, size, pattern, and scale of the imagery, and, is produced by a mixture of features that are too small to be seen individually. For example, grass and water generally appear ‘smooth’ while trees or a forest canopy may appear ‘rough’.

Association

Association is occurrence of features in relation to its surroundings. Sometimes a single feature by itself may not be distinctive enough to permit its identification. It specifies the occurrence of certain objects or features in association of a particular object or feature.

Many features can be easily identified by examining the associated features. For example, a primary school and a high school may be similar flat roofed building structures but it may be possible to identify the high school by its association with an adjacent football field.

Shadow

Shadow is an especially important clue in the interpretation of objects in the following two ways:

• The outline or shape of a shadow provides a profile view of objects, which aids in image interpretation and

• Objects within shadow reflect little light and are difficult to discern on Visual Image Interpretation image, which hinders interpretation.

Taller features cast larger shadows than shorter features. Military image interpreters are often primarily interested in identification of individual items of equipment. Shadow is significant in distinguishing subtle differences that might not be otherwise visible.

Taller objects such as the Qutub Minar cast larger shadow than smaller objects such as buildings and trees (source:earth.google.com)

Site

Site refers to the topographic position, for example, sewage treatment facilities are positioned at low topographic sites near stream or rivers to collect waste flowing through the system from higher locations. The relationship of feature to the surrounding features provides clues towards its identity. You can also consider the example of certain tree species located in areas of specific altitudes. Similarly, identification of landforms can help in deciphering the underlying geology. Often many of the rock types have distinct topographic expressions, for example, some kinds of sedimentary rocks are typically exposed in the form of alternating ridge and valley topography.

Pattern

You have read about the seven image elements. It is now time to discuss about the terrain element which is also a significant element in image interpretation. The terrain elements include drainage, topography/landform, soil, vegetation and land use planning patterns.

Pattern develops in an image due to spatial arrangement of objects. Hence, pattern can be defined as the spatial arrangement of objects in an image. Certain objects can be easily identified because of their pattern. A particular pattern may have its genetic relation with several factors of its origin. For example, urban and rural settlement areas can be easily identified based on the patterns created by the rows of houses or buildings. Similarly, drainage patternhave orderly association with the underlying lithology, structure, soil texture and hydrological characteristics of the ground and hence provide clues about them.

Typical adjectives associated with interpretation elements (source: Bhatta, 2010)

Element	Common adjectives (quantitative and qualitative)
Location	x,y coordinates: longitude and latitude or meters, easting and northing in a map grid
Size	Length, width, perimeters, area: small, medium (intermediate) and large
Shape	An object's geometric characteristics: linear, curvilinear, circular, elliptical, radial, square, rectangular, triangular, hexagonal, pentagonal, amorphous, etc
Shadow	A silhouette caused by solar illumination from the side
Tone/colour	Gray tone: light (bright), intermediate (gray), dark (black) colour = intensity, hue, saturation
Texture	Characteristics placement and arrangement of repetition of tone or colour; smooth, intermediate (medium), rough (coarse), mottled, stippled
Pattern	The spatial arrangement of objects on the ground; systematic, unsystematic or random, linear, curvilinear, rectangular, circular, elliptical, parallel, centripetal, serrated, striated, braided
Height/depth/ volume/aspect	Z-elevation (height), depth (bathymetry), volume, slope, aspect
Site/situation/ association	Site: elevation, slope, aspect, exposure, adjacency to water, transportation, utilities Situation: objects are placed in a particular order or orientation relative to one another association: related phenomena are usually present

****

GEO 296.2: COMPUTER BASICS AND APPLICATIONS

1. Computer components: Hardware and software: CPU, Input and Output devices; Common computer languages, System Software, Application Software and Operating Systems.

2. Representation of data; Numbering Systems; Binary Arithmetic; Basic Logic Gates; Boolean Logic and Reduction Techniques.

3. Computation, Storing and Formatting Spreadsheets: Computation of Rank, Mean, Median, Mode, Standard Deviation, Moving Averages, Sample Variation; Derivation of Correlation, Covariance and regression; Selection of technique and interpretation using MS-Excel and SPSS Environment.

4. Regression, correlation, curve fitting, multivariate analysis.

5. Internet Surfing- generations of data and extraction of information for power-point presentation, Manipulation and editing of graphic files.

Computer Components

1. A. HARDWARE

Computer hardware is the collection of physical parts of a computer system. This includes the computer case, monitor, keyboard, and mouse. It also includes all the parts inside the computer case, such as the hard disk drive, motherboard, video card, and many others. Computer hardware is what you can physically touch.

1. B. SOFTWARE

Computer software, also called software, is a set of instructions and its documentations that tells a computer what to do or how to perform a task. Software includes all different software programs on a computer, such as applications and the operating system.

- **B. i. SYSTEM SOFTWARE**

System software is software designed to provide a platform for other software. Examples of system software include operating systems like macOS, GNU/Linux and Microsoft Windows, computational science software, game engines, industrial automation, and software as service applications.

- **B. ii. APPLICATION SOFTWARE**

Application software (app for short) is a program or group of programs designed for end users. Examples of an application include a word processor, a spreadsheet, an accounting application, a web browser, an email client, a media player, a file viewer, an aeronautical flight simulator, a console game or a photo editor. The collective noun application software refers to all applications collectively. This contrasts with system software, which is mainly involved with running the computer.

1. C. Operating System.

An operating system, or "OS," is software that communicates with the hardware and allows other programs to run. It is comprised of system software, or the fundamental files your computer needs to boot up and function. ... Common desktop operating systems include Windows, OS X, and Linux.

1. D. CPU

A central processing unit, also called a central processor or main processor, is the electronic circuitry within a computer that executes instructions that make up a computer program. The CPU performs basic arithmetic, logic, controlling, and input/output operations specified by the instructions.

1. E. INPUT DEVICE

An input device is any hardware device that sends data to a computer, allowing you to interact with and control it. The picture shows a Logitech trackball mouse, which is an example of an input device.

1. F. OUTPUT DEVICE

An output device is any device used to send data from a computer to another device or user. Most computer data output that is meant for humans is in the form of audio or video.

1. G. COMPUTER LANGUAGE

Computer language is a system of communication with a Computer. There are several type of Computer Language,-

1. G. i. PROGRAMMING LANGUAGE

A programming language is a type of written language that tells computers what to do in order to work. Programming languages are used to make all the computer programs and computer software. ... Usually, the programming language uses real words for some of the commands, so that the language is easier for a human to read.

1. G. ii. GENERAL-PURPOSE LANGUAGE

In computer software, a general-purpose programming language is a programming language designed to be used for writing software in the widest variety of application domains.

1. G. iii. COMMAND LANGUAGE

A command language is a language for job control in computing. It is a domain-specific and interpreted language; common examples of a command language are shell or batch programming languages.

1. G. iv. MACHINE LANGUAGE

Sometimes referred to as machine code or object code, machine language is a collection of binary digits or bits that the computer reads and interprets. ... A computer cannot directly understand the programming languages used to create computer programs, so the program code must be compiled.

1. G. v. ASSEMBLY LANGUAGE

In computer programming, assembly language, often abbreviatedasm, is any low-level programming language in which there is a very strong correspondence between the instructions in the language and the architecture's machine code instructions.

1. G. vi. MARKUP LANGUAGE

In computer text processing, a markup language is a system for annotating a document in a way that is syntactically distinguishable from the text, meaning when the document is processed for display, the markup language is not shown, and is only used to format the text.

1. G. vii. STYLE SHEET LANGUAGE

A style sheet language, or style language, is a computer language that expresses the presentation of structured documents. One attractive feature of structured documents is that the content can be reused in many contexts and presented in various ways.

1. G. viii. CONFIGURATION LANGUAGE

The Configuration Language is an application-level language that allows programmers to define the structure of configuration files, its format, interdependents and limitations among parameters in human-readable view. ... Human-readable format of files for users

1. G. ix. CONSTRUCTION LANGUAGE

Construction language, a general category that includes configuration languages, toolkit languages, and programming languages. Programming language, a formal language designed to communicate instructions to a machine, particularly a computer.

1. G. x. QUERY LANGUAGE

Query language (QL) refers to any computer programming language that requests and retrieves data from database and information systems by sending queries. It works on user entered structured and formal programming command based queries to find and extract data from host databases.

1. G. xi. MODELING LANGUAGE

Modeling language is any graphical or textual computer language that provisions the design and construction of structures and models following a systematic set of rules and frameworks. Modeling language is part of and similar to artificial language.

Representation of Data

2. A. NUMBER SYSTEM

A number system in computer ideology is regarded as the method or system of numbering and representing of digits in the computer 'inner' system.

In other words, it is a technique used in representing numbers in the computer system architecture. The digital computer represents all kinds of data and information in binary numbers. This implies every value/number that you are saving or feeding into/fetching from the computer system memory has a defined number system. The value/data feed in/fetch from can includes but not limited to: audio, graphics, video, text file, numbers etc.

The total number of digits used in a number system is called its base or radix. The base is written after the number as subscript; for instance 10001102 (1000110 base 2), 5610 (56 to base of 10), 718 (71 base 8) etc.

Computer architecture supports following number systems.-

- Binary number system (Base 2)
- Octal number system (Base 8)
- Decimal number system (Base 10)
- Hexadecimal number system (Base 16)
- Let us see each number system one by one-

2. A. i. BINARY NUMBER SYSTEM

A Binary number system has only two digits, which are 0 and 1. Every number (value) is represented with 0 and 1 in this number system. The base of binary number system is 2, because it has only two digits. Though DECIMAL (No 3) is more frequently used in Number representation, BINARY is the number system form which the system/machine accepts.

2. A. ii. OCTAL NUMBER SYSTEM

Octal number system has only eight (8) digits from 0 to 7. Every number (value) is represented with 0,1,2,3,4,5,6 and 7 in this number system. The base of octal number system is 8, because it has only 8 digits.

3. A. iii. DECIMAL NUMBER SYSTEM

Decimal number system has only ten (10) digits from 0 to 9. Every number (value) is represented with 0,1,2,3,4,5,6, 7,8 and 9 in this number system. The base of decimal number system is 10, because it has only 10 digits.

4. A. iv. HEXADECIMAL NUMBER SYSTEM

A Hexadecimal number system has sixteen (16) alphanumeric values from 0 to 9 and A to F. Every number (value) represents with 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E and F in this number system. The base of hexadecimal number system is 16, because it has 16 alphanumeric values. Here, we have 0 to 9, representing 0 – 9 but from 10, we have A is 10, B is 11, C is 12, D is 13, E is 14 and F is 15.

The table above shows the sample representations-

Number system Base Used digits Example

- Binary 2 0, 1(11110000)2
- Octal 8 0, 1,2,3,4,5,6,7 (360)8
- Decimal 10 0,1,2,3,4,5,6,7,8,9 (240)10
- Hexadecimal 16 0,1,2,3,4,5,6,7,8,9, A, B, C, D, E, F (F0)16

Number System Conversions

There are three types of conversion:

• Decimal Number System to Other Base [for example: Decimal Number System to Binary Number System e.g. Base 10 to Base 2 etc.]

• Other Base to Decimal Number System [for example: Binary Number System to Decimal Number System e.g. Base 2 back to Base 10 etc.]

• Other Base to Other Base [for example: Binary Number System to Hexadecimal Number System e.g. Base 2 to Base 16 etc.]

Let's pick them one after the other to see how the computations are done and the underlying logic behind them!

1. Decimal Number System to Other Bases

The under listed are the steps/procedures:

A) Divide the Number (Decimal Number) by the base of target base system (in which you want to convert the number to e.g. Binary (2), Octal (8) OR Hexadecimal (16)).

B) Write the remainder from step 1 as a Least Signification Bit (LSB) to Step last as a Most Significant Bit (MSB); that is, write from down-up.

Example 1: Convert 1234510 to Base 2

Solution 1: Decimal to Binary Conversion Result

Decimal Number is: (12345)10 Binary Number is (11000000111001)2

Example 2: Convert same no (1234510) this time, to Base 8

Solution 2: Decimal to Octal Conversion Result

Decimal Number is: (12345)10 Octal Number is (30071)8

Example 3: Convert 1234510 to Base 16

Solution 3: Decimal to Hexadecimal Conversion Result

Example 1

Decimal Number is: (12345)10 Hexadecimal Number is (3039)16

2. B BINARY ARITHMETIC

In binary number system there are only 2 digits 0 and 1, and any number can be represented by these two digits. The **arithmetic of binary numbers** means the operation of addition, subtraction, multiplication and division. **Binary arithmetic** operation starts from the least significant bit i.e. from the right most side.

2. B. i. BINARY ADDITION

There are four steps in binary addition, they are written below

0 + 0 = 0

0 + 1 = 1

1 + 0 = 1

1 + 1 = 0 (carry 1 to the next significant bit)

```
    1
 10001001
 10010101
---------
100111110
```

2. B.ii. BINARY SUBTRACTION

Here are too four simple steps to keep in memory

0 – 0 = 0

0 – 1 = 1, borrow 1 from the next more significant bit

1 – 0 = 1

1 – 1 = 0

```
 10101010
 10100010
---------
 00001000
```

2. B. iii. BINARY MULTIPLICATION

Binary multiplication may sound like it would be more difficult than binary addition or subtraction – but is actually a simple process. Here are the four steps to be followed, using the same binary numbers 10001001 and 10010101:

0×0=0

1×0=0

0×1=0

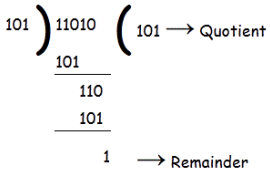
1×1=1 (there is no carry or borrow for this)

```
  1001
X  101
------
  1001
 0000
1001
------
101101
```

2. B. iii. BINARY DIVISION

Binary division is comprised of other two binary arithmetic operations, multiplication and subtraction; an example will explain the operation more easily.

Here 101 is the quotient and 1 is the remainder.

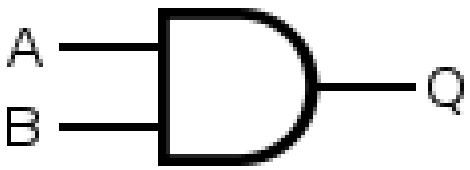


2. C. BASIC LOGIC GATE

A logic gate is an idealized or physical electronic device implementing a Boolean function, a logical operation performed on one or more binary inputs that produces a single binary output.

2. C .i. AND GATE

The AND gate is a basic digital logic gate that implements logical conjunction - it behaves according to the truth table to the right.



INPUT		OUTPUT
A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

A HIGH output (1) results only if all the inputs to the AND gate are HIGH (1). If none or not all inputs to the AND gate are HIGH, a LOW output results. The function can be extended to any number of inputs.

2. C. ii. OR GATE

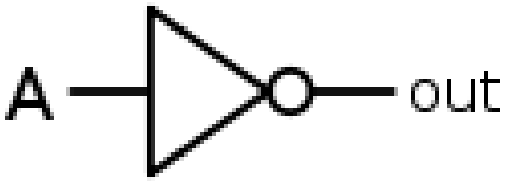
The OR gate is a digital logic gate that implements logical disjunction – it behaves according to the truth table to the right. A HIGH output (1) results if one or both the inputs to the gate are HIGH (1). If neither input is high, a LOW output (0) results.

INPUT		OUTPUT
A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

2. C. ii. NOT GATE

A NOT gate, often called an inverter, is a nice digital logic gate to start with because it has only a single input with simple behavior. A NOT gate performs logical negation on its input. In other words, if the input is true, then the output will be false. Similarly, a false input results in a true output.

INPUT	OUTPUT
A	NOT A
0	1
1	0



2. D. BOOLEAN LOGIC

Named after the nineteenth-century mathematician George Boole, Boolean logic is a form of algebra in which all values are reduced to either TRUE or FALSE. Boolean logic is especially important for computer science because it fits nicely with the binary numbering system, in which each bit has a value of either 1 or 0.

- **Laws of Boolean**

The basic Laws of Boolean Algebra can be stated as follows: Commutative Law states that the interchanging of the order of operands in a Boolean equation does not change its result.

For example: OR operator → A + B = B + A.

AND operator → A * B = B * A.

2. E. REDUCTION TECHNIQUES

In computability theory and computational complexity theory, a reduction is an algorithm for transforming one problem into another problem. A sufficiently efficient reduction from one problem to another may be used to show that the second problem is at least as difficult as the first.

Intuitively, problem A is reducible to problem B if an algorithm for solving problem B efficiently (if it existed) could also be used as a subroutine to solve problem A efficiently. When this is true, solving A cannot be harder than solving B. "Harder" means having a higher estimate of the required

computational resources in a given context (e.g., higher time complexity, greater memory requirement, expensive need for extra hardware processor cores for a parallel solution compared to a single-threaded solution, etc.).

A very simple example of a reduction is from *multiplication* to *squaring*. Suppose all we know how to do is to add, subtract, take squares, and divide by two. We can use this knowledge, combined with the following formula, to obtain the product of any two numbers:

$$a \times b = \frac{\left((a+b)^2 - a^2 - b^2\right)}{2}$$

COMPUTATION, STORING AND FORMATTING SPREADSHEETS

3. A. EXEL ENVIRONMENT

- **Requirements:**

Desktop or laptop with Microsoft Excel installed.

3. A. i. Computation of Rank

A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In mathematics, this is known as a weak order or total preorder of objects.

Calculation- (example)

=Rank (individual cell id, starting cell id: Ending cell id) enter

=Rank (B2, $B$2:$B$11) enter

3. A. ii. Computation of Mean

The statistical mean refers to the mean or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the mean or the average.

Calculation- (example)

=Average (Starting Cell id: Ending Cell id) enter

=Average (B2:B10) enter

3. A. iii. Computation of Median

The median is a simple measure of central tendency. To find the median, we arrange the observations in order from smallest to largest value. If there are an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

Calculation- (example)

=Median (Starting Cell id: Ending Cell id) enter

=Average (B2:B10) enter

3. A. iv. Computation of Mode

The mode of a set of data values is the value that appears most often. If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

Calculation- (example)

=Mode (Starting Cell id: Ending Cell id) enter

=Mode (B2:B10) enter

OR

Data Analysis - Descriptive Statistics – ok – input range (starting cell id: Ending cell id/ $B$2:$B$10) – check column/ row – check labels in first row - check output range - check summary statistics- enter.

3. A. v. Computation of Standard Deviation

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are

further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

Calculation- (example)

=STDEVA (Starting Cell id: Ending Cell id) enter

= STDEVA (B2:B10) enter

3. A. v. Computation of Moving Average

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set.

Calculation- (example)

=MOVING AVERAGE (Ft) = (Actual value of cell under calculation/ No. of observation under calculation)

Months	Temperature	Ft
January	20	(S+J+F)/3
February	18	(J+F+M)/3
March	22	(F+M+A)/3
April	23	(M+A+M)/3
May	28	(A+M+J)/3
June	24	(M+J+J)/3
July	35	(J+J+A)/3
August	39	(J+A+S)/3
September	39	(A+S+J)/3

3. A. vi. Computation of Sample variation

Sampling variation is simply the variation in a statistic from sample to sample. It can be measured by comparing actual samples.

Calculation- (example)

=VAR (Starting Cell id: Ending Cell id) enter

= VAR (B2:B10) enter

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

s^2 = sample variance
x_i = value of i^{th} element
$\bar{x}$ = sample mean
n = sample size

Note: there is several type of Sample variation based on series, such as- one series, two series, more than Two Series.

In the case of two or three series, the methods will be-

Data tab – Click on Data Analysis – Click on Descriptive Statistics – Click on ok – Click on input range – Selects the data set/ input data range – Check the output range –Select a blank cell on work sheet – Check Summary Statistics box – click on ok.

3. A. vii. Computation of Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

Calculation- (example)

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

Hours of Study	Exam. Grade
6	1
3	2
5	1
4	3

Click on data tab – Select data analysis – Click on Correlation – Select Input Range – Select all the cell that has value under calculation – check labels if you select labels cell – Click on ok.

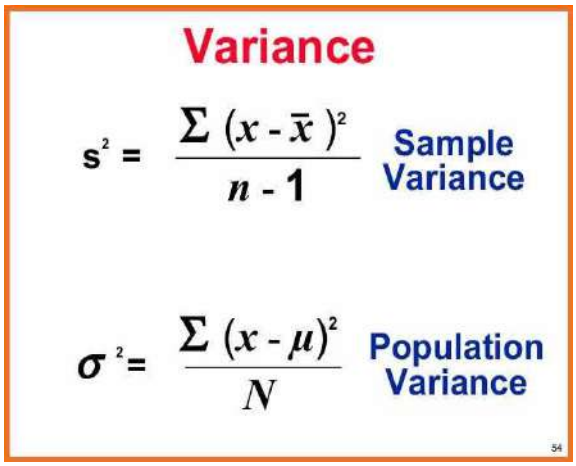
3. A. viii. Computation of Covariance & Correlation

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.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

Calculation- (example)

The 2 step for calculation of Co-variance (Cov) – Sample Variance (S^2) = VAR.S (B2:B11) enter Population Variance (σ^2) = VAR.P (B2:B11) enter



OR

Click on data tab – click on data analysis – select descriptive statistics – select input range – select the entire data – check the label box if you select the label cell – select output range – select a blank cell on worksheet – click on ok.

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

3. A. ix. Computation of Co-efficient of Correlation(r)

Calculation-

$r = S_{xy}/S_x.S_y$ Where, S_{xy} = Co-variance

S_x = Standard Deviation of X

S_y = Standard Deviation of Y

REGRESSION, CORRELATION, CURVE FITTING, MULTIVARIATE ANALYSIS

REGRESSION

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

Straight line: $\gamma i = \beta 0 + \beta 1 xi + \in i. \quad i = 1 \dots \dots n$

Input data – Insert Scatter – Click on layout tab – Select chart title – write chart title – select chart tools – select layout – select trend line – select display equation – select display r^2 (Co-efficient of determination).

Then, Click on data – data analysis – Regression – ok – Input all Y variable – Input all X variable – Check the labels – click on output location – select a blank cell – click on ok.

CORRELATION

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

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

Calculation- (example)

Hours of Study	Exam. Grade
6	1
3	2
5	1
4	3

Click on data tab – Select data analysis – Click on Correlation – Select Input Range – Select all the cell that has value under calculation – check labels if you select labels cell – Click on ok.

SPSS ENVIRONMENT

The full form of SPSS is Software Package for the Social Science. When you go to operate, you should have to mind that there are two sheets, one is Variable View sheet and another one is data view. At

first you have to put variable name, nature, length etc. in the variable view sheet. And next you can go to the data view sheet and put the value. Then you can perform you duty.

Requirements:

Desktop or Laptop with SPSS installed.

i. Computation of Rank

A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In mathematics, this is known as a weak order or total preorder of objects.

- **Calculation Procedure**

Transform – Rank Cases – Select Rank Cases – Check the Display Summary Statistics Box - Check assign rank 1 to (Smallest Value/Largest Value) – Click on Rank Type – Click on Continue – Click on Ties – Select the Rank assigned type – Click on continue – Click on ok.

ii. Computation of Mean& Standard Deviation

The statistical mean refers to the mean or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the mean or the average.

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

- **Calculation Procedure**

Click Analyse - Descriptive Statistics – Descriptive - Drag the variable of interest from the left into the Variables box on the right - Click Options, and select Mean and Standard Deviation. - Press Continue, and then press OK - Result will appear in the SPSS output viewer.

iii. Computation of Median

The median is a simple measure of central tendency. To find the median, we arrange the observations in order from smallest to largest value. If there are an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

- **Calculation Procedure**

Click Analyse - Descriptive Statistics - Frequencies - Move the variable for which you wish to calculate the median into the right hand column - Click the Statistics button - select Median under Central Tendency, and then press Continue - Click OK to perform the calculation.

iv. Computation of Mode

The mode of a set of data values is the value that appears most often. If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

- **Calculation Procedure**

Click Analyse - Descriptive Statistics – Descriptive - Drag the variable of interest from the left into the Variables box on the right - Click Options, and select Mode - Press Continue, and then press OK - Result will appear in the SPSS output viewer.

v. Computation of Moving Average

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set.

- **Calculation Procedure**

Transform – Rank Cases – Select Create time series – Select & Drag the variable upon moving average will calculate in Create Time Series Dialog box – Function – Cantered moving average – Put the order & Span in order & Span box – Click on ok

vi. Computation of Sample variation

Sampling variation is simply the variation in a statistic from sample to sample. It can be measured by comparing actual samples.

- **Calculation Procedure**

Click Analyse - Descriptive Statistics – Descriptive/Frequencies - Drag the variable of interest from the left into the Variables box on the right – Select/ Deselect Display frequencies table - Click Statistics, and select your interest - Press Continue, and then press OK - Result will appear in the SPSS output viewer.

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

s^2 = sample variance
x_i = value of i^{th} element
$\bar{x}$ = sample mean
n = sample size

vii. Computation of Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

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

- **Calculation Procedure**

Click Analyse - Correlation – Select Correlation type (Bivariate/partial etc.) - Drag the variable of interest from the left into the Variables box on the right(X, Y) – Select coefficient type – Select Test of Significance(One tailed/Two tailed) - Check flag Significance correlation – Click on ok - Result will appear in the SPSS output viewer.

viii. Computation of Covariance

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.

- **Calculation Procedure**

Click Analyse - Correlate –Bivariate - Drag the variable of interest from the left into the Variables box on the right(X, Y, Z) – Click on option – Check Mean and Standard deviation box –Check cross product deviations and covariance – Check exclude cases pairwise– Click on continue – Click on ok - Result will appear in the SPSS output viewer.

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

ix. Computation of Co-efficient of Correlation(r)

- **Calculation Procedure**

Click Analyze - Correlate - Bivariate - Select the variables Height and Weight and move them to the Variables box - In the Correlation Coefficients area - select Pearson - In the Test of Significance area - select your desired significance test - two-tailed or one-tailed - Click on ok - Result will appear in the SPSS output viewer.

- r = Sxy/Sx.Sy Where, Sxy = Co-variance
- Sx = Standard Deviation of X
- Sy = Standard Deviation of Y

---

4. Regression, Correlation, Curve Fitting

REGRESSION

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

Straight line: $\gamma i = \beta 0 + \beta 1 xi + \in i. \quad i = 1 \ldots \ldots n$

- **Calculation Procedure**

Click Analyse - Regression – Linear – Select &Drag the variable of interest from the left into the Variables box on the right(X, Y,) – Click on option – Enter use of probability F & removal – Check include constant in equation – Check exclude cases list wise – Click on continue – Click on plot – Dependent - Click on continue – Statistics – Check model fit – Check estimate - Click on continue - Click on ok - Result will appear in the SPSS output viewer.

---

5. INTERNET SURFING, DATA EXTRACTION& POWER POINT PRESENTATION

---

5. A. INTERNET SURFING

On the World Wide Web, surfing means to move from one Web page to another, usually in an undirected manner. When surfing, the user typically visits pages based on what interests him/her at the moment. ... Surfing is a favourite pastime for millions of people around the world who have access to the Internet.

5. B. DATA MANIPULATION

Data manipulation is the process of changing data to make it easier to read or be more organized. For example, a log of data could be organized in alphabetical order, making individual entries easier to locate. Data manipulation is often used on web server logs to allow a website owner to view their most popular pages as well as their traffic sources.

Users in the accounting field or similar fields often manipulate data to figure out product costs, sales trends, or potential tax obligations. Stock market analysts are frequently using data manipulation to predict trends in the stock market and how stocks might perform in the near future.

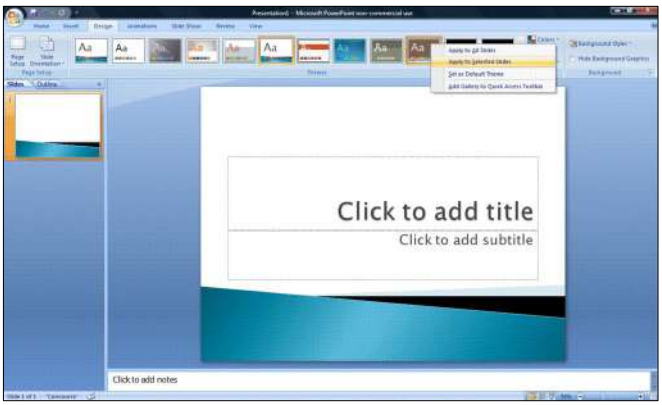
5. C. POWERPOINT PRESENTATION

A PowerPoint presentation is a presentation created using Microsoft PowerPoint software. The presentation is a collection of individual slides that contain information on atopic. PowerPoint presentations are commonly used in business meetings and for training and educational purposes.

PROCEDURE:

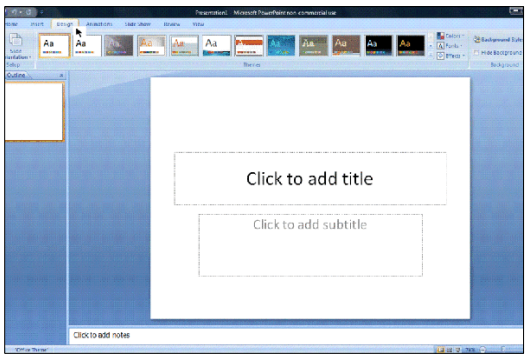
Step 1: Launch the PowerPoint Program

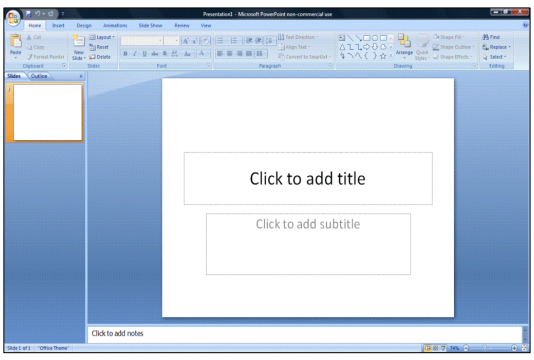
When you launch the PowerPoint program, you may be prompted to pick what kind of document you want to create. Choose to create a blank presentation. If it does not ask you this, a blank presentation will automatically launch.



Step 2: Choosing a Design

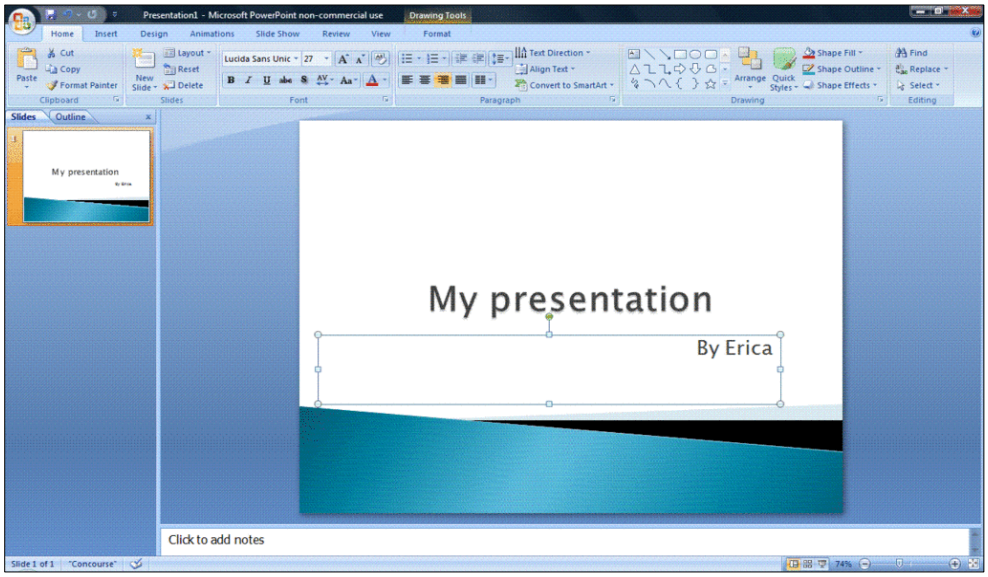
The next thing you want to do is decide what design you want for the presentation. To do this, go to the 'Design' tab at the top of the page. Scroll through all the options and decide which one looks best for the presentation you want. To get a preview of what the design will look like before applying it to the presentation, hover over the design you want to preview. This design will be automatically continued throughout the rest of your presentation. Once you have more than one slide, you can add a different design for just one slide. To do this, select the slide you want to change the design on by clicking on it. It will pop-up as the big slide in the screen. Then you can right-click the design you want for this slide and select 'Apply to Selected Slide'. It will appear on that slide, but will not change the design of the other slides.





Step 3: Create Title Page

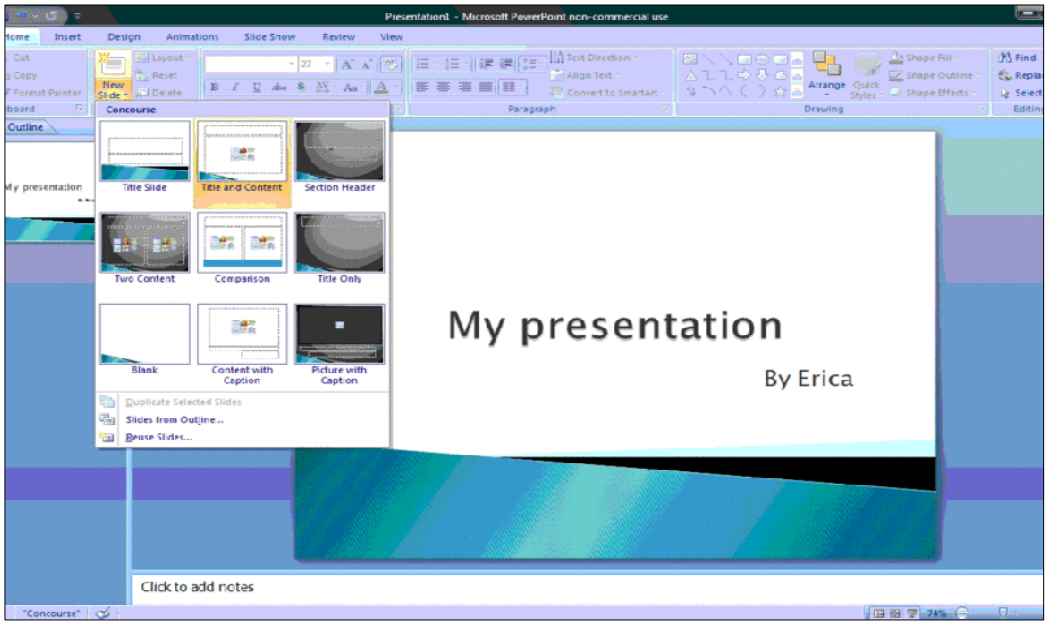
Click the first box that says 'Click to add title' and add the title of your presentation. Click the bottom box to add your name, or any other subtitle that you choose. Once you have your text in the boxes, you can change their font, size, color, etc. with the toolbar options at the top. You can change the size of the text box by selecting it, and then dragging the corners of the box. To move the text boxes, select the box, and move your arrow over the border of the box. A four-arrow icon will appear, and clicking with this icon will allow you to move the text boxes wherever you choose.



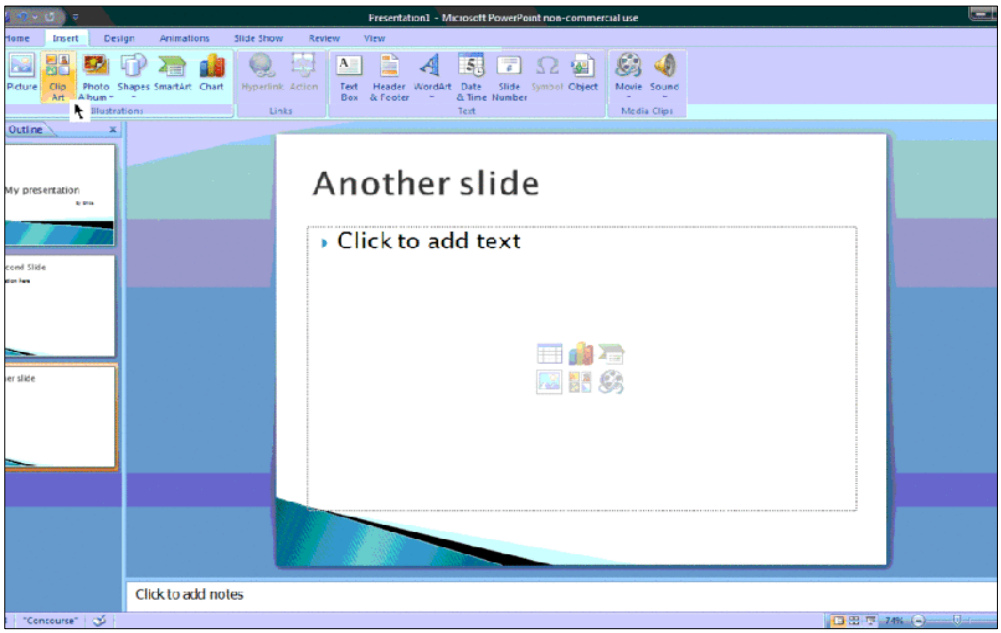
Step 4: Add More Slides

Chances are, you are going to need more than one slide. There are a few ways you can add more slides. Notice that there is a separate area to the left of the screen where your first slide is located. The first way to add a slide is to right-click the area under where your first slide is located and selects 'New Slide'. A new slide will appear. The second way to add another slide it to click 'New Slide' in the toolbar above the slides. This button is divided into two parts,. The top will insert a new slide with a default layout. You can also click the bottom half of this button, which will allow you to choose what type of layout you want. You can choose a slide with two text-boxes and a title, one text-box, only a title, and many other options. You will see your new slide appear to the left under the first, as well become the large slide that you can edit. The design you picked earlier will have carried over to this slide.

The design will carry over for the rest of the slides you create unless you decide to change just one, like described earlier. The guideline layout you chose will appear, and you can then add in your information.



Step 5: Add Charts, Pictures, and Graphs etc.

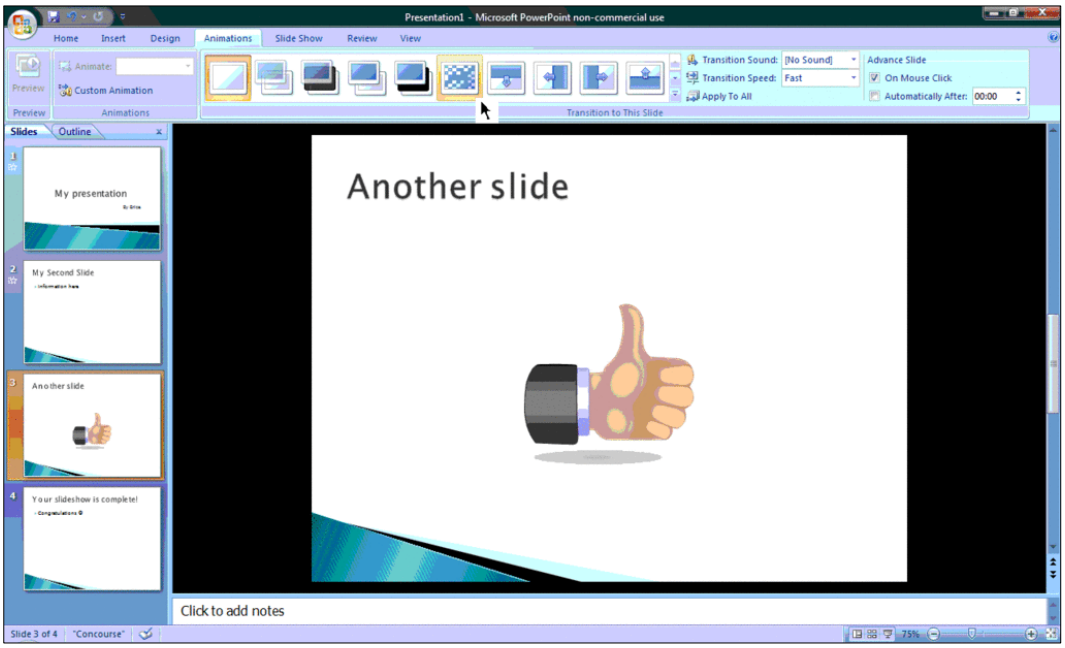


If you want to insert a chart, picture, graph, or any other graphic, click on the 'Insert' tab at the top of the window. Here you will see buttons of all the options of what you can insert into your slide.

Click the designated box and insert what it is you want to have on that slide. A second way you can insert pictures and graphs is when you have an empty text or image box. Little pictures of the same options you saw in the toolbox will show up in the middle of the box, and you can click any of these to insert as well. Once you have your chart or picture, you can add a border or edit it however you want in the 'Format' tab.

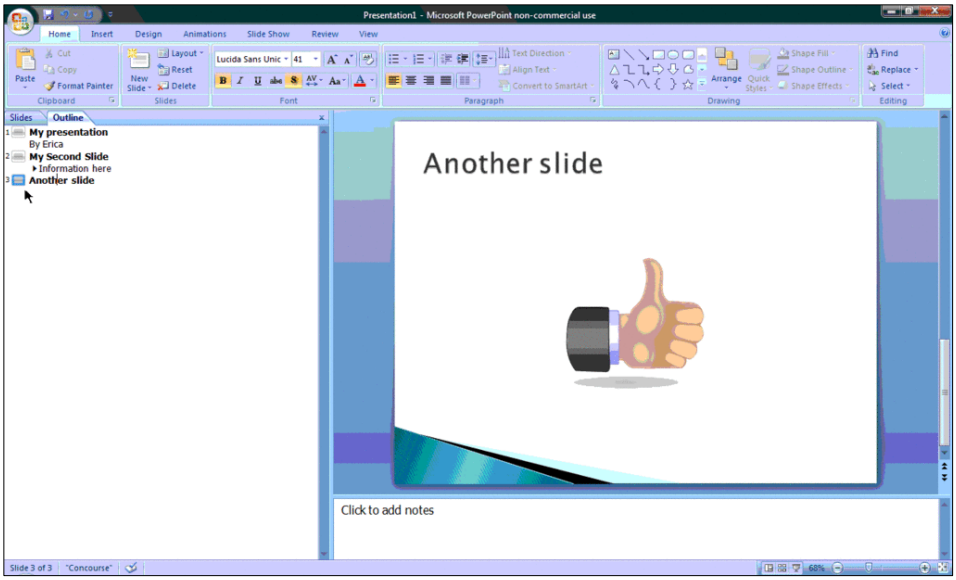
Step 6: Add Transitions

To add transitions in between your slides, click the 'Animations' tab at the top of the page. Here you can scroll through all the options of transitions, and hover over them to see a preview. Select the slide you want the transition applied to, and then click the transition you chose. You can do this for every slide, selecting the same or different transitions.



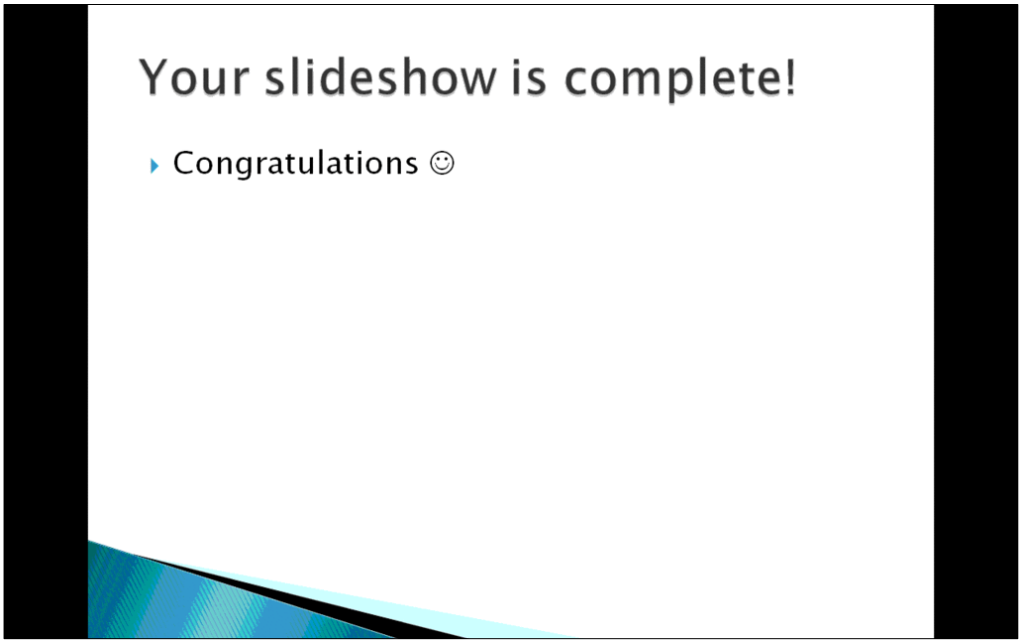
Step 7: Changing the Order

Once you have all your slides made, you can change the order of the slides. To do this, click and drag the slides from where they are to where you want them in the order. Another possibility, which is particularly useful if you presentation is longer, is to click the 'Outline' button. You can find this small button above the left area where all your slides are located smaller, directly to the right of the 'Slides' button. Here you will see a list of all your slides and you can click and drag your slides to where you want them.



Step 8: Play the Presentation

Once you have all your slides completed and in the order you want, view your slideshow. Click the 'Slide Show' tab at the top of the page and select 'From Beginning'. You can go through your entire slideshow, and change slides by clicking or pressing the right arrow.



****

DISCLAIMER

This self-learning material is based on different books, journals and web sources.