KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University Established Under Section 3 of UGC Act 1956)
Coimbatore – 641 021.

DEPARTMENT OF Computer Applications

18CAP304 **STATISCAL COMPUTING**

SEMESTER-III

L	T	P	C
4	0	0	4

SCOPE

On successful completion of this course the learner gains knowledge about the concept of statistical measures, relationship between variables, hypothesis testing, theory of estimation, experimental designs and quality control charts and its applications with exact statistical treatment.

COURSE OBJECTIVE

To make the learners to understand the concepts of statistical measures, relationship between variables, hypothesis testing, theory of estimation, experimental designs and quality control charts with computational applications.

COURSE OUTCOMES

On successful completion of this course the students will able to

- Exhibit knowledge of statistical terms.
- Differentiate between the two branches of statistics.
- Identify the position of a data value in a data set using various measures of position such as percentiles, deciles and quartiles.
- Describe data using the measures of variation such as range, variance, and standard deviation.
- Find the relationship between variable using correlation and regression
- Determine inferential statistics such as Hypothesis testing and estimation
- Design of models for Experiments such as CRD, RBD and LSD
- To familiar with the quality control concepts and know how to draw different types of Quality Control Charts such as $\overline{X}$, R charts.p and np charts

UNIT I

Statistical Measures: Introduction to descriptive Statistics-basic definitions, frequency distribution, Measure of Central Tendency: Mean Median and Mode. Measure of Dispersion: Absolute and relative measure of dispersion: Range, Mean deviation, Quartile deviation, Standard deviation and corresponding relative measures.

UNIT II

Correlation and Regression: Types of Correlation – Simple and Multiple, Positive and Negative, Linear and Non-Linear, Partial and Total. Methods of calculating correlation coefficient: Scatter diagram, Karl Pearson and Spearmen (Rank) correlation coefficient, Regression: Types, lines and equations, Linear Regression - least square method of solving regression equations, X on Y and Y on X.

UNIT III

Testing of Hypothesis: Introduction to Inferential Statistics: Null and alternative hypothesis, Type I and Type II errors, Standard error, level of significance, acceptance and rejection regions and procedure for testing hypothesis. Large sample test - Z test - tests for means, variances and proportions, Small sample tests based on t, F and Chi- square distributions.

UNIT IV

Estimation and Design of Experiment: Point Estimation - characteristics of estimation - interval estimation - interval estimates of mean, standard deviation and\proportion. Design of Experiments: Completely Randomized Design (CRD), Randomized Block Design (RBD) and Latin Square design (LSD) Models

UNIT V

Statistical Quality Control (SQC): Statistical basis for control charts, control limits. Control charts for variables - $\overline{X}$, R charts. Charts for defectives – p and np charts. Chart for defects – C chart. Acceptance Sampling – single and double sampling plans.

SUGGESTED READINGS

1. T.Veerarajan, ”Fundamentals of Mathematical Statistics”, Yesdee Publishing Pvt Ltd,2017.
2. R.S.N.Pillai, Bagavathy, “Statistics”, S. Chand & Company Ltd, New Delhi, 2002.
3. T N Srivastava and ShailajaRego., 2012, 2e, Statistics for Management, McGraw Hill
4. Education, New Delhi.
5. Steven K Thompson., 2012, Sampling, John wiley and sons inc.
6. Montgomery Douglas C., 2008. Introduction to Quality Control, Sixth Edition, John Wiley and Sons.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF MATHEMATICS

Staff name: M.Sangeetha
Subject Name: Statistical Computing **Subject Code:18CAP304**
Semester:III **Class: II MCA**

S.No	Lecture Duration Period	Topics to be Covered	Support Material/Page Nos
		UNIT-I	
1.	1	Statistics - Meaning, definition, Basic concept of statistics and its applications	R4: Chap: 2,Pg.No: 13-21
2.	1	Measure of central tendencies	R4: Chap: 4,Pg.No: 124-137
3.	1	Problems on Measure of central tendencies	R4: Chap: 5,Pg.No: 237-241
4.	1	Measure of dispersion: Mean Median and Mode	R5: Chap: 4,Pg.No:224-232
5.	1	Measure of Dispersion and Absolute and relative measure of dispersion	R1: Chap: 4,Pg.No:4.45-4.47 R6: Chap: 5,Pg.No:5.26-5.30
6.	1	Range, Mean deviation and quartile deviation	R4 : chap :13, Pg No : 479-501
7.	1	Standard deviation and Problems on Standard deviation and corresponding measures	R4 : chap :13, Pg No : 479-508
8.	1	Recapitulation and Discussion on important questions from previous year ESE question paper	
		Total Hours	**8 Hours**
		UNIT-II	
1.	1	Types of Correlation	R4 : chap :20, Pg No : 810-814

2.	1	Simple and Multiple, Positive and Negative,	R5 : chap :8, Pg No : 269-272
3.	1	Linear and Non-Linear, Partial and Total methods of calculating correlation coefficient	R5 : chap :8, Pg No : 273-280
4.	1	Scatter diagram	R5 : chap :8, Pg No : 282-285
5.	1	Karl Pearson and Spearmen (Rank) correlation coefficient	R4 : chap :20, Pg No : 825-830
6.	1	Problems on Karl Pearson correlation coefficient	R4 : chap :20, Pg No : 832-842
7.	1	Problems on Spearmen (Rank) correlation coefficient	R4 : chap :20, Pg No : 843-845
8.	1	Recapitulation and Discussion on important questions from previous year ESE question paper	
		Total Hours	**8 Hours**
UNIT-III			
1.	1	Introduction to Inferential Statistics: Null and alternative hypothesis,	R1: Chap: 17 : P: NO :721-723
2.	1	Problems on Null hypothesis and alternative hypothesis	R1: Chap: 13: P: NO :566 R1:chap:17:P.No:722-723
3.	1	Type I and Type II errors, Standard error and problems on testing hypothesis	R1: Chap: 17 : P: NO :635-637 R1: Chap: 14: P: NO :638
4.	1	level of significance, Problems on level of significance	R1: Chap:14: P: NO :601-606 R1: Chap: 15: P: NO :635
5.	1	acceptance and rejection regions	R1: Chap: 16:P:NO:686-690
6.	1	procedure for testing hypothesis	R1: Chap: P:NO:691-696
7.	1	Problems on testing hypothesis and large sample test	R1: Chap: 16:P:NO:686-696
8.	1	Z test - tests for means	R1: Chap: P:NO:697-699
9.	1	F and Chi- square distributions	R1: Chap: 16:P:NO:699-702
10.	1	Recapitulation and Discussion on important questions from previous year ESE question paper	
		Total Hours	**10 Hours**
UNIT-IV			

1.	1	Concept of Point Estimation	R5:chap 16 , Pg No : 550
2.	1	Problems on Point Estimation	R5:chap 16 , Pg No : 551-553
3.	1	characteristics of estimation	R5:chap 16 , Pg No : 554-556
4.	1	Interval estimation of mean	R5:chap 16 , Pg No : 558-563
5.	1	standard deviation and proportion	R5:chap 16 , Pg No : 564-569
6.	1	Design of Experiments: Completely Randomized Design (CRD)	R5:chap 18 , Pg No : 643-648
7.	1	Randomized Block Design (RBD) and Concept of Randomized Block Design (RBD)	R5:chap 18 , Pg No : 649-655
8.	1	Comparison between CRD and RBD	R5:chap 18 , Pg No :656-660
9.	1	Latin Square design (LSD) Models	R5:chap 18 , Pg No :661-668
10.	1	Concept of Latin Square design	R5:chap 18 , Pg No :669-676
11.	1	Comparison between CRD ,RBD and LSD	R5:chap 16 , Pg No : 551-553
12.	1	Recapitulation and Discussion on important questions from previous year ESE question paper	
	Total Hours		**12 Hours**

UNIT-V			
1.	1	Basic concept of SQC methods	R5:chap 18 , Pg No : 643-648
2.	1	Statistical basis for control charts and Some examples SQC methods	R5:chap 18 , Pg No : 648-649
3.	1	control limits for variables	R5:chap 18 , Pg No : 650-652
4.	1	$\overline{X}$, R charts and charts for defectives	R5:chap 18 , Pg No : 653-658
5.	1	p , np charts , Tests of Significance F test and Application	R5:chap 18 , Pg No :659-663
6.	1	Chart for defects ,C chart and Acceptance sampling	R5:chap 18 , Pg No :664-668
7.	1	Single sampling plans and double sampling plans	R5:chap 18 , Pg No :669-676

8.	1	Discussion on Previous year Question Papers	
9.	1	Discussion on Previous year Question Papers	
10.	1	Recapitulation and Discussion on important questions from previous year ESE question paper	
		Total Hours	**10 Hours**
		Total Planned Hours	**48 Hours**

SUGGESTED READINGS

1 Veerarajan, "Fundamentals of Mathematical Statistics", Yesdee Publishing Pvt Ltd,2017.
2 R.S.N.Pillai, Bagavathy, "Statistics", S. Chand & Company Ltd, New Delhi, 2002.
3 T N Srivastava and Shailaja Rego., 2012, 2e, Statistics for Management, Mc Graw Hill Education, New Delhi.
4 Steven K Thompson., 2012, Sampling, John wiley and sons inc.
5 Montgomery Douglas C., 2008. Introduction to Quality Control, Sixth Edition, John Wiley and Sons.

UNIT – I

SYLLABUS

Statistical Measures: Introduction to descriptive Statistics-basic definitions, frequency distribution, Measure of Central Tendency: Mean Median and Mode. Measure of Dispersion: Absolute and relative measure of dispersion: Range, Mean deviation, Quartile deviation, Standard deviation and corresponding relative measures.

Measures of Central Tendency:

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Characteristics for a good or an ideal average :

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing.

Arithmetic mean or mean :

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values $x_1, x_2 \ldots x_n$ then the mean, $\bar{x}$, is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

This formula is for the ungrouped or raw data.

Example 1 :

Calculate the mean for 2, 4, 6, 8, 10

Solution:

$$\bar{x} = \frac{2+4+6+8+10}{5}$$

$$= \frac{30}{5} = 6$$

Short-Cut method :

Under this method an assumed or an arbitrary average (indicated by A) is used as the basis of calculation of deviations from individual values. The formula is

$$\bar{x} = A + \frac{\sum d}{n}$$

where, A = the assumed mean or any value in x

d = the deviation of each value from the assumed mean

Example 2 :

A student' s marks in 5 subjects are 75, 68, 80, 92, 56. Find his average mark.

Solution:

	X	d=x-A
	75	7
A	68	0
	80	12
	92	24
	56	-12
Total		31

$$\bar{x} = A + \frac{\sum d}{n}$$

$$= 68 + \frac{31}{5}$$

$$= 68 + 6.2$$

$$= 74.2$$

Grouped Data :

The mean for grouped data is obtained from the following formula:

$$\bar{x} = \frac{\sum fx}{N}$$

where x = the mid-point of individual class

f = the frequency of individual class

N = the sum of the frequencies or total frequencies.

Short-cut method :

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

where $d = \frac{x - A}{c}$

A = any value in x

N = total frequency

c = width of the class interval

Example 3:

Given the following frequency distribution, calculate the arithmetic mean

Marks	64	63	62	61	60	59
Number of Students	8	18	12	9	7	6

Solution:

X	F	fx	d=x-A	fd
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	–1	–9
60	7	420	–2	–14
59	6	354	–3	–18
	60	3713		- 7

Direct method

$$\bar{x} = \frac{\sum fx}{N} = \frac{3713}{60} = 61.88$$

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{N} = 62 - \frac{7}{60} = 61.88$$

Example 4 :

Following is the distribution of persons according to different income groups. Calculate arithmetic mean.

Income Rs(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of persons	6	8	10	12	7	4	3

Solution:

Income C.I	Number of Persons (f)	Mid X	$d = \frac{x - A}{c}$	Fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	A 35	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
	50			-20

$$\text{Mean} = \bar{x} = A + \frac{\sum fd}{N}$$

$$= 35 - \frac{20}{50} \times 10$$

$$= 35 - 4$$

$$= 31$$

Merits and demerits of Arithmetic mean :

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Median :

The median is that value of the variate which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median.

Ungrouped or Raw data :

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle value .If the number of values are even, median is the mean of middle two values.

By formula

$$\text{Median} = \text{Md} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{item.}$$

Example 11:

When odd number of values are given. Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

The middle value is the 5^{th} item i.e., 25 is the median

Using formula

$$\text{Md} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{item.}$$

$$= \left(\frac{9+1}{2}\right)^{\text{th}} \text{item.}$$

$$= \left(\frac{10}{2}\right)^{\text{th}} \text{item}$$

$$= 5^{\text{th}} \text{ item}$$

$$= 25$$

Example 12 :

When even number of values are given. Find median for the following data

5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (ie) mean of (10,12) ie

$$= \left(\frac{10+12}{2}\right) = 11$$

$$\therefore \text{median} = 11.$$

Using the formula

$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{item.}$$

$$= \left(\frac{8+1}{2}\right)^{th} \text{item.}$$

$$= \left(\frac{9}{2}\right)^{th} \text{item} = 4.5^{th} \text{item}$$

$$= 4^{th} \text{item} + \left(\frac{1}{2}\right)(5^{th} \text{item} - 4^{th} \text{item})$$

$$= 10 + \left(\frac{1}{2}\right)[12\text{-}10]$$

$$= 10 + \left(\frac{1}{2}\right) \times 2$$

$$= 10 + 1$$

$$= 11$$

Example 13:

The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.

Serial No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

Indicate in which subject is the level of knowledge higher ?

Solution:

For such question, median is the most suitable measure of central tendency. The mark in the two subjects are first arranged in increasing order as follows:

Serial No	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	28	30	32	35	46	47	52	53	55	60
Marks in Accountancy	24	25	31	32	43	45	57	72	80	84

$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{item} = \left(\frac{10+1}{2}\right)^{\text{th}} \text{item} = 5.5^{\text{th}} \text{ item}$$

$$= \frac{Value\ of\ 5^{th}\ item + value\ of\ 6^{th}\ item}{2}$$

$$\text{Md (Statistics)} = \frac{46+47}{2} = 46.5$$

$$\text{Md (Accountancy)} = \frac{43+45}{2} = 44$$

There fore the level of knowledge in Statistics is higher than that in Accountancy.

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution , cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency : (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series:

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N+1}{2}\right)$

Step3: See in the cumulative frequencies the value just greater than $\left(\frac{N+1}{2}\right)$

Step4: Then the corresponding value of x is median.

Example 14:

The following data pertaining to the number of members in a family. Find median size of the family.

Number of members **x**	1	2	3	4	5	6	7	8	9	10	11	12
Frequency F	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	f	cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

Median = size of $\left(\frac{N+1}{2}\right)^{th}$ item

$$= \text{size of} \left(\frac{60+1}{2} \right)^{\text{th}} \text{item}$$

$$= 30.5^{\text{th}} \text{ item}$$

The cumulative frequencies just greater than 30.5 is 38.and the value of x corresponding to 38 is 6.Hence the median size is 6 members per family.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N}{2} \right)$

Step3: See in the cumulative frequency the value first greater than $\left(\frac{N}{2} \right)$, Then the corresponding class interval is called the Median class. Then apply the formula

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where l = Lower limit of the median class

m = cumulative frequency preceding the median

c = width of the median class

f =frequency in the median class.

N=Total frequency.

Example 15:

The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.

Income group (in Rs)	Number of workers
Below 100	1
100-150	20
150-200	42
200-250	55
250-300	62
300-350	45
350-400	30
400-450	25
450-500	15
500-550	18
550-600	10
600 and above	2
	325

Calculate median income

Solution:

Income group (Class-interval)	Number of workers (Frequency)	Cumulative frequency c.f
Below 100	1	1
100-150	20	21
150-200	42	63
200-250	55	118
250-300	62	180
300-350	45	225
350-400	30	255
400-450	25	280
450-500	15	295
500-550	18	313
550-600	10	323
600 and above	2	325
	325	

$$\frac{N}{2}=\frac{325}{2}=162.5$$

Here $l = 250$, N = 325, f = 62, c = 50, m = 118

$$\text{Md} = 250+\left(\frac{162.5-118}{62}\right)\times 50$$

$$= 250+35.89$$

$$= 285.89$$

Example 16:

Following are the daily wages of workers in a textile. Find the median.

Wages (in Rs.)	Number of workers
less than 100	5
less than 200	12
less than 300	20
less than 400	32
less than 500	40
less than 600	45
less than 700	52
less than 800	60
less than 900	68
less than 1000	75

Solution :

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 100, hence the width of the class interval equal to 100.

Class interval	f	c.f
0-100	5	5
100-200	7	12
200-300	8	20
300- 400	12	32
400-500	8	40
500-600	5	45
600-700	7	52
700-800	8	60
800-900	8	68
900-1000	7	75
	75	

$$\left(\frac{N}{2}\right) = \left(\frac{75}{2}\right) = 37.5$$

$$\text{Md} = l + \left(\frac{\frac{N}{2} - m}{f}\right) \times \text{c}$$

$$= 400 + \left(\frac{37.5 - 32}{8}\right) \times 100 = 400 + 68.75 = 468.75$$

Merits of Median :

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median :

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations.

Mode :

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it.

Computation of the mode:

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 29:

2 , 7, 10, 15, 10, 17, 8, 10, 2

$\therefore$ Mode = $M_0 = 10$

In some cases the mode may be absent while in some cases there may be more than one mode.

Example 30:

1. 12, 10, 15, 24, 30 (no mode)
2. 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

$\therefore$ the modes are 7 and 10

Grouped Data:

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution :

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula.

$$\text{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

l = Lower limit of the model class

$$\Delta_1 = f_1 - f_0$$
$$\Delta_2 = f_1 - f_2$$

f_1 = frequency of the modal class
f_0 = frequency of the class preceding the modal class
f_2 = frequency of the class succeeding the modal class

The above formula can also be written as

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Example 31:

Calculate mode for the following :

C- I	f
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

Solution:

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

Here $l=200, f_1=150, f_0=91, f_2=87, C=50$

$$\text{Mode} = M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 200 + \frac{150\text{-}91}{2\times150-91-87}\times 50$$

$$= 200 + \frac{2950}{122}$$

$$= 200 + 24.18 = 224.18$$

Determination of Modal class :

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases

i. If the maximum frequency is repeated
ii. If the maximum frequency occurs in the beginning or at the end of the distribution
iii. If there are irregularities in the distribution, the modal class is determined by the method of grouping.

Steps for Calculation :

We prepare a grouping table with 6 columns

1. In column I, we write down the given frequencies.
2. Column II is obtained by combining the frequencies two by two.
3. Leave the 1^{st} frequency and combine the remaining frequencies two by two and write in column III
4. Column IV is obtained by combining the frequencies three by three.
5. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V
6. Leave the 1st and 2^{nd} frequencies and combine the remaining frequencies three by three and write in column VI

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class use the formula to calculate the modal value.

Example 32:

Calculate mode for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	9	12	15	16	17	15	10	13

Grouping Table

C I	f	2	3	4	5	6
0- 5	9					
5-10	12	21				
10-15	15		27	36		
15-20	16	31			**43**	
20-25	**17**		**33**			**48**
25-30	15	**32**		**48**		
30-35	10		25		42	38
35-40	13	23				

Analysis Table

Columns	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
1					1			
2					1	1		
3				1	1			
4				1	1	1		
5		1	1	1				
6			1	1	1			
Total		1	2	4	5	2		

The maximum occurred corresponding to 20-25, and hence it is the modal class.

$$\text{Mode} = \text{Mo} = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Here $l = 20$; $\Delta_1 = f_1 - f_0 = 17 - 16 = 1$

$\Delta_2 = f_1 - f_2 = 17 - 15 = 2$

$$\therefore M_0 = 20 + \frac{1}{1+2} \times 5$$

$$= 20 + 1.67 = 21.67$$

MEASURES OF DISPERSION

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties

1.It should be rigidly defined

2. It should be based on all the items.

3. It should not be unduly affected by extreme items.

4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate

Absolute and Relative Measures :

There are two kinds of measures of dispersion, namely
1.Absolute measure of dispersion
2.Relative measure of dispersion.

The various absolute and relative measures of dispersion are listed below.

Absolute measure	Relative measure
1. Range	1.Co-efficient of Range
2.Quartile deviation	2.Co-efficient of Quartile deviation
3.Mean deviation	3. Co-efficient of Mean deviation
4.Standard deviation	4.Co-efficient of variation

7.3 Range and coefficient of Range:

7.3.1 Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range $= L - S$.
Where L = Largest value.
S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class
S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.
S = Mid value of the lowest class.

7.3.2 Co-efficient of Range :

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$

Example1:

Find the value of range and its co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

Solution:

L=11, S = 4.

Range $= L - S = 11 - 4 = 7$

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$

$$= \frac{11-4}{11+4}$$

$$= \frac{7}{15} = 0.4667$$

Example 2:

Calculate range and its co efficient from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution:

L = Upper boundary of the highest class.
= 75

S = Lower boundary of the lowest class.
= 60

Range = L – S = 75 – 60 = 15

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$

$$= \frac{75-60}{75+60}$$

$$= \frac{15}{135} = 0.1111$$

7.3.3 Merits and Demerits of Range :

Merits:

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, et c., range is most widely used.

Demerits:

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

7.6 Standard Deviation and Coefficient of variation:

7.6.1 Standard Deviation :

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

The standard deviation is denoted by the Greek letter σ (sigma)

7.6.2 Calculation of Standard deviation-Individual Series :

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean
b) Deviation taken from Assumed mean

a) Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series ($\bar{x}$)
2. Find out the deviation of each value from the mean ($x = X - \overline{X}$)
3. Square the deviations and take the total of squared deviations $\sum x^2$

4. Divide the total ($\sum x^2$) by the number of observation $\left(\frac{\sum x^2}{n}\right)$

The square root of $\left(\frac{\sum x^2}{n}\right)$ is standard deviation.

Thus $\sigma = \sqrt{\left(\frac{\sum x^2}{n}\right)}$ or $\sqrt{\frac{\sum (x-\bar{x})^2}{n}}$

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value.

Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, We apply short –cut method; deviations are taken from an assumed mean. The formula is:

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

Where d-stands for the deviation from assumed mean = (X-A)

Steps:

1. Assume any one of the item in the series as an average (A)
2. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the deviations $\sum d$
3. Square the deviations; i.e., d^2 and add up the squares of deviations, i.e, $\sum d^2$
4. Then substitute the values in the following formula:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

Note: We can also use the simplified formula for standard deviation.

$$\sigma = \frac{1}{n}\sqrt{n\sum d^2 - (\sum d)^2}$$

For the frequency distribution

$$\sigma = \frac{c}{N}\sqrt{N\sum fd^2 - (\sum fd)^2}$$

Example 9:

Calculate the standard deviation from the following data.

14, 22, 9, 15, 20, 17, 12, 11

Solution:

Deviations from actual mean.

Values (X)	$X - \overline{X}$	$(X - \overline{X})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
120		140

$$\overline{X} = \frac{120}{8} = 15$$

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{140}{8}}$$

$$= \sqrt{17.5} = 4.18$$

Example 10:

The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

Student Nos :	1	2	3	4	5	6	7	8	9	10
Marks :	43	48	65	57	31	60	37	48	78	59

Solution: (Deviations from assumed mean)

Nos.	Marks (x)	d=X-A (A=57)	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
n = 10		$\sum d=-44$	$\sum d^2=1952$

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{1952}{10} - \left(\frac{-44}{10}\right)^2}$$

$$= \sqrt{195.2 - 19.36}$$

$$= \sqrt{175.84} = 13.26$$

7.6.3 Calculation of standard deviation:

Discrete Series:

There are three methods for calculating standard deviation in discrete series:

(a) Actual mean methods
(b) Assumed mean method
(c) Step-deviation method.

(a) Actual mean method:

Steps:

1. Calculate the mean of the series.
2. Find deviations for various items from the means i.e.,
   $x - \bar{x} = d$.
3. Square the deviations ($= d^2$) and multiply by the respective frequencies(f) we get fd^2
4. Total to product ($\sum fd^2$) Then apply the formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}}$$

(b) Assumed mean method:

Here deviation are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

Steps:

1. Assume any one of the items in the series as an assumed mean and denoted by A.
2. Find out the deviations from assumed mean, i.e, X-A and denote it by d.
3. Multiply these deviations by the respective frequencies and get the $\sum fd$
4. Square the deviations (d^2).
5. Multiply the squared deviations ($d^{2)}$ by the respective frequencies (f) and get $\sum fd^{2.}$
6. Substitute the values in the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

Where $d = X - A$, $N = \sum f$.

Example 11:

Calculate Standard deviation from the following data.

X :	20	22	25	31	35	40	42	45
f :	5	12	15	20	25	14	10	6

Solution:

Deviations from assumed mean

x	f	d = x –A (A = 31)	d^2	fd	fd^2
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210
45	6	14	196	84	1176
	N=107			$\sum fd=167$	$\sum fd^2$ =6037

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$= \sqrt{\frac{6037}{107} - \left(\frac{167}{107}\right)^2}$$

$$= \sqrt{56.42 - 2.44}$$

$$= \sqrt{53.98} = 7.35$$

7.6.4 Calculation of Standard Deviation –Continuous series:

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

The formula is,

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$d = \frac{m - A}{C}$$, C- Class interval.

Steps:

1. Find out the mid-value of each class.
2. Assume the center value as an assumed mean and denote it by A
3. Find out $d = \frac{m - A}{C}$
4. Multiply the deviations d by the respective frequencies and get Σfd
5. Square the deviations and get d^2
6. Multiply the squared deviations (d^2) by the respective frequencies and get Σfd^2
7. Substituting the values in the following formula to get the standard deviation

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C$$

Example 13:

The daily temperature recorded in a city in Russia in a year is given below.

Temperature C^0	No. of days
-40 to –30	10
-30 to –20	18
-20 to –10	30
-10 to 0	42
0 to 10	65
10 to 20	180
20 to 30	20
	365

Calculate Standard Deviation.

Solution:

Temperature	Mid value (m)	No. of days f	$d = \frac{m-(-5^n)}{10^n}$	fd	fd^2
-40 to -30	-35	10	-3	-30	90
-30 to -20	-25	18	-2	-36	72
-20 to -10	-15	30	-1	-30	30
-10 to - 0	-5	42	0	0	0
0 to 10	5	65	1	65	65
10 to 20	15	180	2	360	720
20 to 30	25	20	3	60	180
		N=365		Σfd = 389	Σfd^2 =1157

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{1157}{365} - \left(\frac{389}{365}\right)^2} \times 10$$

$$= \sqrt{3.1699 - 1.1358} \times 10$$

$$= \sqrt{2.0341} \times 10$$

$$= 1.4262 \times 10$$

$$= 14.26^\circ c$$

7.6.6 Merits and Demerits of Standard Deviation:

Merits:

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits:

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

7.6.7 Coefficient of Variation :

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiply it by 100. symbolically,

$$\text{Coefficient of variation (C.V)} = \frac{\sigma}{\bar{X}} \times 100$$

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Example 15:

In two factories A and B located in the same industrial area, the average weekly wages (in rupees) and the standard deviations are as follows:

Factory	Average	Standard Deviation	No. of workers
A	34.5	5	476
B	28.5	4.5	524

1. Which factory A or B pays out a larger amount as weekly wages?
2. Which factory A or B has greater variability in individual wages?

Solution:

Given $N_1 = 476$, $\overline{X}_1 = 34.5$, $\sigma_1 = 5$

$N_2 = 524,\ \overline{X}_2 = 28.5,\ \sigma_2 = 4.5$

1. Total wages paid by factory A

$= 34.5 \times 476$

$=$ Rs.16.422

Total wages paid by factory B

$= 28.5 \times 524$

$=$ Rs.14,934.

Therefore factory A pays out larger amount as weekly wages.

2. C.V. of distribution of weekly wages of factory A and B are

$$C.V.(A) = \frac{\sigma_1}{\bar{X}_1} \times 100$$

$$= \frac{5}{34.5} \times 100$$

$$= 14.49$$

$$C.V\ (B) = \frac{\sigma_2}{\bar{X}_2} \times 100$$

$$= \frac{4.5}{28.5} \times 100$$

$$= 15.79$$

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A

Example 16:

Prices of a particular commodity in five years in two cities are given below:

Price in city A	Price in city B
20	10
22	20
19	18
23	12
16	15

Which city has more stable prices?

Solution:

Actual mean method

City A			City B		
Prices (X)	Deviations from $\bar{X}$=20 dx	dx^2	Prices (Y)	Deviations from $\bar{Y}$ =15 dy	dy^2
20	0	0	10	-5	25
22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
Σx=100	Σdx=0	Σdx^2=30	Σy=75	Σdy=0	Σdy^2 =68

City A: $\overline{X} = \frac{\Sigma x}{n} = \frac{100}{5} = 20$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n}} = \sqrt{\frac{\Sigma dx^2}{n}}$$

$$= \sqrt{\frac{30}{5}} = \sqrt{6} = 2.45$$

$$C.V(x) = \frac{\sigma_x}{\overline{x}} \times 100$$

$$= \frac{2.45}{20} \times 100$$

$$= 12.25\ \%$$

City B: $\overline{Y} = \frac{\Sigma y}{n} = \frac{75}{5} = 15$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\overline{y})^2}{n}} = \sqrt{\frac{\Sigma dy^2}{n}}$$

$$= \sqrt{\frac{68}{5}} = \sqrt{13.6} = 3.69$$

$$C.V.(y) = \frac{\sigma_y}{\overline{y}} \text{ x } 100$$

$$= \frac{3.69}{15} \times 100$$

$$= 24.6\ \%$$

City A had more stable prices than City B, because the coefficient of variation is less in City A.

POSSIBLE QUESTIONS

PART –B(SIX MARKS)

1. Calculate the geometric mean for the following data:

x :	12	13	14	15	16	17
f :	5	4	4	3	2	1

2. Find the standard deviation of the following distribution:

Age :	20-25	25-30	30-35	35-40	40-45	45-50
No of persons:	170	110	80	45	40	35

3. Calculate the Median for the following.

Hourly Wages (in Rs.)	40 - 50	50 – 60	60 - 70	70 - 80	80 - 90	90 - 100
Number of Employees	10	20	15	30	15	10

4.Calculate the Standard Deviation for the following Frequency Distribution.

X	25	35	45	55	65
F	7	12	15	10	6

5. Calculate the Mean for the following Continuous Frequency Distribution.

Hourly Wages (in Rs.)	40 - 50	50 – 60	60 - 70	70 - 80	80 - 90	90 - 100
Number of Employees	10	20	15	30	15	10

6.Weekly Wages of a laborer are given below .Calculate Quartile Deviation and also the coefficient of Quartile Deviation

Wages(Rs)	15	30	45	60	75	80	Total
No.of Weeks	1	4	8	21	10	8	52

7. Calculate the range and its coefficient for the following data.

8 10 5 9 12 11

8. Coefficient of variation of two series are 75% and 90% and their standard deviations are 15 and 18 respectively. Find the mean .

9.Calculate the mean and standard deviation for the following data.

X:	6	9	12	15	18
F:	7	12	13	10	8

10.Calculate the Mean for the following Continuous Frequency Distribution.

Hourly Wages (in Rs.)	40 - 50	50 – 60	60 - 70	70 - 80	80 - 90	90 - 100
Number of Employees	10	20	15	30	15	10

11.Calculate mean and Standard Deviation of following frequency distribution of marks.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No.of students	5	12	30	45	50	37	21

PART –C (TEN MARKS)

1.The following data give the details about salaries (in thousands of rupees) of seven employees randomly selected from a Pharmaceutical Company.

Serial No.	1	2	3	4	5	6	7
Salary per Annum ('000)	89	57	104	73	26	121	81

Calculate the Standard Deviation and Coefficient of variance of the given data.

UNIT – II

SYLLABUS

Correlation and Regresson:Types of correlation-simple and Multiple,Positive and Negative,Linear and Non –Linear,partial and Total.Methods of calculating correlation coefficient:Scatter diagram,Karl pearson and spearmen (Rank)Correlation coefficient,Regression:Types,lines and equations,Linear regression-Least square method of solving regression-Least square method of solving regression equations ,X on Y and Y on X.

CORRELATION

Introduction:

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Definitions:

1. Correlation Analysis attempts to determine the degree of relationship between variables- Ya-Kun-Chou.
2. Correlation is an analysis of the covariation between two or more variables.- A.M.Tuttle.

Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject)

independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

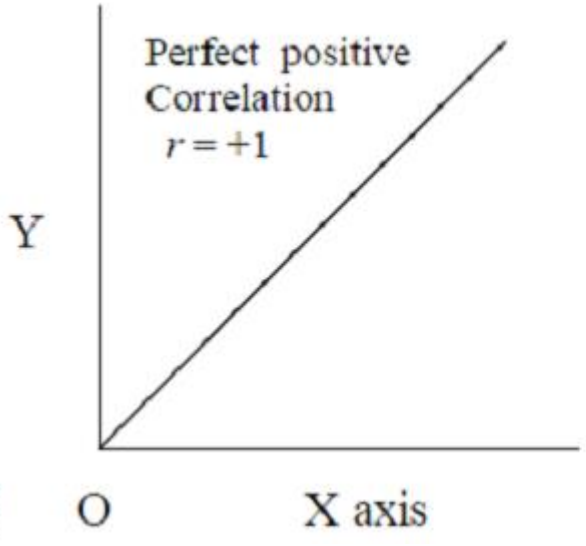
Uses of correlation:

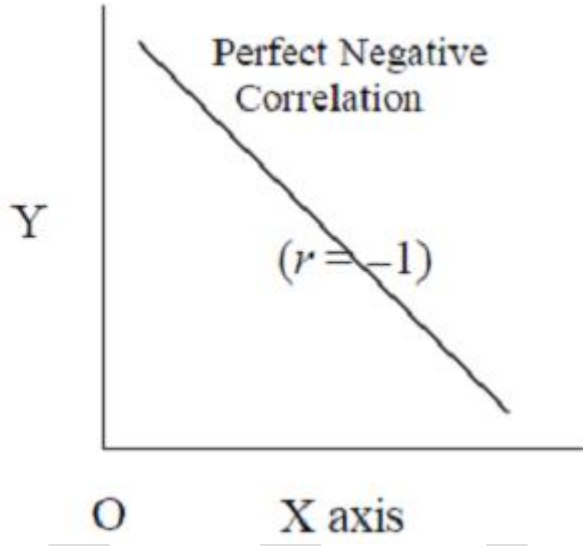
1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

Scatter Diagram:

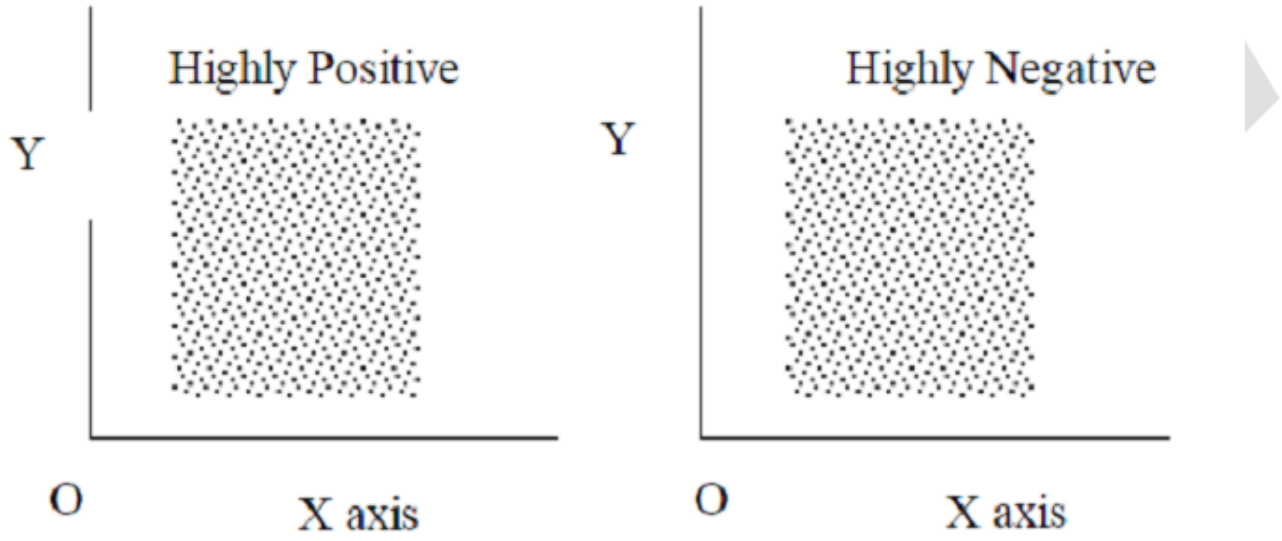
It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

1. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is
   Perfect positive correlation. We denote this as $r = +1$





1. If all the plotted dots lie on a straight line falling from upper left hand corner to lower right hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.
2. If the plotted points in the plane form a band and they show a rising trend from the lower left hand corner to the upper right hand corner the two variables are highly positively correlated.



1. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.
2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.

Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

Types of Correlation:

Correlation is classified into various types. The most important ones are

i) Positive and negative.
ii) Linear and non-linear.
iii) Partial and total.
iv) Simple and Multiple.

Positive and Negative Correlation:

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ' r' .

Co-variation:

The covariation between the variables x and y is defined as

Cov(x,y) = $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ where $\bar{x}, \bar{y}$ are respectively means of x and y and ' n' is the number of pairs of observations.

Karl pearson' s coefficient of correlation:

Karl pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. It is most widely used method in practice and it is known as pearsonian coefficient of correlation. It is denoted by 'r'. The formula for calculating 'r' is

(i) $r = \dfrac{Cov(x,y)}{\sigma_x . \sigma_y}$ where σ_x, σ_y are S.D of x and y respectively.

(ii) $r = \dfrac{\sum xy}{n\ \sigma_x\ \sigma_y}$

(iii) $r = \dfrac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}}$, $X = x - \bar{x}$, $Y = y - \bar{y}$

when the deviations are taken from the actual mean we can apply any one of these methods. Simple formula is the third one.

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

Steps:

1. Find the mean of the two series x and y.
2. Take deviations of the two series from x and y.

   $X = x - \bar{x}$, $Y = y - \bar{y}$

3. Square the deviations and get the total, of the respectiv squares of deviations of x and y and denote by ΣX^2 , ΣY^2 respectively.
4. Multiply the deviations of x and y and get the total and Divide by n. This is covariance.
5. Substitute the values in the formula.

$$r = \frac{\text{cov}(x,y)}{\sigma x.\sigma y} = \frac{\Sigma(x-\bar{x})\ (y\ -\ \bar{y})/n}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}$$

The above formula is simplified as follows

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2.\Sigma Y^2}}, \quad X = x-\bar{x}\ , Y = y-\bar{y}$$

Example 1:

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

Solution:

x	Y	$X = x - \bar{x}$ $X = x - 67$	X^2	$Y = y - \bar{y}$ $Y = y - 68$	Y^2	XY
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
469	476	0	28	0	34	25

$$\bar{x} = \frac{469}{7} = 67 \; ; \; \bar{y} = \frac{476}{7} = 68$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}} = \frac{25}{30.85} = 0.81$$

Since r = + 0.81, the variables are highly positively correlated. (ie) Tall fathers have tall sons.

Working rule (i)

We can also find r with the following formula

We have $r = \dfrac{Cov(x, y)}{\sigma_x . \sigma_y}$

$$Cov(x,y) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} = \frac{\Sigma(xy + \bar{x}y - \bar{y}x - \bar{x}\bar{y})}{n}$$

$$= \frac{\Sigma xy}{n} - \frac{\bar{y}\Sigma x}{n} - \frac{\bar{x}\Sigma y}{n} + \frac{\Sigma \bar{x}\bar{y}}{n}$$

$$\text{Cov}(x,y) = \frac{\Sigma xy}{n} - \cancel{\bar{y}\bar{x}} - \bar{x}\bar{y} + \cancel{\bar{x}\bar{y}} = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

$$\sigma^2 x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2, \quad \sigma^2 y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

$$\text{Now } r = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$$

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sqrt{\left(\frac{\Sigma x^2}{n} - \bar{x}^2\right)} \sqrt{\left(\frac{\Sigma y^2}{n} - \bar{y}^2\right)}}$$

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

Note: In the above method we need not find mean or standard deviation of variables separately.

Example 2:

Calculate coefficient of correlation from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597

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

$$r = \frac{9 \times 597 - 45 \times 108}{\sqrt{(9\times 285 - (45)^2).(9\times 1356 - (108)^2)}}$$

$$r = \frac{5373 - 4860}{\sqrt{(2565 - 2025).(12204 - 11664)}}$$

$$= \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540} = 0.95$$

Working rule (ii) (shortcut method)

We have $r = \dfrac{Cov(x,y)}{\sigma_x . \sigma_y}$

where Cov(x,y) = $\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$

Take the deviation from x as x – A and the deviation from y as y – B

$$Cov(x,y) = \frac{\Sigma\ [(x-A)-(\bar{x}-A)]\ [(y-B)-(\bar{y}-B)]}{n}$$

$$= \frac{1}{n}\ \Sigma\ [(x-A)\ (y-B)\ -\ (x-A)\ (\bar{y}-B)$$

$$-\ (\bar{x}-A)(y-B)+(\bar{x}-A)(\bar{y}-B)]$$

$$= \frac{1}{n}\ \Sigma\ [(x-A)\ (y-B)\ -\ (\bar{y}-B)\ \frac{\Sigma(x-A)}{n}$$

$$-\ (\bar{x}-A)\frac{\Sigma(y-B)}{n}+\frac{\Sigma(\bar{x}-A)(\bar{y}-B)}{n}$$

$$= \frac{\Sigma(x-A)(y-B)}{n}\ -\ (\bar{y}-B)\ \left(\bar{x}-\frac{nA}{n}\right)$$

$$-\ (\bar{x}-A)\ \left(\bar{y}-\frac{nB}{n}\right)+(\bar{x}-A)\ (\bar{y}-B)$$

$$= \frac{\Sigma(x-A)(y-B)}{n} - (\bar{y}-B)\ (\bar{x}-A)$$

$$- \cancel{(\bar{x}-A)\ (\bar{y}-B)} + \cancel{(\bar{x}-A)\ (\bar{y}-B)}$$

$$= \frac{\Sigma(x-A)(y-B)}{n} - (\bar{x}-A)\ (\bar{y}-B)$$

Let $x-A=u$; $y-B=v$; $\bar{x}-A=\bar{u}$; $\bar{y}-B=\bar{v}$

$$\therefore \text{Cov}(x,y) = \frac{\Sigma uv}{n} - \bar{u}\bar{v}$$

$$\sigma\sigma_x^2 = \frac{\Sigma u^2}{n} - \bar{u}^2 = \sigma u^2$$

$$\sigma\sigma_y^2 = \frac{\Sigma v^2}{n} - \bar{v}^2 = \sigma v^2$$

$$\therefore r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{\left[n\Sigma u^2 - (\Sigma u)^2\right].\left[(n\Sigma v^2) - (\Sigma v)^2\right]}}$$

Limitations:

1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
2. Extreme items of variables are being unduly operated on correlation coefficient.
3. Existence of correlation does not necessarily indicate cause-effect relation.

Interpretation:

The following rules helps in interpreting the value of ' r' .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

Rank Correlation:

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc.The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as $r = 1 - \frac{6\Sigma D^2}{n^3 - n}$ r = rank correlation coefficient.

Note: Some authors use the symbol ρ for rank correlation.
ΣD^2 = sum of squares of differences between the pairs of ranks.
n = number of pairs of observations.

The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5^{th} rank, the common rank to be assigned to each item is $\frac{5+6}{2} = 5.5$ which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}(m^3\text{-}m)$. A slightly different formula is used when there is more than one item having the same value.

The formula is

$$r = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots.]}{n^3 - n}$$

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example 6:

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

Price of tea	Rank	Price of coffee	Rank	D	D^2
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0
70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
					$\Sigma D^2 = 6$

$$r = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 6}{7^3 - 7}$$

$$= 1 - \frac{36}{336} = 1 - 0.1071$$

$$= 0.8929$$

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

REGRESSION

In mathematics, regression is one of the important topics in statistics. The process of determining the relationship between two variables is called as regression. It is also one of the statistical analysis methods that can be used to assessing the association between the two different variables.

9.1 Introduction:

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the ' explained' variable and the other the ' independent' or the ' predicting' variable. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product, consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

9.1.1 Definition:

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

9.2 Types Of Regression:

The regression analysis can be classified into:

a) Simple and Multiple
b) Linear and Non –Linear
c) Total and Partial

a) Simple and Multiple:

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than

two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as $y = f(x,z)$.

b) Linear and Non-linear:

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predective value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

c) Total and Partial:

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

9.3 Linear Regression Equation:

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is

$$Y = a + bX \ldots\ldots (1)$$

And X on Y is

$$X = a + bY \ldots\ldots (2)$$

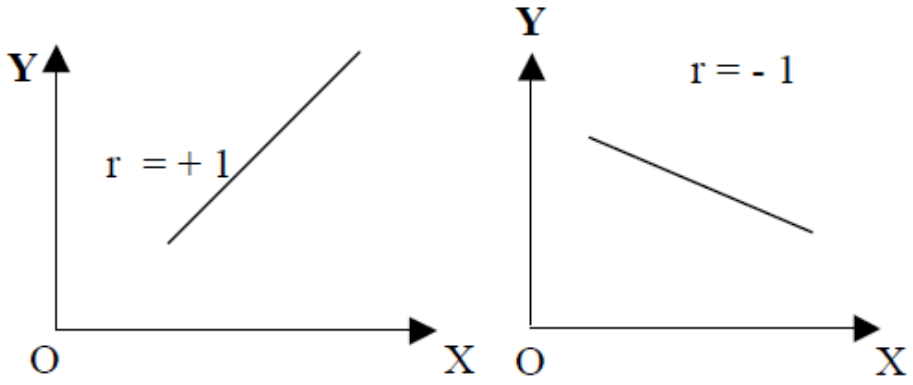
a, b are constants.

From (1) We can estimate Y for known value of X.
(2) We can estimate X for known value of Y.

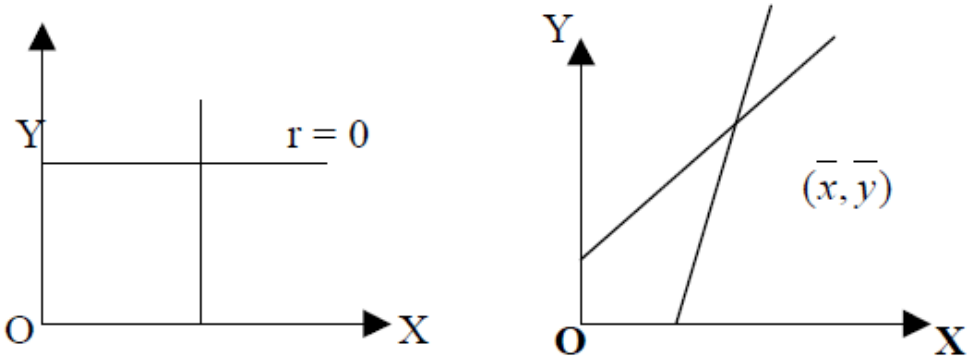
9.3.1 Regression Lines:

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

For perfect correlation, positive or negative i.e., $r = \pm 1$, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.



Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X-axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y-axis will touch the mean value of Y.



9.3.2 Principle of 'Least Squares' :

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

(i) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

$\sum(X - Xc) = 0$ or $\sum (Y - Yc) = 0$

Where Xc and Yc are the values obtained by regression analysis.

(ii) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

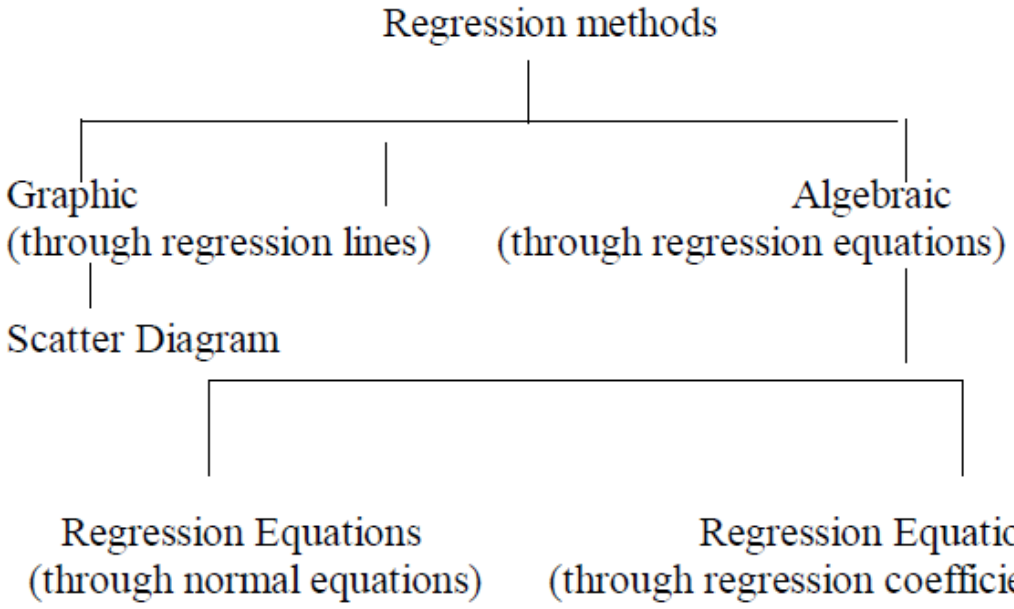
$\sum(Y - Yc)^2 < \sum (Y - Ai)^2$

Where Ai = corresponding values of any other straight line.

(iii) The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $\bar{x}, \bar{y}$.

9.4 Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:



9.4.1 Graphic Method:

Scatter Diagram:

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

9.4.2 Algebraic Methods:

(i) Regression Equation.

The two regression equations

for X on Y; $X = a + bY$

And for Y on X; $Y = a + bX$

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, $X = a + bY$

The normal equations are

$$\sum X = na + b \sum Y \text{ and}$$

$$\sum XY = a\sum Y + b\sum Y^2$$

For the equation, $Y = a + bX$, the normal equations are

$$\sum Y = na + b\sum X \text{ and}$$

$$\sum XY = a\sum X + b\sum X^2$$

From these normal equations the values of a and b can be determined.

Example 1:

Find the two regression equations from the following data:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Solution:

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

Regression equation of Y on X is $Y = a + bX$ and the normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values, we get

$$40 = 5a + 30b \ldots\ldots (1)$$
$$214 = 30a + 220b \ldots\ldots (2)$$

Multiplying (1) by 6

$$240 = 30a + 180b \ldots\ldots (3)$$

(2) – (3) $\quad -26 = 40b$

$$\text{or } b = -\frac{26}{40} = -0.65$$

Now, substituting the value of 'b' in equation (1)

$$40 = 5a - 19.5$$
$$5a = 59.5$$
$$a = \frac{59.5}{5} = 11.9$$

Hence, required regression line Y on X is $Y = 11.9 - 0.65\,X$.

Again, regression equation of X on Y is

$X = a + bY$ and

The normal equations are

$$\sum X = na + b\sum Y \text{ and}$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Now, substituting the corresponding values from the above table, we get

$$30 = 5a + 40b \ldots(3)$$
$$214 = 40a + 340b \ldots(4)$$

Multiplying (3) by 8, we get

$$240 = 40a + 320\,b \ldots(5)$$

(4) – (5) gives

$$-26 = 20b$$
$$b = -\frac{26}{20} = -1.3$$

Substituting $b = -1.3$ in equation (3) gives

$$30 = 5a - 52$$
$$5a = 82$$
$$a = \frac{82}{5} = 16.4$$

Hence, Required regression line of X on Y is

$X = 16.4 - 1.3Y$

(ii) Regression Co-efficents:

The regression equation of Y on X is $y_e = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Here, the regression Co.efficient of Y on X is

$$b_1 = b_{yx} = r\frac{\sigma_y}{\sigma_x}$$
$$y_e = \bar{y} + b_1(x - \bar{x})$$

The regression equation of X on Y is

$$X_e = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Here, the regression Co-efficient of X on Y

$$b_2 = b_{xy} = r\frac{\sigma_x}{\sigma_y}$$
$$X_e = \bar{X} + b_2(y - \bar{y})$$

If the deviation are taken from respective means of x and y

$$b_1 = b_{yx} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{\sum xy}{\sum x^2} \quad \text{and}$$

$$b_2 = b_{xy} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})^2} = \frac{\sum xy}{\sum y^2}$$

where $x = X - \overline{X}, y = Y - \overline{Y}$

If the deviations are taken from any arbitrary values of x and y (short – cut method)

$$b_1 = b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - \left(\sum u\right)^2}$$

$$b_2 = b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - \left(\sum v\right)^2}$$

where u = x – A : v = Y-B

A = any value in X

B = any value in Y

9.5 Properties of Regression Co-efficient:

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. correlation coefficient is the geometric mean of the regression coefficients ie, $r = \pm\sqrt{b_1 b_2}$
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. Symbolically $\frac{b_1 + b_2}{2} \geq r$

7. If r=0, the variables are uncorrelated , the lines of regression become perpendicular to each other.
8. If r= ± 1, the two lines of regression either coincide or parallel to each other
9. Angle between the two regression lines is $\theta = \tan^{-1}\left[\frac{m_1 - m_2}{1 + m_1 m_2}\right]$

where m_1 and,m_2 are the slopes of the regression lines X on Y and Y on X respectively.

10. The angle between the regression lines indicates the degree of dependence between the variables.

Example 2:

If 2 regression coefficients are $b_1=\frac{4}{5}$ and $b_2=\frac{9}{20}$ .What would be the value of r?

Solution:

The correlation coefficient , $r = \pm\sqrt{b_1 b_2}$

$$= \sqrt{\frac{4}{5} \text{x} \frac{9}{20}}$$

$$= \sqrt{\frac{36}{100}} = \frac{6}{10} = 0.6$$

Example 3:

Given $b_1 = \frac{15}{8}$ and $b_2 = \frac{3}{5}$, Find r

Solution:

$r = \pm\sqrt{b_1 b_2}$

$= \sqrt{\frac{15}{8} \text{x} \frac{3}{5}}$

$= \sqrt{\frac{9}{8}} = 1.06$

It is not possible since *r*, cannot be greater than one. So the given values are wrong

Example 4:

Compute the two regression equations from the following data.

X	1	2	3	4	5
Y	2	3	5	4	6

If x =2.5, what will be the value of y?

Solution:

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
1	2	-2	-2	4	4	4
2	3	-1	-1	1	1	-1
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
15	20	20		10	10	9

$$\overline{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{20}{5} = 4$$

Regression Co efficient of Y on X

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{9}{10} = 0.9$$

Hence regression equation of Y on X is

$$Y = \overline{Y} + b_{yx}(X - \overline{X})$$

$$= 4 + 0.9\ (X - 3)$$

$$= 4 + 0.9X - 2.7$$

$$= 1.3 + 0.9X$$

when X = 2.5

$$Y = 1.3 + 0.9 \times 2.5$$

$$= 3.55$$

Regression co efficient of X on Y

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{9}{10} = 0.9$$

So, regression equation of X on Y is

$$X = \overline{X} + b_{xy}(Y - \overline{Y})$$

$$= 3 + 0.9\,(Y - 4)$$

$$= 3 + 0.9Y - 3.6$$

$$= 0.9Y - 0.6$$

Example 6:

In a correlation study, the following values are obtained

	X	Y
Mean	65	67
S.D	2.5	3.5

Co-efficient of correlation = 0.8
Find the two regression equations that are associated with the above values.

Solution:

Given,

$\overline{X} = 65$, $\overline{Y} = 67$, $\sigma_x = 2.5$, $\sigma_y = 3.5$, r = 0.8

The regression co-efficient of Y on X is

$$b_{yx} = b_1 = r\frac{\sigma_y}{\sigma_x}$$

$$= 0.8 \times \frac{3.5}{2.5} = 1.12$$

The regression coefficient of X on Y is

$$b_{xy} = b_2 = r\,\frac{\sigma_x}{\sigma_y}$$

$$= 0.8 \times \frac{2.5}{3.5} = 0.57$$

Hence, the regression equation of Y on X is

$$Y_e = \overline{Y} + b_1(X - \overline{X})$$

$$= 67 + 1.12\ (X-65)$$

$$= 67 + 1.12\ X - 72.8$$

$$= 1.12X - 5.8$$

The regression equation of X on Y is

$$X_e = \overline{X} + b_2(Y - \overline{Y})$$
$$= 65 + 0.57\ (Y\text{-}67)$$
$$= 65 + 0.57Y - 38.19$$
$$= 26.81 + 0.57Y$$

9.7 Uses of Regression Analysis:

1. Regression analysis helps in establishing a functional relationship between two or more variables.
2. Since most of the problems of economic analysis are based on cause and effect relationships, the regression analysis is a highly valuable tool in economic and business research.
3. Regression analysis predicts the values of dependent variables from the values of independent variables.
4. We can calculate coefficient of correlation (r) and coefficient of determination (r^2) with the help of regression coefficients.
5. In statistical analysis of demand curves, supply curves, production function, cost function, consumption function etc., regression analysis is widely used.

9.8 Difference between Correlation and Regression:

S.No	Correlation	Regression
1.	Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	Regression means going back and it is a mathematical measure showing the average relationship between two variables
2.	Both the variables X and Y are random variables	Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables.
3.	It finds out the degree of relationship between two variables and not the cause and effect of the variables.	It indicates the causes and effect relationship between the variables and establishes functional relationship.

4.	It is used for testing and verifying the relation between two variables and gives limited information.	Besides verification it is used for the prediction of one value, in relationship to the other given value.
5.	The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1	Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
6.	There may be spurious correlation between two variables.	In regression there is no such spurious regression.
7.	It has limited application, because it is confined only to linear relationship between the variables.	It has wider application, as it studies linear and non-linear relationship between the variables.
8.	It is not very useful for further mathematical treatment.	It is widely used for further mathematical treatment.
9.	If the coefficient of correlation is positive, then the two variables are positively correlated and vice-versa.	The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

POSSIBLE QUESTIONS

PART – B(SIX MARKS)

1.Calculate Karl Pearson's correlation coefficient between the marks in English and Hindi obtained by 10 students:

Marks in English :	10	25	13	25	22	11	12	25	21	20
Marks in Hindi:	12	22	16	15	18	18	17	23	24	17

2.From the following data calculate the regression equations taking deviation of items from the mean of X and Y series:

X:	6	2	10	4	8
Y:	9	11	5	8	7

3.Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute Rank Correlation Coefficient.

X	25	20	28	22	40	60	20
Y	40	30	50	30	20	10	30

4.You are given the following data:

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8

Correlation coefficient between X and Y = 0.66
i) Find the two regression equations. ii) Find r.

5.Explain the different methods to calculating correlation coefficient between two Variables?

6.From the data given below find the two Regression Equations.

X	10	12	13	12	16	15
Y	40	38	43	45	37	43

i) Estimate Y when X = 20.
ii) Estimate X when Y = 25.

7.Find Karl Pearson's Co-efficient of correlation

X :	100	101	102	102	100	99	97	98	96	95
Y :	98	99	99	97	95	92	95	94	90	91

(OR)

8.Marks obtained by 7 students in Accountancy (X) and Statistics (Y) are given below. Compute Rank Correlation Coefficient.

X	15	20	28	12	40	60	20
Y	40	30	50	30	20	10	30

9.Explain different between correlation and regression .

10.For the following data calculate the rank correlation coefficient between X and Y.

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y :	12	9	6	10	3	5	4	7	8	2	11	1

PART- C (TEN Marks)

1. Calculate the two regression equation from the following data:

X	:	10	12	13	12	16	15
Y	:	40	38	43	45	37	43

2. Find two regression equations for the following two series, what is most likely value of X when Y = 20 and most likely value of Y when X = 22.

X:	35	25	29	31	27	24	33	36
Y:	23	27	26	21	24	20	29	30

UNIT – III

Testing of Hypothesis: Introduction to Inferential Statistics: Null and alternative hypothesis, Type I and Type II errors, Standard error, level of significance, acceptance and rejection regions and procedure for testing hypothesis. Large sample test - Z test - tests for means, variances and proportions, Small sample tests based on t, F and Chi- square distributions.

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 a 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.

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

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 $x_1, x_2, x_3, ..., x_n$ 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 $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

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

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

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. $s = \sqrt{\dfrac{1}{n-1}\sum (x - \bar{x})^2}$.

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

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

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

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

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 $x_1, x_2, x_3, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2,

for large sample, sample standard deviation s 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 $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

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

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

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

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

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

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

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

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

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

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

- **As σ is unknown,**

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Step 2: If μ_0 falls into the above confidence intervals, then

do *not* reject H_0. Otherwise, reject H_0.

Example 1:

The average starting salary of a college graduate is $19000 according to government's report. The average salary of a random sample of 100 graduates is $18800. The standard error is 800. Is the government's report reliable as the level of significance is 0.05. Find the p-value and test the hypothesis in

(a) with the level of significance $\alpha = 0.01$. The other report by some institute indicates that the average salary is $18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \text{ vs. } H_a : \mu \neq \mu_0 = 19000,$$
$$n = 100, \bar{x} = 18800, s = 800, \alpha = 0.05$$

Then,

$$|z| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| = \left| \frac{18800 - 19000}{800/\sqrt{100}} \right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96 .$$

Therefore, reject H_0.

(b)

$$\text{p - value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, ***not*** reject H_0.

(c)

$$H_0 : \mu = \mu_0 = 18900 \text{ vs } H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 18800 \pm z_{0.025} \frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, ***not*** reject H_0 .

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \text{ vs. } H_a : u \neq 40.$$

based on

(a) classical hypothesis test
(b) p-value
(c) confidence interval.
[solution:]

$$\bar{x} = 38,\ s = 7,\ u_0 = 40,\ n = 49,\ z = \frac{\bar{x} - u_0}{s/\sqrt{n}} = \frac{38 - 40}{7/\sqrt{49}} = -2.$$

(a)

$$|z| = 2 > 1.96 = z_{0.025}$$

we reject H_0 .

(b)

$$p-value = P(|Z| > |z|) = P(|Z| > 2) = 2*(1-0.9772) = 0.0456 < 0.05 = \alpha$$

we reject H_0 .

(c)

$100\times(1-\alpha)\% = 95\%$ confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 38 \pm z_{0.025}\frac{7}{\sqrt{49}} = 38 \pm 1.96 = [36.04, 39.96].$$

Since $40 \notin [36.04, 39.96]$, we reject H_0 .

Hypothesis Testing for the Mean (Small Samples)

For samples of size less than 30 and when σ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean μ.

Using the t-Test for a Mean μ when the sample is small		
Procedure	**Equations**	**Example 4**
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0 : \mu \geq 16500$ $H_a : \mu < 16500$ $n = 14, \bar{x} = 15700, s = 1250$
Specify the level of significance	Specify α	$\alpha = 0.05$
Identify the degrees of freedom and sketch the sampling distribution	$d.f = n - 1$	$d.f. = 13$
Determine any critical values. If test is left tailed, use One tail, α column with a negative sign. If test is right tailed, use One tail, α column with a positive sign. If test is two tailed, use Two tails, α column with a negative and positive sign.	Table 5 (*t*-distribution) in appendix B	The test is left-tailed. Since test is left tailed and $d.f = 13$, the critical value is $t_0 = -1.771$
Determine the rejection regions.	The rejection region is $t < t_0$	The rejection region is $t < -1.771$
Find the standardized test statistic	$t = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} \approx \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$	$t = \dfrac{15700 - 16500}{1250/\sqrt{14}} \approx -2.39$
Make a decision to reject or fail to reject the null hypothesis	If t is in the rejection region, reject H_0, Otherwise do not reject H_0	Since $-2.39 < -1.771$, reject H_0
Interpret the decision in the context of the original claim.		Reject claim that mean is at least 16500.

Chi-square Tests and then F -Distribution

Goodness of Fit

DEFINITION :A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

H_0: The distribution fits the proposed proportions

H_1: The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$$E_i = np_i$$

where *n* is the number of trials (the sample size) and p_i is the assumed probability of the *i*th category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequency of each category and *E* represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true* .

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5.

Performing the Chi-Square Goodness-of-Fit Test (p 496)		
Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0: Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6-1 = 5$
Find the critical value	χ^2_α: Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 15.086$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Survey results, n = 500 Classical O= 8 E = .04*500 = 20 Country O = 210 E = .36*500 = 180 Gospel O = 7 E = .11*500 = 55 Oldies O = 10 E = .02*500 = 10 Pop O = 75 E = .18*500 = 90 Rock O= 125 E = .29*500 = 145 Substituting $\chi^2 = 22.713$
Make the decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 22.713 > 15.086 we reject the null hypothesis Equivalently $P(X \geq 22.713) < 0.01$so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		Music preferences differ from the radio station's claim.

Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)

The data from the example above (Example 2 p 497) will be used.

Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns 'Music Types', 'Observed' and 'Distribution' are entered in the gray row at the top.)

Select **Calc->Calculator**, **Store the results in** C4, and calculate the **Expression** C3*500, click **OK**, Name C4 'Expected' since it now contains the expected frequencies

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10
Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

Sum of C5

Sum of C5 = 22.7132

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select **Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic 22.7132 for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

Cumulative Distribution Function

Chi-Square with 5 DF

x	P(X <= x)
22.7132	0.999617

$P(X \leq 22.7132) = 0.999617$ So the P-value = 1 – 0.999617 = 0.000383. This is less that $\alpha = 0.01$ so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: $\alpha = 0.05$
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2_{0.05}(d.f.=5)=11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where O is the actual number of M&M's of each color in the bag and E is the proportions specified under H_0 times the total number.
Reject H_0 if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c* contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals B_j.

	A_1	**A_2**	**A_3**	**A_4**	**A**
B_1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B_2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B_3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum\ of\ row\ i) * (sum\ of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550} \approx 25.64$	$\frac{300*87}{550} \approx 47.45$	$\frac{300*193}{550} \approx 105.27$	$\frac{300*180}{550} \approx 98.18$	$\frac{300*43}{550} \approx 23.45$	300
Large	$\frac{250*47}{550} \approx 21.36$	$\frac{250*87}{550} \approx 39.55$	$\frac{250*193}{550} \approx 87.73$	$\frac{250*180}{550} \approx 81.82$	$\frac{250*43}{550} \approx 19.55$	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequencies and *E* represents the expected frequencies.

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0: CEO's ages are independent of company size H_1: CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$
Find the critical value.	χ_α^2 :Obtain from Table 6, Appendix B	$\chi_\alpha^2 \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi_\alpha^2$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above
Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   Solution:

   There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion $p = 5/20 = 0.25$. Using the PROB-VALUE method the steps in this test are:

   1) H_0: □ = 0.5 and H_A: □ ¹ 0.5
   2) We will use the *Z*-distribution
   3) We will use the 5%-level, thus □ = 0.05
   4) The test statistic is $z = (0.25 - 0.5) / \sqrt{0.25/20} = -2.24$
   5) Table A-4 shows that $P(|Z| > 2.24) \approx 0.025$.
   6) Because PROB-VALUE < □, we reject H_0. We conclude □ is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

   Solution:

Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$P(X \geq 8) = 0.1208 + 0.0537 + 0.0161 + 0.0029 + 0.002 = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE > □□□ Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

Solution:

(a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference.

We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level □ = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5}+\frac{39^2}{5}+\frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the □2 distribution with $3-1=2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5
2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $H_0 : \eta = 0$ and $H_A : \eta \neq 0$ □□We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with □ = 0.5. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the □ = 10% level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

	On the Floodplain	Off the Floodplain
Insured	50	10
No Insurance	15	25

Test a relevant hypothesis.

Solution:

We will do a χ^2 test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

	On the Floodplain	Off the Floodplain
Insured	50 (39)	10 (21)
No Insurance	15 (26)	25 (14)

The corresponding χ^2 value is 22.16. Table A-8 shows that with 1 degree of freedom, $P(\chi^2 > 20)$ is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any $\alpha < 0.0005$), we can reject the null hypothesis.

The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

Day	Percentage of sunshine	Day	Percentage of sunshine	Day	Percentage of sunshine
1	75	11	21	21	77
2	95	12	96	22	100
3	89	13	90	23	90
4	80	14	10	24	98
5	7	15	100	25	60
6	84	16	90	26	90
7	90	17	6	27	100
8	18	18	0	28	90
9	90	19	22	29	58
10	100	20	44	30	0

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

Solution:

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are $n_x = 21$ sunny days, and $n_y = 9$ cloudy days. Because $n_x < 20$, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

9. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the χ^2 test with $k = 6$ classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

x_i	$S(x_i)$	$F(x_i)$	$\lvert S(x_i)-F(x_i)\rvert$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017
4.4	0.060	0.032	0.028
…	…	…	…
5.9	0.780	0.692	0.088
…	…	...	…

6.7	0.960	0.960	0.000
6.8	0.980	0.972	0.008
6.9	1.000	0.981	0.019

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the χ^2 table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

Group	Minimum	Maximum	O_j	E_j	$(O_j-E_j)^2/E_j$
1	4.000	4.990	9	3.3	10.13
2	5.000	5.490	10	17.0	2.89
3	5.500	5.990	20	21.7	0.14
4	6.000	6.990	11	7.0	2.24

The observed Chi-square value is 15.4. With $k - p - 1 = 4 - 2 - 1 = 1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the χ^2 test to be reliable.

Nonparametric Hypothesis Testing

EARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Identify and cite examples of situations in which nonparametric tests of hypothesis are appropriate.
2. Explain the logic of nonparametric hypothesis testing for ordinal variables as applied to the Mann-Whitney *U* and runs tests.
3. Perform Mann-Whitney *U* and runs tests using the five-step model as a guide, and correctly interpret the results.
4. Select an appropriate nonparametric test.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

5. **FORMULA 1**
6.
7. where: n_1 = number of cases in sample 1
8. n_2 = number of cases in sample 2
9. $\sum R_1$ = the sum of the ranks for sample 1
10. For our sample problem above

$$U = (12)(12) + \frac{12(13)}{2} - 173.5$$

11.

$$U = 144 + \frac{156}{2} - 173.5$$

12.
13. $U = 48.5$
14.
15. Note that we could have computed the *U* by using data from sample 2. This alternative solution, which we will label *U*′ (*U* prime), would have resulted in a larger value for *U*. The smaller of the two values, *U* or *U*′, is always taken as the value of *U*. Once *U* has been calculated, *U*′ can be quickly determined by means of Formula 2:
16. The alternative or research hypothesis is usually a statement to the effect that the two populations are different. This form for H_1 would direct the use of a two-tailed test. It is perfectly possible to use one-tailed tests with Mann-Whitney *U* when a direction for the difference can be predicted, but, to conserve space and time, we will consider only the two-tailed case.

17. In step 3, we will take advantage of the fact that, when total sample size (the combined number of cases in the two samples) is greater than or equal to 20, the sampling distribution of *U* approximates normality. This will allow us to use the *Z*-score table (see Appendix A of the textbook) to find the critical region as marked by *Z* (critical).
18. To compute the Mann-Whitney *U* test statistic (step 4), the necessary formulas are
19. **FORMULA 3** $Z \text{ (obtained)} = \dfrac{U - \mu_U}{\sigma_U}$
20. **FORMULA 5** $\sigma_U = \sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$
21. We now have all the information we need to conduct a test of significance for *U*.

22. <u>Step 2. Stating the Null Hypothesis</u>

23. H_0: The populations from which the samples are drawn are identical on the variable of interest.
24. (H_1: The populations from which the samples are drawn are different on the variable of interest.)
25. **Step 3. Selecting the Sampling Distribution and Establishing the Critical Region**.
26. Sampling distribution = *Z* distribution
27. Alpha = 0.05
28. *Z* (critical) = ± 1.96
29. **Step 4. Calculating the Test Statistic**. With *U* equal to 48.5, □U equal to 72, and σ_U of 17.32,

30. $Z \text{ (obtained)} = \dfrac{U - \mu_U}{\sigma_U}$

31. $Z \text{ (obtained)} = \dfrac{48.5 - 72}{17.32}$

32. *Z* (obtained) = -1.36

33. **Step 5. Making a Decision**. The test statistic, a *Z* (obtained) of -1.36, does not fall in the critical region as marked by the *Z* (critical) of ± 1.96. Therefore, we fail to reject the null of no difference. Male students are not significantly different from female students in terms of their level of satisfaction with the social life available on campus. Note that if we had used the *U´* value of 95.5 instead of *U* in computing the test statistic, the value of *Z* (obtained) would have been +1.36, and our decision to fail to reject the null would have been exactly the same.
34. **Making Assumptions**.
35. Model: Independent random sampling
36. Level of measurement is ordinal
37. **Step 2. Stating the Null Hypothesis.**
38. H_0: The two populations are identical on level of pain.
39. (H_1: The two populations are different on level of pain.)

POSSIBLE QUESTIONS

PART – B (SIX MARKS)

1.If $x \geq 1$ is the critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$ on the basis of a single observation from the population $f(x, \theta) = \theta e^{-\theta x}$; $0 \leq x < \infty$; $\theta > 0$ and $\theta = 0$ elsewhere ,obtain the values of the sizes of type 1 error ($\propto$) and type 2 error (β) and hence show that $1-\beta = \propto^2$

2.If $x \geq 1$ is the critical region for testing $H_0 : \theta = 2$ against the alternative $\theta = 1$ on the basis of the single observation from the population

3.If the density function of arandom variable X is given by $f(x, \theta) = 1/\theta\, e^{-x/\theta}$; $0 \leq x \leq \infty$; $\theta > 0$ and $\theta = 0$;elsewhere.test $H_0 : \theta = 2$ against $H_1 : \theta = 1$,using the random sample $\{x_1, x_2, x_3\}$ of size 3 .Given that critical region $W = \{(x_1, x_2, x_3)\ (x_1 + x_2 + x_3 \geq 15\}$ Find the significance level of the test.

4.An Urn contains either 3 red and 5 blue balls or 5 red and 3 blue balls .3 balls are drawn at random from the Urn. If less than 3 red balls are obtained it will be concluded that the Urn contains 3 red and 5 blue balls. Calculate the values of the Type I and Type II errors

5.The following are measurements of the breaking strength of a certain kind of 2 – inch cotton ribbon in pounds :

163 165 160 189 161 171 158 151 169 162 163 139 172 165 148 166 172 163 187 173.Use the sign test to test the null hypothesis $\mu = 160$ against the alternative hypothesis $\mu > 160$ at the 0.05 level of significance .

6.Random Sample of 400 men and 600 women were asked whether they would like to have a flyover near their residence .200 men and 325 women were in favour of the proposal . test the hypothesis that proportions of men and women in favour of the proposal are same against that they are not at 5 % level .

7.The specification for a certain kind of ribbon call for a mean breaking strength of 185 pounds. If five oieces randomly selected from different rolls have breaking strength of 171.6’ 191.8 , 178.3 ,184.9 and 189.1 pounds , testing the null hypothesis $\mu = 185$ pounds against the alternative hypothesis $\mu < 185$ pounds at the 0.05 level of significance .

8.Suppose that it is known from experience that the standard deviation of the weight of 8 –ounce packages of cookies made by a certain bakery is 0.16 ounce .To check whether the true average

9.A discrete random variable X follows a binomial distribution B (10,P) where $p \in \{1/4, 1/2\}$ .If the member of a random sample of size 1 is less than or equal to 3 we reject $H_0 : p = ½$ accept $H_1 : p = ¼$ .Find the small test based on t.

10.State and discuss about Critical and acceptance region.

PART – C (TEN MARKS)

11.A Company has the head office at calculate and a branch at Bombay .The personal director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose .Out of a sample of 500 worker at Calcutta 62% forward the new plan .At Bombay Out of a sample of 400 workers,41% were against the new plan .Is there any significance difference between the two groups in their attitude towards the new plan at 5 % level.

UNIT – IV

SYLLABUS

Estimation and Design of Experiment: Point Estimation - characteristics of estimation - interval estimation -interval estimates of mean, standard deviation and\proportion. Design of Experiments: Completely Randomized Design (CRD), Randomized Block Design (RBD) and Latin Square design (LSD) Models

Design of Experiment

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way. The designing of experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid.

It is important to understand first the basic terminologies used in the experimental design.

Experimental unit:

For conducting an experiment, the experimental material is divided into smaller parts and each part is referred to as experimental unit. The experimental unit is randomly assigned to a treatment is the experimental unit. The phrase "randomly assigned" is very important in this definition.

Experiment:

A way of getting an answer to a question which the experimenter wants to know.

Treatment

Different objects or procedures which are to be compared in an experiment are called treatments.

Sampling unit:

The object that is measured in an experiment is called the sampling unit. This may be different from the experimental unit.

Factor:

A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termed as fixed factor if all the levels of interest are included in the experiment.

A factor is termed as random factor if all the levels of interest are not included in the experiment and those that are can be considered to be randomly chosen from all the levels of interest.

Replication:

It is the repetition of the experimental situation by replicating the experimental unit.

Experimental error:

The unexplained random part of variation in any experiment is termed as experimental error. An estimate of experimental error can be obtained by replication.

Treatment design:

A treatment design is the manner in which the levels of treatments are arranged in an experiment.

Example: (Ref.: Statistical Design, G. Casella, Chapman and Hall, 2008)

Suppose some varieties of fish food is to be investigated on some species of fishes. The food is placed in the water tanks containing the fishes. The response is the increase in the weight of fish. The experimental unit is the tank, as the treatment is applied to the tank, not to the fish. Note that if the experimenter had taken the fish in hand and placed the food in the mouth of fish, then the fish would have been the experimental unit as long as each of the fish got an independent scoop of food.

Design of experiment:

One of the main objectives of designing an experiment is how to verify the hypothesis in an efficient and economical way. In the contest of the null hypothesis of equality of several means of normal populations having same variances, the analysis of variance technique can be used. Note that such techniques are based on certain statistical assumptions. If these assumptions are violated, the outcome of the test of hypothesis then may also be faulty and the analysis of data may be meaningless.So the main question is how to obtain the data such that the assumptions are met and the data is readily available for the application of tools like analysis of variance. The designing of such mechanism to obtain such data is achieved by the design of experiment. After obtaining the sufficient experimental unit, the treatments are allocated to the experimental units in a random fashion. Design of experiment provides a method by which the treatments are placed at random on the experimental units in sucha way that the responses are estimated with the utmost precision possible.

Principles of experimental design:

There are three basic principles of design which were developed by Sir Ronald A. Fisher.

(i) Randomization

(ii) Replication

(iii) Local control

(i) Randomization

The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor that may affect the experiment. In the development of analysis of variance, we assume that the errors are random and independent. In turn, the observations also become random. The principle of randomization ensures this.

The random assignment of experimental units to treatments results in the following outcomes.

a) It eliminates the systematic bias.

b) It is needed to obtain a representative sample from the population.

c) It helps in distributing the unknown variation due to confounded variables throughout the experiment and breaks the confounding influence.

Randomization forms a basis of valid experiment but replication is also needed for the validity of the experiment.

If the randomization process is such that every experimental unit has an equal chance of receiving each treatment, it is called a complete Randomization.

(ii) Replication:

— In the replication principle, any treatment is repeated a number of times to obtain a valid and more reliable estimate than which is possible with one observation only. Replication provides an efficient way of increasing the precision of an experiment. The precision increases with the increase in the number of observations. Replication provides more observations when the same treatment is used, so it increases precision. For example, if variance of x is σ^2 than variance of sample mean $\bar{x}$ based on n observation is σ^2. So as n increases, $Var(\bar{x})$ decreases.

(ii) Local control (error control)

The replication is used with local control to reduce the experimental error. For example, if the experimental units are divided into different groups such that they are homogeneous within the blocks, than the variation among the blocks is eliminated and ideally the error component will contain the variation due to the treatments only. This will in turn increase the efficiency.

Complete and incomplete block designs:

In most of the experiments, the available experimental units are grouped into blocks having more or less identical characteristics to remove the blocking effect from the experimental error. Such design are termed as **block designs**.

The number of experimental units in a block is called the **block size**.

If size of block = number of treatments and each treatment in each block is randomly allocated, then it is a **full replication** and the design is called as **complete block design**.

In case, the number of treatments is so large that a full replication in each block makes it too heterogeneous with respect to the characteristic under study, then smaller but homogeneous blocks can be used. In such a case, the blocks do not contain a full replicate of the treatments. Experimental designs with blocks containing an incomplete replication of the treatments are called **incomplete block designs**.

Completely randomized design (CRD)

The CRD is the simplest design. Suppose there are *v* treatments to be compared.

- All experimental units are considered the same and no division or grouping among them exist.
- In CRD, the *v* treatments are allocated randomly to the whole set of experimental units, without making any effort to group the experimental units in any way for more homogeneity.
- Design is entirely flexible in the sense that any number of treatments or replications may be used.
- Number of replications for different treatments need not be equal and may vary from treatment to treatment depending on the knowledge (if any) on the variability of the observations on individual treatments as well as on the accuracy required for the estimate of individual treatment effect.

Example: Suppose there are 4 treatments and 20 experimental units, then

- the treatment 1 is replicated, say 3 times and is given to 3 experimental units
- the treatment 2 is replicated, say 5 times and is given to 5 experimental units
- the treatment 3 is replicated, say 6 times and is given to 6 experimental units

and

- finally, the treatment 4 is replicated [20-(6+5+3)=]6 times and is given to the remaining 6 experimental units

- All the variability among the experimental units goes into experimented error.
- CRD is used when the experimental material is homogeneous.
- CRD is often inefficient.
- CRD is more useful when the experiments are conducted inside the lab.
- CRD is well suited for the small number of treatments and for the homogeneous experimental material.

LAYOUT OF CRD

Following steps are needed to design a CRD:

- Divide the entire experimental material or area into a number of experimental units, say *n*.
- Fix the number of replications for different treatments in advance (for given total number of available experimental units).

No local control measure is provided as such except that the error variance can be reduced by choosing a homogeneous set of experimental units.

Procedure

Let the v treatments are numbered from 1,2,...,v and n be the number of replications required for i^{th}

i

treatment such that $\sum_{i=1}^{v} n_i = n$.

- Select n_1 units out of n units randomly and apply treatment 1 to these n_1 units. (**Note**: This is how the randomization principle is utilized is CRD.)
- Select n_2 units out of $(n - n_1)$ units randomly and apply treatment 2 to these n_2 units.
- Continue with this procedure until all the treatments have been utilized.
- Generally equal number of treatments are allocated to all the experimental units unless no practical limitation dictates or some treatments are more variable or/and of more interest.

Analysis

There is only one factor which is affecting the outcome – treatment effect. So the set-up of one-way analysis of variance is to be used.

y_{ij} : Individual measurement of j^{th} experimental units for i^{th} treatment $i = 1,2,...,v$, $j = 1,2,..., n_i$

. y_{ij} : Independently distributed following $N(\mu + \alpha_i, \sigma^2)$ with $\sum_{i}^{v} n_i \alpha_i = 0$.

μ : overall mean

^{ith} treatment effect

$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_v = 0$

H_1 : All $\alpha_i's$ are not equal.

The data set is arranged as follows:

Treatments			
1	2	...	v
y_{11}	y_{21}	...	y_{v1}
y_{12}	y_{22}	...	y_{v2}
y_{1n}	y_{2n}	...	y_{vn}
T_1	T_2	...	T_v

where $T_i = \sum_{j=1}^{n} y_{ij}$ is the treatment total due to i^{th} effect,

Randomized Block Design

If large number of treatments are to be compared, then large number of experimental units are required. This will increase the variation among the responses and CRD may not be appropriate to use. In such a case when the experimental material is not homogeneous and there are *v* treatments to be compared, then it may be possible to

- group the experimental material into blocks of sizes *v* units.
- Blocks are constructed such that the experimental units within a block are relatively homogeneous and resemble to each other more closely than the units in the different blocks.
- If there are *b* such blocks, we say that the blocks are at *b* levels. Similarly if there are *v* treatments, we say that the treatments are at *v* levels. The responses from the *b* levels of blocks and *v* levels of treatments can be arranged in a two-way layout. The observed data set is arranged as follows:

		Blocks					Block Totals
		1	2	i		b	
Treatments	1	y_{11}	y_{21}	y_{i1}		y_{b1}	$B_1 = y_{o1}$
	2	y_{12}	y_{22}	y_{i2}		y_{b2}	$B_2 = y_{o2}$
	j	y_{1j}	y_{2j}	y_{ij}		y_{bj}	$B_j = y_{oj}$
	v	y_{1v}	y_{2v}	y_{iv}		y_{bv}	$B_b = y_{ob}$
Treatment Totals		$T_1 = y_{1o}$	$T_2 = y_{2o}$	$T_i = y_{io}$		y_{vo}	Grand Total G= y_{oo}

Layout:

A two-way layout is called a randomized block design (RBD) or a randomized complete block design (RCB) if within each block, the v treatments are randomly assigned to v experimental units such that each of the $v!$ ways of assigning the treatments to the units has the same probability of being adopted in the experiment and the assignment in different blocks are statistically independent.

The RBD utilizes the principles of design - randomization, replication and local control - in the following way:

1. Randomization:

- Number the v treatments 1,2,…,v.
- Number the units in each block as 1, 2,...,v.
- Randomly allocate the v treatments to v experimental units in each block.

2. Replication

Since each treatment is appearing in the each block, so every treatment will appear in all the blocks. So each treatment can be considered as if replicated the number of times as the number of blocks. Thus in RBD, the number of blocks and the number of replications are same.

3. Local control

Local control is adopted in RBD in following way:

- First form the homogeneous blocks of the xperimental units.
- Then allocate each treatment randomly in each block.

The error variance now will be smaller because of homogeneous blocks and some variance will be parted away from the error variance due to the difference among the blocks.

Example:

Suppose there are 7 treatment denoted as $T_1, T_2, .., T_7$ corresponding to 7 levels of a factor to be included in 4 blocks. So one possible layout of the assignment of 7 treatments to 4 different blocks in a RBD is as follows

Block 1	T_2	T_7	T_3	T_5	T_1	T_4	T_6
Block 2	T_1	T_6	T_7	T_4	T_5	T_3	T_2
Block 3	T_7	T_5	T_1	T_6	T_4	T_2	T_3
Block 4	T_4	T_1	T_5	T_6	T_2	T_7	T_3

Analysis

Let

y_{ij} : Individual measurements of j^{th} treatment in i^{th} block, $i = 1,2,...,b$, $j = 1,2,...,v$.

y_{ij} 's are independently distributed following $N(\mu + \beta_i + \tau_j, \sigma^2)$

where μ : overall mean effect

β_i : i^{th} block effect

τ_j : j^{th} treatment effect

such that $\sum_{i=1}^{b} \beta_i = 0, \sum_{j=1}^{v} \tau_j = 0$.

There are two null hypotheses to be tested.

- related to the block effects

  $H_{0B}: \beta_1 = \beta_2 = ... = \beta_b = 0.$

- related to the treatment effects

  $H_{0T}: \tau_1 = \tau_2 = ... = \tau_v = 0.$

The linear model in this case is a two-way model as

$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$, $i = 1,2,.., b; j = 1,2,.., v$

where ε_{ij} are identically and independently distributed random errors following a normal distribution with mean 0 and variance σ^2 .

The tests of hypothesis can be derived using the likelihood ratio test or the principle of least squares. The use of likelihood ratio test has already been demonstrated earlier, so we now use the principle of least squares.

Minimizing $S = \sum_{i=1}^{b}\sum_{j=1}^{v} \varepsilon_{ij}^2 = \sum_{i=1}^{b}\sum_{j=1}^{v} (y_{ij} - \mu - \beta_i - \tau_j)^2$

and solving the normal equation

$\frac{\partial S}{\partial \mu} = 0, \frac{\partial S}{\partial \beta_i} = 0, \frac{\partial S}{\partial \tau_j} = 0$ for all $i = 1,2,.., b, j = 1,2,.., v$.

The fitted model is

$$y_{ij} = \hat{\mu} + \hat{\beta}_i + \hat{\tau}_j + \hat{\varepsilon}_{ij}$$
$$= \bar{y}_{oo} + (\bar{y}_{io} - \bar{y}_{oo}) + (\bar{y}_{oj} - \bar{y}_{oo}) + (y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo}).$$

Squaring both sides and summing over i and j gives

$$\sum_{i=1}^{b}\sum_{j=1}^{v} (\bar{y}_{ij} - \bar{y}_{oo})^2 = v\sum_{i=1}^{b} (\bar{y}_{io} - \bar{y}_{oo})^2 + b\sum_{j=1}^{v} (\bar{y}_{oj} - \bar{y}_{oo})^2 + \sum_{i=1}^{b}\sum_{j=1}^{v} (y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo})^2$$

or $TSS = SSBl + SSTr + SSE$

with degrees of freedom partitioned as bv

$- 1 = (b - 1) + (v - 1) + (b - 1)(v - 1)$.

The reason for the number of degrees of freedom for different sums of squares is the same as in the case of CRD.

$G = \sum_{i=1}^{b}\sum_{j=1}^{v} y_{ij}$: Grand total of all the observation.

Reject H_{0B} *if* $F_{be} > F_{\alpha}((b-1),(b-1)(v-1)$

Reject H_{0T} *if* $F_{Tr} > F_{\alpha}((v-1),(b-1)(v-1))$

If H_{0B} is accepted, then it indicates that the blocking is not necessary for future experimentation.

If H_{0T} is rejected then it indicates that the treatments are different. Then the multiple comparison tests are used to divide the entire set of treatments into different subgroup such that the treatments in the same subgroup have the same treatment effect and those in the different subgroups have different treatment effects.

The analysis of variance table is as follows

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Blocks	b - 1	*SSBl*	*MSBl*	F_{Bl}
Treatments	v - 1	*SSTr*	*MSTr*	F_{Tr}
Errors	$(b$ - 1$)(v$ - 1$)$	*SSE*	*MSE*	
Total	bv - 1	*TSS*		

Latin Square Design

The treatments in the RBD are randomly assigned to b blocks such that each treatment must occur in each block rather than assigning them at random over the entire set of experimental units as in the CRD. There are only two factors – block and treatment effects – which are taken into account and the total number of experimental units needed for complete replication are bv where b and v are the numbers of blocks and treatments respectively.

If there are three factors and suppose there are b, v and k levels of each factor, then the total number of experimental units needed for a complete replication are bvk. This increases the cost of experimentation and the required number of experimental units over RBD.

In Latin square design (LSD), the experimental material is divided into rows and columns, each having the same number of experimental units which is equal to the number of treatments. The treatments are allocated to the rows and the columns such that each treatment occurs once and only once in the each row and in the each column.

In order to allocate the treatment to the experimental units in rows and columns, we take the help from Latin squares.

Latin Square:

A Latin square of order p is an arrangement of p symbols in p^2 cells arranged in p rows and p columns such that each symbol occurs once and only once in each row and in each column. For example, to write a Latin square of order 4, choose four symbols – A, B, C and D. These letters are Latin letters which are used as symbols. Write them in a way such that each of the letters out of A, B, C and D occurs once and only once is each row and each column. For example, as

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

This is a Latin square.

We consider first the following example to illustrate how a Latin square is used to allocate the treatments and in getting the response.

Example:

Suppose different brands of petrol are to be compared with respect to the mileage per liter achieved in motor cars.

Important factors responsible for the variation in the mileage are

- difference between individual cars.
- difference in the driving habits of drivers.

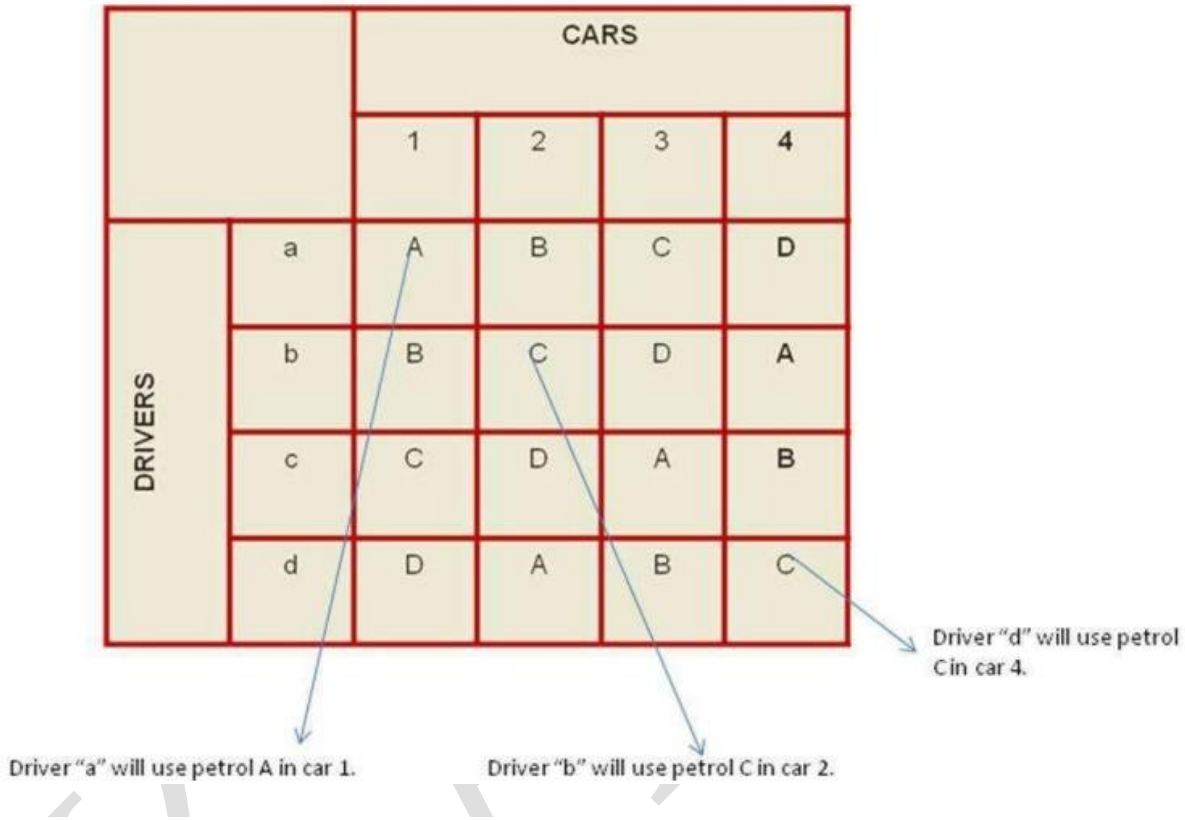
We have three factors – cars, drivers and petrol brands. Suppose we have

- 4 types of cars denoted as 1, 2, 3, 4.
- 4 drivers that are represented by as a, b, c, d.
- 4 brands of petrol are indicated by as A, B, C, D.

Now the complete replication will require 4 × 4 × 4 = 64 number of experiments. We choose only 16 experiments. To choose such 16 experiments, we take the help of Latin square. Suppose we choose the following Latin square:

A B C D
B C D A
C D A B
D A B C

Write them in rows and columns and choose rows for cars, columns for drivers and letter for petrol brands. Thus 16 observations are recorded as per this plan of treatment combination (as shown in the next figure) and further analysis is carried out. Since such design is based on Latin square, so it is called as a Latin square design.



Another choice of a Latin square of order 4 is

C B A D

B C D A

A D C B

D A B C

This will again give a design different from the previous one. The 16 observations will be recorded again but based on different treatment combinations.

Since we use only 16 out of 64 possible observations, so it is an incomplete 3 way layout in which each of the 3 factors – cars, drivers and petrol brands are at 4 levels and the observations are recorded only on 16 of the 64 possible treatment combinations.

Thus in a LSD,

- the treatments are grouped into replication in two-ways
  - once in rows and
  - and in columns,
- rows and columns variations are eliminated from the within treatment variation.
  - In RBD, the experimental units are divided into homogeneous blocks according to the blocking factor. Hence it eliminates the difference among blocks from the experimental error.
  - In LSD, the experimental units are grouped according to two factors. Hence two effects (like as two block effects) are removed from the experimental error.
  - So the error variance can be considerably reduced in LSD.

The LSD is an incomplete three-way layout in which each of the the three factors, viz, rows, columns and treatments, is at v levels each and observations only on v^2 of the v^3 possible treatment combinations are taken. Each treatment combination contains one level of each factor.

The analysis of data in a LSD is conditional in the sense it depends on which Latin square is used for allocating the treatments. If the Latin square changes, the conclusions may also change.

We note that Latin squares play an important role is a LSD, so first we study more about these Latin squares before describing the analysis of variance.

Standard form of Latin square

A Latin square is in the standard form if the symbols in the first row and first columns are in the **natural order** (Natural order means the order of alphabets like A, B, C, D,…).

Given a Latin square, it is possible to rearrange the columns so that the first row and first column remain in natural order.

Example: Four standard forms of 4 ×4 Latin square are as follows.

A B C D	A B C D	A B C D	A B C D
B A D C	B C D A	B D A C	B A D C
C D B A	C D A B	C A D B	C D A B
D C A B	D A B C	D C B A	D C B A

For each standard Latin square of order p, the p rows can be permuted in $p!$ ways. Keeping a row fixed, vary and permute $(p - 1)$ columns in $(p - 1)!$ ways. So there are $p!(p - 1)!$ different Latin squares.

For illustration

Size of square	Number of Standard squares	Value of $p!(1 - p)!$	Total number of different squares
3 x 3	1	12	12
4 x 4	4	144	576
5 x 5	56	2880	161280
6 x 6	9408	86400	812851250

Conjugate:

Two standard Latin squares are called conjugate if the rows of one are the columns of other .

For example

A B C D		A B C D
B C D A	and	B C D A
C D A B		C D A B
D A B C		D A B C

are conjugate. In fact, they are **self conjugate**.

A Latin square is called **self conjugate** if its arrangement in rows and columns are the same.

Transformation set:

A set of all Latin squares obtained from a single Latin square by permuting its rows, columns and symbols is called a transformation set.

From a Latin square of order p, $p!(p-1)!$ different Latin squares can be obtained by making $p!$ permutations of columns and $(p-1)!$ permutations of rows which leaves the first row in place. Thus

Number of different Latin squares of order p in a transformation set = $p!(p-1)!$ X number of standard Latin squares in the set

Orthogonal Latin squares

If two Latin squares of the same order but with different symbols are such that when they are superimposed on each other, every ordered pair of symbols (different) occurs exactly once in the Latin square, then they are called orthogonal.

Greco-Latin square:

A pair of orthogonal Latin squares, one with Latin symbols and the other with Greek symbols forms a Greco-Latin square.

For example

A B C D	$\alpha\ \beta\ \gamma\ \delta$
B A D C	$\delta\ \gamma\ \beta\ \alpha$
C D A B	$\beta\ \alpha\ \delta\ \gamma$
D C B A	$\gamma\ \delta\ \alpha\ \beta$

is a Greco-Latin square of order 4.

Greco Latin squares design enables to consider one more factor than the factors in Latin square design. For example, in the earlier example, if there are four drivers, four cars, four petrol and each petrol has four varieties, as α , β , γ and δ , then Greco-Latin square helps in deciding the treatment combination as follows:

		Cars			
		1	2	3	4
Driver	a	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
	b	$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
	c	$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
	d	$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

Now

$A\alpha$ means: Driver 'a' will use the α variant of petrol A in Car 1.

$B\gamma$ means: Driver 'c' will use the γ variant of petrol B in Car 4

and so on.

Mutually orthogonal Latin square

A set of Latin squares of the same order is called a set of mutually orthogonal Latin square (or a hyper Greco-Latin square) if every pair in the set is orthogonal. The total number of mutually orthogonal Latin squares of order p is at most $(p - 1)$.

Analysis of LSD (one observation per cell)

In designing a LSD of order p,

- choose one Latin square at random from the set of all possible Latin squares of order p.
- Select a standard latin square from the set of all standard Latin squares with equal probability.
- Randomize all the rows and columns as follows:
  - Choose a random number, less than p, say n_1 and then 2^{nd} row is the n_1^{th} row.
  - Choose another random number less than p, say n_2 and then 3^{rd} row is the n_2^{th} row and so on.
  - Then do the same for column.
- For Latin squares of order less than 5, fix first row and then randomize rows and then randomize columns. In Latin squares of order 5 or more, need not to fix even the first row. Just randomize all rows and columns.

Example:

Suppose following Latin square is chosen

A B C D E
B C D E A
D E A B C
E A B C D
C D E A B

Now randomize rows, e.g., 3rd row becomes 5th row and 5th row becomes 3rd row . The Latin square becomes

A B C D E
B C D E A
C D E A B
E A B C D
D E A B C.

Now randomize columns, say 5th column becomes 1st column, 1st column becomes 4th column and 4th column becomes 5th column

E B C A D
A C D B E
D A B E C
C E A D B
B D E C A

Now use this Latin square for the assignment of treatments.

y_{ijk} : Observation on k^{th} treatment in i^{th} row and j^{th} block, $i = 1,2,...,v$, $j = 1,2,...,v$, $k = 1,2,...,v$.

Triplets (I, j, k) take on only the v^2 values indicated by the chosen particular Latin square selected for the experiment.

y_{ijk} 's are independently distributed as $N(\mu + \alpha_i + \beta_j + \tau_k, \sigma^2)$.

Decision rules:

Reject H_{0R} at level α if $F_R > F_{1-\alpha;\,v(-1),(v-1)(v-2)}$

Reject H_{0C} at level α if $F_C > F_{1-\alpha;(v-1),(v-1)(v-2)}$

Reject H_{0T} at level α if $F_T > F_{1-\alpha;(v-1),(v-1)(v-2)}$.

If any null hypothesis is rejected, then use multiple comparison test.

The analysis of variance table is as follows

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F
Rows	v - 1	*SSR*	*MSR*	F_R
Columns	v - 1	*SSC*	*MSC*	F_C
Treatments	v - 1	*SSTr*	*MSTr*	F_T
Error	$(v - 1)(v - 2)$	*SSE*	*MSE*	
Total	$v^2 - 1$	*TSS*		

Missing plot techniques:

It happens many time in conducting the experiments that some observation are missed. This may happen due to several reasons. For example, in a clinical trial, suppose the readings of blood pressure are to be recorded after three days of giving the medicine to the patients. Suppose the medicine is given to 20 patients and one of the patient doesn't turn up for providing the blood pressure reading. Similarly, in an agricultural experiment, the seeds are sown and yields are to be recorded after few months. Suppose some cattle destroys the crop of any plot or the crop of any plot is destroyed due to storm, insects etc.

In such cases, one option is to
- somehow estimate the missing value on the basis of available data,
- replace it back in the data and make the data set complete.

Now conduct the statistical analysis on the basis of completed data set as if no value was missing by making necessary adjustments in the statistical tools to be applied. Such an area comes under the purview of "missing data models" and lot of development has taken place. Several books on this issue have appeared, e.g.

- Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley.
- Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, London etc.

We discuss here the classical missing plot technique proposed by Yates which involve the following steps:

- Estimate the missing observations by the values which makes the error sum of squares to be minimum.
- Substitute the unknown values by the missing observations.
- Express the error sum of squares as a function of these unknown values.
- Minimize the error sum of squares using principle of maxima/minima, i.e., differentiating it with respect to the missing value and put it to zero and form a linear equation.
- Form as many linear equation as the number of unknown values (i.e., differentiate error sum of squares with respect to each unknown value).
- Solve all the linear equations simultaneously and solutions will provide the missing values.
- Impute the missing values with the estimated values and complete the data.

- Apply analysis of variance tools.
- The error sum of squares thus obtained is corrected but treatment sum of squares are not corrected.
- The number of degrees of freedom associated with the total sum of squares are subtracted by the number of missing values and adjusted in the error sum of squares. No change in the degrees of freedom of sum of squares due to treatment is needed.

Missing observations in RBD

One missing observation:

Suppose one observation in $(i, j)^{th}$ cell is missing and let this be x. The arrangement of observations in RBD then will look like as follows:

	Blocks							Block Total
		1	2		i		b	
Treatments	1	y_{11}	y_{21}	…	y_{i1}	…	y_{b1}	$B_1 = y_{o1}$
	2	y_{12}	y_{22}	…	y_{i2}	…	y_{b2}	$B_2 = y_{o2}$
	.	.	.		.		.	.
	.	.	.		.		.	.
	.	.	.		.		.	.
	j	y_{1j}	y_{2j}	…	$y_{ij} = x$	…	y_{bj}	$B_j = y'_{oj} + x$
	.		.		.		.	.
	.		.		.		.	.
	.		.		.		.	.
	v	y_{1v}	y_{2v}	…	y_{iv}	…	y_{bv}	$B_b = y_{ob}$
Treatment Totals		$T_1 = y_{1o}$	$T_2 = y_{2o}$		$T_i = y'_{io} + x$		y_{vo}	Grand Total $G = y'_{oo} + x$

Two missing observations:

If there are two missing observation, then let they be x and y.

- Let the corresponding row sums (block totals) are $(R_1 + x)$ and $(R_2 + y)$.
- Column sums (treatment totals) are $(C_1 + x)$ and $(C_2 + y)$.
- Total of known observations is S.

Then

$$SSE = x^2 + y^2 - \frac{1}{b}[(R_1 + x)^2 + (R_2 + y)^2] - \frac{1}{v}[(C_1 + x)^2 + (C_2 + y)^2] + \frac{1}{bv}(S + x + y)^2$$
$$+ \text{terms independent of } x \text{ and } y.$$

Now differentiate SSE with respect to x and y, as

$$\frac{\partial(SSE)}{} \quad \frac{R}{} \quad \frac{+x}{} \quad \frac{C}{} \quad \frac{+x}{} \quad \frac{S+}{} \quad \frac{x+y}{}$$

Adjustments to be done in analysis of variance

(i) Obtain the within block sum of squares from incomplete data.

(ii) Subtract correct error sum of squares from (i) . This given the correct treatment sum of squares.

(iii) Reduce the degrees of freedom of error sum of squares by the number of missing observations.

(iv) No adjustments in other sum of squares are required.

Missing observations in LSD

Let

- x be the missing observation in $(i, j, k)^{th}$ cell,

  i.e., y_{ijk} , $i = 1,2,.., v$, $j = 1,2,.., v$, $k = 1,2,..,v$.
- R: Total of known observations in i^{th} row
- C: Total of known observations in j^{th} column
- T: Total of known observation receiving the k^{th} treatment.
- S: Total of known observations

Now

Correction factor $(CF) = \frac{(S + x)_2}{v2}$

Total sum of squares $(TSS) = x^2$ + term which are constant with respect to x - CF

Row sum of squares $(SSR) = \frac{(R + x)_2}{v}$ + term which are constant with respect to x - CF

Column sum of squares $(SSC) = \frac{(C + x)_2}{v}$ + term which are constant with respect to x - CF

Treatment sum of squares$(SSTr) = \frac{(T + x)^2}{v}$ + term which are constant with respect to x - CF

Sum of squares due to error $(SSE) = TSS - SSR - SSC - SSTr$

Adjustment to be done in analysis of variance:

Do all the steps as in the case of RBD.

To get the correct treatment sum of squares, proceed as follows:

- Ignore the treatment classification and consider only row and column classification .
- Substitute the estimated values at the place of missing observation.

- Obtain the error sum of squares from complete data, say SSE_1 .
- Let SSE_2 be the error sum of squares based on LSD obtained earlier.
- Find corrected treatment sum of squares = $SSE_2 - SSE_1$.
- Reduce of degrees of freedom of error sum of squares by the number of missing values.

POSSIBLE QUESTIONS

PART-B (SIX MARKS)

1.Write the procedure for analyzing an experimental design using RBD and how RBD is differs from CRD.

2.Explain the missing plot technique. And how to estimate the missing value (plot) in LSD when one value is missing?

3.Explain the steps involved in a systematic approach to the planning and implementation of experiments.

4.Describe the procedure for analyzing an experimental design using orthogonal form of LCD.

5.Describe the five different broad categories of experimental problems in brief.

6.How layout can be made for design LSD and write the procedure for analyzing an experimental design using LSD.

7.What are three fundamental principles in experimental design? Explain them in detail.

8.How layout can be made for design a CRD and write the procedure for analyzing an experimental design using CRD.

9.What are three fundamental principles in experimental design? Explain them in detail.

PART-C (TEN MARKS)

1.How layout can be made for design a CRD and write the procedure for analyzing an experimental design using CRD.

UNIT – V

Statistical Quality Control (SQC): Statistical basis for control charts, control limits. Control charts for variables - $\overline{X}$, R charts. Charts for defectives – p and np charts. Chart for defects – C chart. Acceptance Sampling – single and double sampling plans.

Introduction

What is quality?

"The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" -- ASQC

- User-based:
- Product-based:
- Manufacturing-based:

Why quality is important?

- Costs and market share
- Company's reputation
- Product liability
- International implications

Two parts of quality management

- Quality Control:
- Quality Assurance:

Traditional concept of quality management:

- Responsibility of quality control dept. only
- Rely on the inspection process
- Satisfied with meeting specifications

I. Statistical Process Control

- graphical presentation of samples of process output over time

- used to monitoring (production) process and detect quality problems

Natural and Assignable Variations

	Natural	Assignable
Charact.:		
Causes:		
Action:		

Idea Behind Control Charts:

If (production) process is normal

→ only natural variations exist

→ samples of output is Normally distributed

→ within 3 std. 99.7% of time

Therefore,

→ If not within 3 std. ==> assignable variations exist!

UCL and LCL are set to correspond to the 3 std. lines

Procedures of using control charts:

1. take samples regularly
2. calculate proper statistics
3. plot the statistics on the control chart
4. Analyze the pattern of plots and draw conclusion

In control and out of control

- In control: plots Normally distributed, unbiased, no patterns
  → indicating no assignable variations exist

- Out of control:
  - one plot outside UCL or LCL (for all charts)
  - 2 of 3 consecutive plots out of 2 std. Line (for X-chart)
  - 7 consecutive plots on one side (for X-chart)

  → indicating assignable variations exist, sign of quality problems.

Types of control chart:

- Variable Charts: for continuous quality measure
  - X-chart: process average
  - R-chart: process dispersion and variation

- Attribute Charts: for attribute quality measure
  - p-chart: defective rate
  - c-chart: number of defectives

V. Construct and Use Control Charts

Construct X-chart

1. based on some process information:

- CL = specification (target value)
- UCL = CL + 3 std of process /√n
- LCL = CL - 3 std of process /√n

- n: sample size

2. based only on past samples

- CL = average of past samples
- UCL = CL + A_2*range average of past samples
- LCL = CL - A_2*range average of past samples

- A_2: found from the table of your textbook

3. Differences between 1. and 2.

Construct R-chart (based on past samples)

- CL = average range of past samples
- UCL = D_4*average range of past samples
- LCL = D_3*average range of past samples > 0

- D_4 and D_3: found from the table of your textbook

Example 1

Samples taken from a process for making aluminum rods have an average of 2cm. The sample size is 16. The process variability is approximately normal and has a std. of 0.1cm. Design an X-chart for this process control.

Example 2

Five samples of drop-forged steel handles, with four observations in each sample, have been taken. The weight of each handle in the samples is given below (in ounces). Use the sample data to construct an X-chart and an R-chart to monitor the future process.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
10.2	10.3	9.7	9.9	9.8
9.9	9.8	9.9	10.3	10.2
9.8	9.9	9.9	10.1	10.3
10.1	10.4	10.1	10.5	9.7

Use X-chart and R-chart

- Calculate averages and ranges of new samples
- Plot on the X-chart and R-chart, respectively

Example 2 (continued)

Five more samples of the handles are taken

Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
10.4	10.5	9.9	10.3	9.9
9.8	9.9	9.9	10.4	10.4
9.9	9.9	9.9	10.6	10.5
10.3	10.5	10.3	10.5	9.9

Is the process in control (changed)?

Construct p-chart

- CL = average defective rate of past samples
- UCL = CL + 3 std of past sample defective rates
- LCL = CL - 3 std of past sample defective rates > 0

Use p-chart

- Calculate defective rates of new samples
- Plot on the p-chart

Example 3

A good quality lawnmower is supposed to start at the first try. In the third quarter, 50 craftsman lawnmowers are started every day and an average of 4 did not start. In the fourth quarter, the number of lawnmower did not start (out of 50) in the first 6 days are 4, 5, 4, 6, 7, 6, respectively. Was the quality of lawnmower changed in the fourth quarter?

Construct c-chart

- CL = average # of defectives in past products
- UCL = CL + 3 std of # of defectives in past products
- LCL = CL - 3 std of # of defectives in past products > 0

Use c-chart

- Count # of defective in new products
- Plot on the c-chart

Example 4

There have been complaints that the sports page of the Dubuque Register has lots of typos. The last 6 days have been examined carefully, and the number of typos/page is recorded below. Is the process in control?

Day	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Typos	2	1	5	3	4	0

VI. Acceptance Sampling

Accept or reject a lot (input components or finished products) based on inspection of a sample of products in the lot

Role of Inspection

- Involved in all stages of production process
- Inspection itself does not improve quality
- Destructive and nondestructive inspection

Why sampling instead of 100% inspection?

- Destructive test
- Worker's morale
- Cost consideration

Single Acceptance Sampling Plan:

1. Take a sample of size n from a lot with size N
2. Inspect the sample 100%
3. If # of defective > c, reject the whole lot; otherwise, accept it.

- need to determine n and c.

Operating Characteristic (OC) Curves

To evaluate how well a single acceptance sampling plan discriminates between good and bad lots.

Draw OC curve approximately for a sampling plan with n and c

Idea:

The number of defectives in a sample of size n with defective rate p follows a Poisson distribution approximately with parameter $\lambda = np$, when p is small, n is large, and N is much larger.

→ P(acceptance) = Prob(# def. <= c)

 Prob(# def. <= c, λ) based on Poisson distribution

Procedure:

1. Create a series of p = 1% to 10%.
2. Calculate $\lambda = np$ for each p.
3. Use the Poisson table of Appendix B to find P (acceptance) for each λ and c.
4. Link P(acceptance) to form a curve.

Example 5:

A single sampling plan with n=100 and c=3 is used to inspect a shipment of 10000 computer memory chips. Draw the OC curve for the sampling plan.

P(%)	1	2	3	4	5	6	7	8	9	10
$\lambda = np$										
P(acceptance)										

Concepts related to the OC Curve

- AQL: Acceptable quality level, the defective rate that a consumer is happy to accept (considers as a good lot)
- LTPD: Lot tolerance percent defective, the maximum defective rate that a consumer is willing to accept
- Consumer's risk: the probability that a lot containing defective rate exceeding the LTPD will be accepted.
- Producer's risk: the probability that a lot containing the AQL will be rejected.

Example 5 continued:

The buyer of the memory chip requires that the consumer's risk is limited to 5% at LTPD = 8%. The producer requires that the producer's risk is no more than 5% at AQL = 2%. Does the single sampling plan meet both consumer and producer's requirements?

Sensitivity of OC curve, consumer's risk, and producer's risk to N, n, c.

- Changing n, keeping c constant:
- Changing c, keeping n constant:
- Changing both n and c, keeping c/n constant:
- Changing N:

Average Outgoing Quality (AOQ)

The quality after inspection (by a single sampling plan), measured in defective rate, assuming all defectives in the rejected lot are replaced

$AOQ = p\ P_a(N-n)/N \approx p\ P_a$

P_a = P(acceptance for a lot with defective rate p) can be found from the OC curve

Example 5 continued:

The average defective rate of the memory chip is about 5% (based on the past data). Calculate the AOQ of the memory chip after it is inspected by the sampling plan in Example 5.

Other Sampling Plans

- Double sampling plan

- Given n: sample size

c_1: acceptable level of the first sample

c_2: acceptable level of both samples

Procedure:

Example:

n = 100, $c_1 = 4$, $c_2 = 7$, # of defective in the first sample = 5.

- Sequential sampling plan

- Given n: sample size

upper and lower limits of number of defectives allowed

Procedure:

1. Count # of total defectives found in all previous samples
2. If # of defectives > upper boundary, reject the lot
3. If # of defectives <= lower boundary, accept the lot
4. Otherwise, take a new sample and repeat.

Advantages of double and sequential samplings:

- Psychologically:
- Cost: less inspection for the same accuracy

Factors for Computing Control Chart Limits for X and R Charts

Sample Size (n)	Mean Factor (A2)	Upper Range (D4)	Lower Range (D3)
2	1.880	3.268	0
3	1.023	2.574	0
4	0.729	2.282	0
5	0.577	2.115	0
6	0.483	2.004	0
7	0.419	1.924	0.076
8	0.373	1.864	0.136
9	0.337	1.816	0.184
10	0.308	1.777	0.223

Summary on Statistical Quality Control

This focuses on the three major topics of traditional statistical tools: calculating means and ranges, Statistical Process Control (SPC) and acceptance sampling. Sources of variation in a process are discussed. The SPC section includes discussions about the usage and types of control charts for attributes and variables, along with process capability. Finally, acceptance sampling is the use of sampling to determine whether to accept or reject a lot.

Questions and Answers

1. *Explain the three categories of statistical quality control (SQC). How are they different, what different information do they provide, and how can they be used together?*

The three categories of SQC are traditional statistical tools, acceptance sampling and statistical process control (SPC). Traditional statistical tools are descriptive statistics, such as the mean and range, used to describe quality characteristics. Acceptance sampling is a process of taking a random sample or portion of a batch and deciding whether to accept or reject the whole batch or lot. SPC is a process that uses samples to determine whether a process is functioning normally or not. Traditional statistical tools describe the quality characteristics, but do not tell us whether quality is good or bad. Acceptance sampling tells us whether an entire batch or lot produced should be accepted or rejected after the goods have been produced, while SPC tracks the process over time to ensure it is functioning properly. These tools can be used together effectively. We use the traditional statistical tools as inputs into SPC, which is updated frequently enough to ensure that quality problems are caught in a timely manner. Finally, after a batch has been produced, we use acceptance sampling to determine whether or not the batch can be sold to the customer.

2. *Describe three recent situations in which you were directly affected by poor product or service quality.*

I purchased a bag of flour that I did not open right away. I placed in it the kitchen cabinet. Weeks later, I found bugs in many food items in the cabinet. When I examined the bag of flour, I found dead bugs in the glued seal and inside the bag. Because of the infestation, I had to throw out a number of food items from the cabinet. I also had to treat the kitchen with a compound that would get rid of the remaining bugs.

I did not notice that the bag of sliced beef was expired when I purchased it from a grocery store. A few days later, I opened it and ate some beef. Within an hour, I became very sick. I then looked at the package only to determine that it had expired 3 weeks earlier! I was very upset. I do accept some responsibility in that I did not always check expiration dates on items I purchased. However, that package should not have been available for sale. I now

check expiration dates every time. In addition, when I went to another facility of the same grocery store, I found that every package of sliced beef for sale was expired by at least a week. So I took all of the packages of beef up to the help desk to give them to a manager and complain. He apologized saying that he would deal with the problem.

My mother came to visit me to help me unpack from my move. Her luggage did not arrive with her flight that arrived late on the evening of August 1. She was assigned a file reference number by the airline baggage service after standing in line for at least 15 minutes. I started calling the baggage call center the next evening since we had not heard from the airline nor received the luggage. Our frustration was further compounded by the fact that every time I called the 1-800 number, it was busy. And believe me when I say that I tried many times. The following day (August 3rd), I decided to try the flight reservation number in hopes of speaking with someone. I ended up speaking to someone from that area many times. They would try calling the contracted baggage delivery service after they found out that the luggage had been turned over to this service by the airline. They were not able to get an answer from them. They gave me their phone number as well. When I was able to speak to someone there, I would get different answers, such as the need to find more information or that the luggage had already been delivered. I kept calling both phone numbers on the next day as well in hopes of getting more information. Finally, on August 5th, the luggage was delivered after midnight. My mother did not have her luggage for four full days of a six-day trip. To make matters worse, we had to deal with the frustration of not knowing what was going on and of continuing need to spend time trying to gather information.

3. *Discuss the key differences between common and assignable causes of variation. Give examples.*

Common causes of variation are random, which means that there is not a specific reason for the variation, such as a malfunctioning machine. If you look at 2-liter bottles of a soft drink on a shelf at a retailer, you will notice that they are not filled to the exact same level. That is to be expected, since a machine cannot fill each one to exactly 2 liters every time. Assignable causes of variation are not random. Some examples of assignable causes are a machine in need of repair and defective raw materials from our supplier.

4. *Describe a quality control chart and how it can be used. What are the upper and lower control limits? What does it mean if an observation falls outside the control limits?*

Samples of product or services are plotted on the control chart over time. We then interpret the chart to determine whether the variation in the process is normal or abnormal. We need to use the chart because most products and services exhibit some variation. If the variation is normal, the process is assumed to be "in control." Therefore, we do not need to fix the process.

If the variation is abnormal, this process is assumed to be "out of control." Therefore, we need to determine what is causing this variation, such as a malfunctioning machine or untrained worker. Control charts do not tell us how many defects are being produced by the process. The charts tell us whether it is likely that the process has changed. If the process has changed or is out of control, then it is likely that we are producing defects.

A control chart is a diagram with a center line, an upper control limit and a lower control limit. The upper control limit is normally equal to the mean of the sample data plus three standard deviations. The lower control limit is normally equal to the mean minus three standard deviations. If an observation falls above the upper control limit or below the lower control limit, the process is assumed to be out of control. This is because the chart is based on the normal distribution, which states that 99.7% of the plots will fall within three standard deviations of the mean. Since it is highly unlikely that a plot will fall outside three standards deviations of the mean, the process is likely to be out of control.

5. *Explain the differences between x-bar and R-charts. How can they be used together and why would it be important to use them together?*

The x-bar chart is used to detect variations in the mean of the process, while the R-chart is used to detect changes in the variability of the process. The x-bar and R-charts are used when the data is a variable, meaning that we can collect data using decimal points, such as 16.5 ounces. Examples of variables are weight, height and temperature.

The x-bar and R-charts should be used together. Think about preparing a Thanksgiving turkey in the oven. What can go wrong with the temperature of the oven if it is set at 350 degrees? The average temperature during cooking could be 250 degrees instead. On the other hand, the temperature could average 350 degrees, but actually fluctuate during cooking time between 200 and 500 degrees. Either way, the turkey will not be properly cooked in the oven. The inaccurate average temperature would have been detected by the x-bar chart. The changes in the temperature would have been detected by the R-chart. We use these charts together by plotting the average of the sample on the x-bar chart and the range (high temperature in the sample minus the low temperature) on the R-chart. Then, we first interpret the R-chart. If it is out of control, then the process variation is out of control. The next step would be to investigate the cause of this problem. There is no need to interpret the x-bar chart if the R-chart is out of control. If the variation is out of control, it is not possible to make conclusions about the average because the variation would probably change the average. If the R-chart is in control, then we interpret the x-bar chart. If it is out of control, then the process average is out of control.

6. *Explain the use of p-charts and c-charts. When would you use one rather than the other? Give examples of measurements for both p-charts and c-charts.*

The p-charts and c-charts are both used when the data is an attribute. Data is an attribute when we ask a yes or no question or count the number of defects. For example, is the bottle of Coca-Cola full or not? How many of the Hershey's kisses in the bag are not covered in foil? The p-chart is used to determine whether the proportion of defective units in a sample is in control or not. The c-chart is used to determine whether the number of defects on each item is in control or not. The key difference is that the sample size for the c-chart is always one. In other words, each plot represents the number of defects on one item, such as the number of spelling errors in a report.

7. *Explain what is meant by process capability. Why is it important? What does it tell us? How can it be measured?*

Management or regulations set acceptable levels of variation in order to determine if a product is defective or not. For example, a product is not defective if it is filled to 16 ounces plus or minus one ounce. For this product, the upper specification limit would be 17 ounces, while the lower specification limit would be 15 ounces. Process capability tells us whether or not the process itself is capable of manufacturing product that has a high probability of falling within the specification limits (is not defective). It is important for a company to produce quality products. Process capability is measured by comparing the specifications to the actual variation in the process. The process capability index is the width of the specifications divided by the width of the process variation. If the process capability index is less than one, then the process is not capable of producing within specifications. The higher the index, the more capable the process. The index can be used to determine how many defects are produced on average.

8. *Describe the process of acceptance sampling. What types of sampling plans are there? What is acceptance sampling used for?*

In acceptance sampling, we determine if a set number of items, such as fifteen out of a batch of 100, are defective or not. Then we compare the number of defects to a preset maximum number of acceptable defects (c) to determine whether to accept or reject the whole batch. If the number of actual defects is less than or equal to the preset number, then the entire batch is accepted. Otherwise, it is rejected. Acceptance sampling is used to determine whether the level of quality is acceptable or not. The three types of sampling plans are single sampling, double sampling and multiple sampling.

9. *Describe the concept of six-sigma quality. Why is such a high quality level important?*

It is referred to as six sigma because the specification limits are six sigma away from the mean. We know that the normal process variation is three sigma away from the mean. So, if we divide the width of the specification limits, which is six sigma plus six sigma, divided by the width of the process variation, which is three sigma plus three sigma, then the process capability index is calculated as 2. Therefore, a process with a six-sigma quality level has a process capability index of 2. This means that the process will make about 3 defective items for every million produced if the process is centered on the mean. A process is centered on the mean when the process mean and the desired mean are equal. If they are not equal, then the process has shifted to the left or right, thus causing more defects in the tail of the distribution. An example will clarify this issue. Assume that the following data was collected from the process of filling jars of spaghetti sauce that has a six-sigma quality level when the process is centered:

- Process mean = 16.5 ounces
- Desired mean = 16 ounces (what is printed on the jar)
- Process standard deviation = 0.5 ounces
- Upper specification limit = 17 ounces
- Lower specification limit = 15 ounces

This process would produce jars that are filled above 17 ounces since the process mean is greater than the desired mean. (It might be helpful to show this information on a graph.)

This level of quality is important because customers demand a high level of quality. In the past, parts per thousand defective was the measure used by companies. Now the measure is parts per million (ppm) defective.

THE SEVEN TOOL OF QUALITY

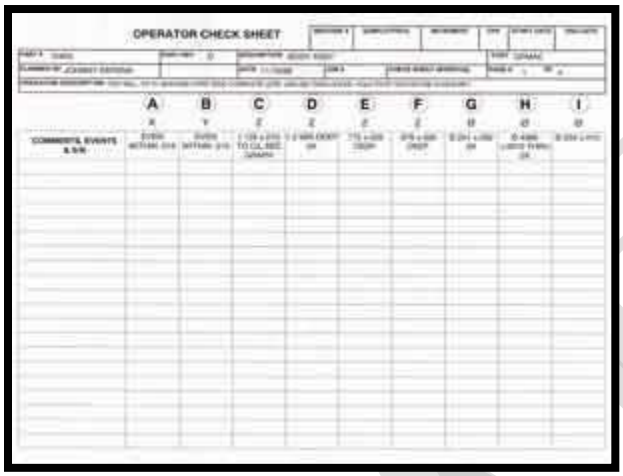
I. Seven quality control tools

There is general agreement on the utility of seven quality control tools. Two of fathers of total quality management (Deming and Ishikawa) recommend that the seven tools be utilized to ensure that all attempts at process improvement include:

- Discovery
- Analysis
- Improvement
- Monitoring
- Implementation
- Verification

1. Checksheet

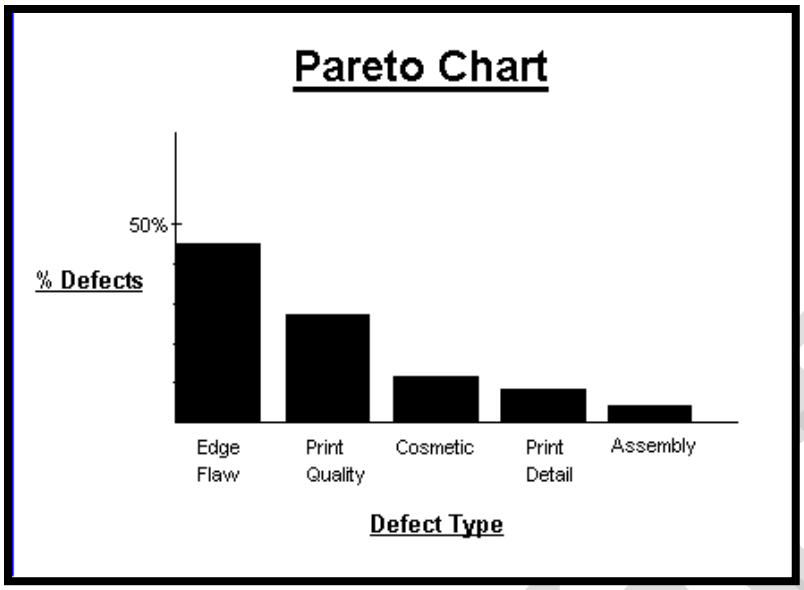
The function of a checksheet is to present information in an efficient, graphical format. This may be accomplished with a simple listing of items. However, the utility of the checksheet may be significantly enhanced, in some instances, by incorporating a depiction of the system under analysis into the form.



2. Pareto chart

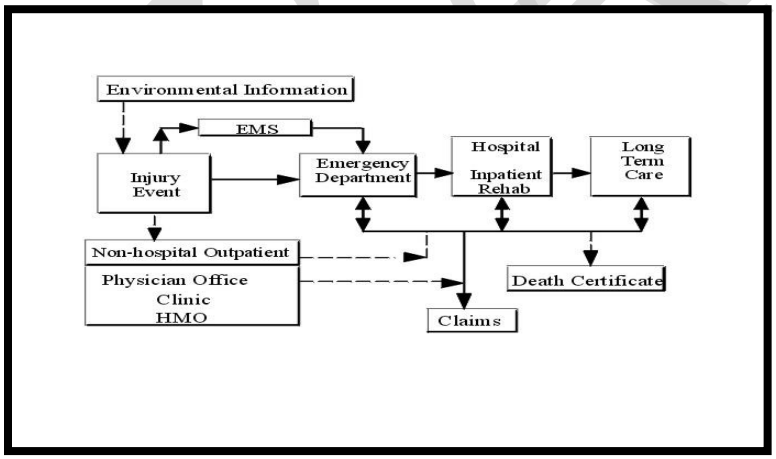
Pareto charts can be used to identify those factors that have the greatest cumulative effect on the system, and thus screen out the less significant factors in an analysis. Ideally, this allows the user to focus attention on a few important factors in a process.

They are created by plotting the cumulative frequencies of the relative frequency data (event count data), in descending order. When this is done, the most essential factors for the analysis are graphically apparent, and in an orderly format.



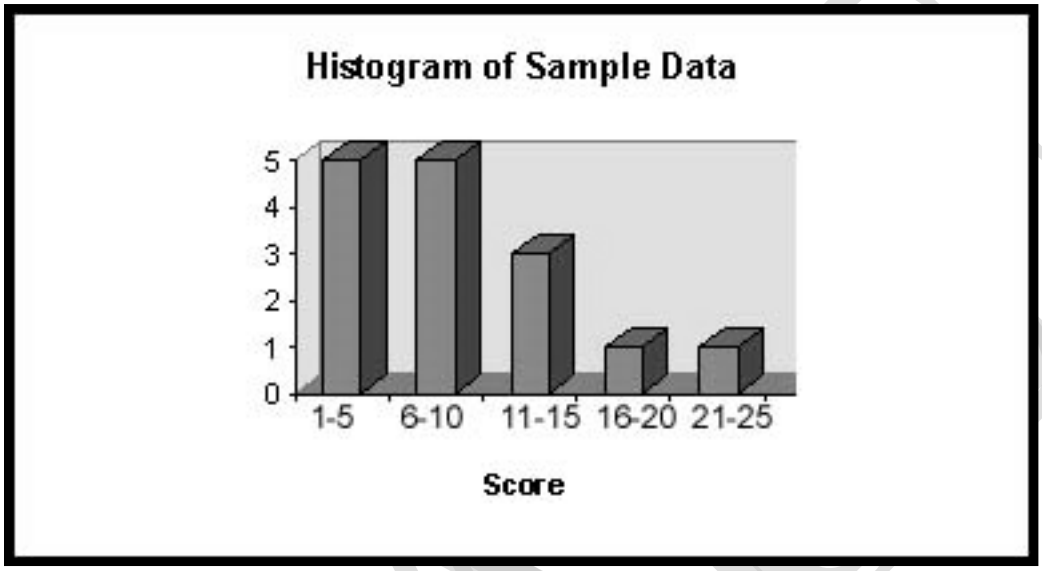
3. Flowchart

Flowcharts are pictorial representations of a process. By breaking the process down into its constituent steps, flowcharts can be useful in identifying where errors are likely to be found in the system.



4. Histogram

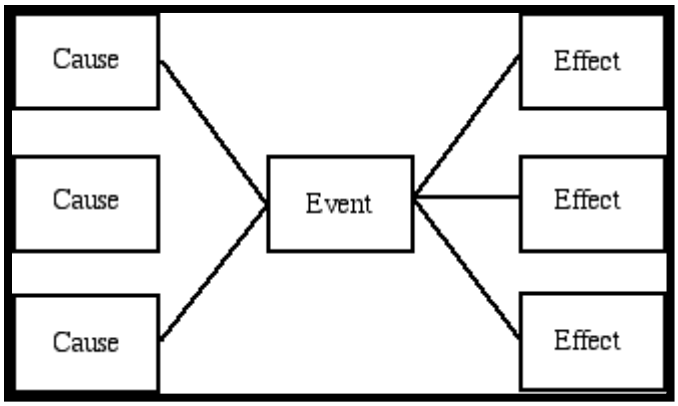
Histograms provide a simple, graphical view of accumulated data, including its dispersion and central tendency. In addition to the ease with which they can be constructed, histograms provide the easiest way to evaluate the distribution of data.



5. Cause and effect diagram

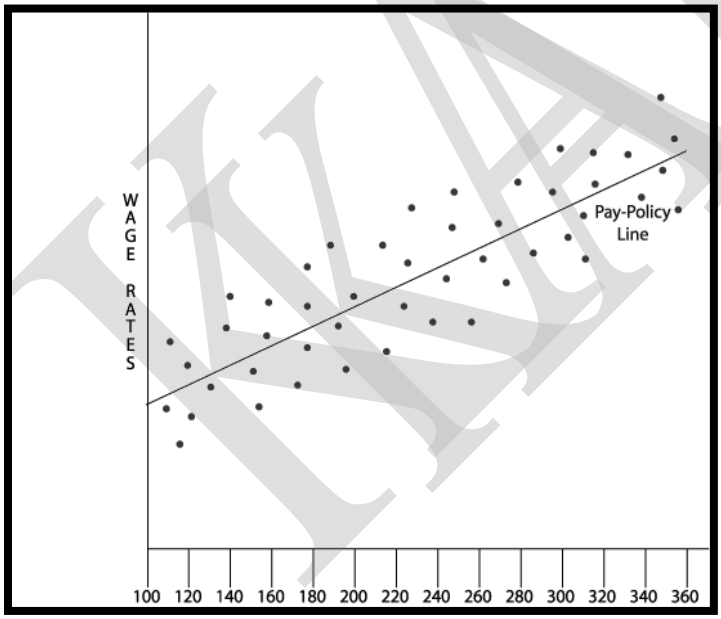
This tool, also called an Ishikawa diagram (or fish bone diagram), is used to associate multiple possible causes with a single effect. Thus, given a particular effect, the diagram is constructed to identify and organize possible causes for it.

The primary branch represents the effect (the quality characteristic that is intended to be improved and controlled) and is typically labelled on the right side of the diagram. Each major branch of the diagram corresponds to a major cause (or class of causes) that directly relates to the effect. Minor branches correspond to more detailed causal factors. This type of diagram is useful in any analysis, as it illustrates the relationship between cause and effect in a rational manner.



6. Scatter diagram

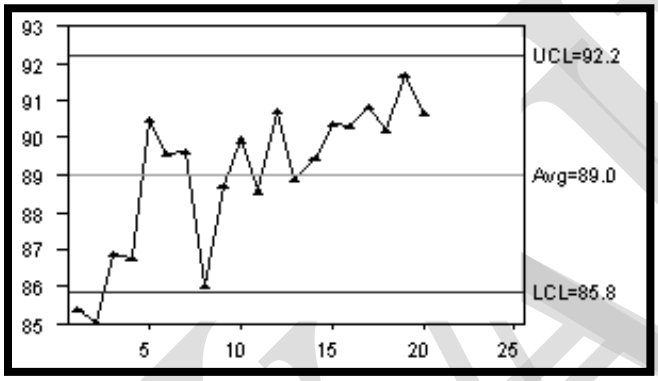
Scatter diagrams are graphical tools that attempt to depict the influence that one variable has on another. A common diagram of this type usually displays points representing the observed value of one variable corresponding to the value of another variable



7. Control chart

The control chart is the fundamental tool of statistical process control, as it indicates the range of variability that is built into a system. Thus, it helps determine whether or not a process is operating consistently or if a special cause has occurred to change the process mean or variance.

The bounds of the control chart are marked by upper and lower control limits that are calculated by applying statistical formulas to data from the process. Data points that fall outside these bounds represent variations due to special causes, which can typically be found and eliminated. On the other hand, improvements in common cause variation require fundamental changes in the process.



II. The New Seven Tools of Quality

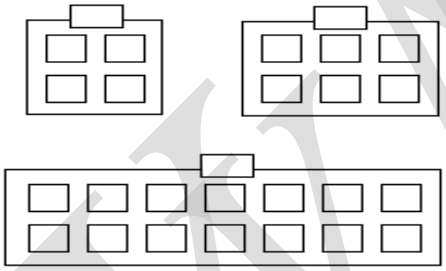
1. The Affinity Diagram

This Affinity Diagram is a very useful tool to use when brainstorming is the main goal. This is an especially effective tool because it allows the participants to be both creative and logical.

By generating ideas the participants are tapping into their creative side, and organizing those ideas allows them to exercise logic. There are 3 main instances when an affinity diagram is especially useful. First, when the problem is complex or hard to understand. Second, when the problem is very large and could appear to be overwhelming. Last, when support and involvement of another group is required. There are six basic steps to creating an affinity diagram:

1. Identify the problem or issue
2. Each person writes issues related to problem on note card or sticky notes
3. Organize the cards or sticky notes into logical piles
4. Name each pile with a header
5. Draw an affinity diagram
6. Discuss the piles created

Below is an example of an affinity diagram:

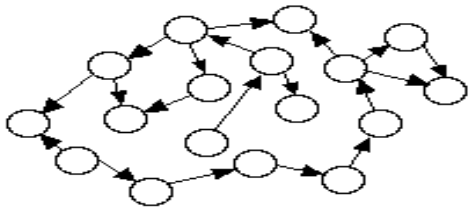


2. The Interrelationship Digraph

The main purpose of the interrelationship digraph is to depict the relationships between different issues. Often times this digraph is used in conjunction with the affinity diagram. It can be very powerful in that it reveals the impact one issue can have on other issues. There are seven steps to create an interrelationship diagram:

1. Identify problem or issue
2. Write each element that relates to problem in a box
3. Draw arrows from the element that influences to the element that is influenced
4. Draw the strongest influence if two elements impact each other
5. Count the arrows
6. Elements with the most outgoing arrows will be root causes or drivers
7. The elements with the most incoming arrows will be key outcomes or results

Below is an example of an interrelationship digraph:

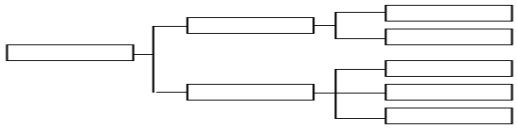


3. Tree Diagram

A tree diagram is often used to discover the steps needed to solve a given problem. It always the user(s) to gain further insight into the problem and helps the team focus on specific tasks to complement the tasks at hand to solve the problem. There are five major steps in creating a tree diagram. They are:

1. Determine the main goal
2. Be concise
3. Brainstorm the main tasks involved in solving the problem and add them to the tree
4. Brainstorm subtask that can also be added to the tree
5. Do this until all possibilities have been exhausted

Below is an example of a tree diagram:

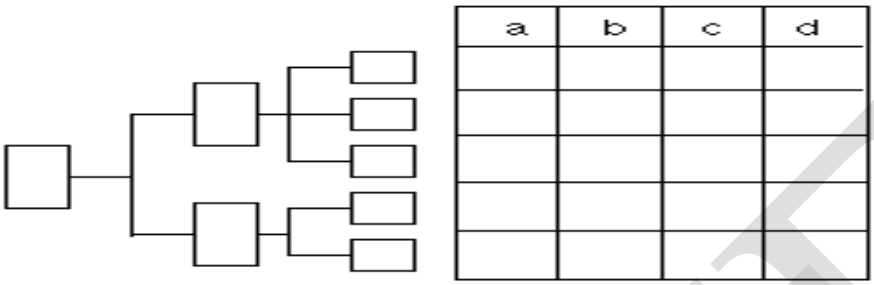


4. Prioritization Grid

A prioritization grid is typically used to make decisions that require analysis of several criteria. These situations could have several options that need to be compared and several criteria that need to be considered. There are eight steps to develop a prioritization grid. They are:

1. Identify your goal
2. Rank the criteria in order from least important to most important
3. Assign each criterion a weight for each option, and be sure the sum of all weights equals one
4. Sum the individual rating for each criterion to come to an overall ranking. Divide by the number of options to find an average ranking.
5. Rank order each option with respect to the criteria. Average the rankings and apply a completed ranking
6. Multiply the criteria weight by its associated criterion rank for each criterion in the matrix. The result in each cell of the matrix is called an importance score
7. Sum the importance scores for each alternative
8. Rank the alternatives in order of importance

Below is an example of a prioritization grid:



5. Matrix Diagram

The matrix diagram is a good tool to use to compare the efficiency and effectiveness of alternatives based on the relationship between two criteria. It uses criteria and symbols to visually depict the relationship between For example, a user could analyze the relationship between cost and performance. Matrix diagrams can be used with up to four dimensions. There are several styles of matrix diagrams. The most common styles are the L-shape, the T- shape, and the Y-shape. There are five steps in constructing a matrix diagram.

1. Decide the factors that are the most important to make the decision
2. Select the style of matrix that will help the best
3. Select the symbols to be used to represent the relationships
4. Complete the matrix using the determined factors and symbols
5. Analyze the completed matrix

Below is an example of a matrix diagram:

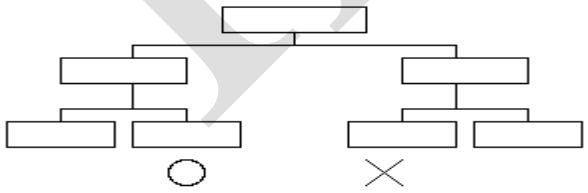
a	b	c	d

6. Process Decision Program Chart

The Process Decision Program Chart is a good tool to use for contingency planning. It helps to realize what could go wrong or problems associated with the implementation of programs and improvements. There are four main steps to creating a Process Decision Program Chart. They are:

1. List the steps in the process you wish to analyze
2. List what could go wrong at each step
3. List the counter measures to the problems
4. Evaluate the counter measures by placing an O for feasible or an X for not feasible

Below is an example of a Process Decision Program Chart

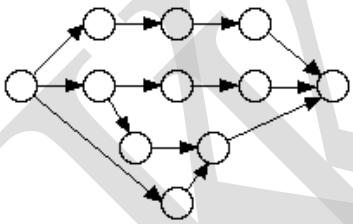


7. Activity Network Diagrams

The Activity Network Diagram is also known as the PERT diagram or the critical path diagram. It evaluates the time it takes from the beginning of a process to the end of a process and determines where there is slack time, and what parts of the process can be improved in relation to time. There are several steps in developing an Activity Network Diagram. They are:

1. List all tasks
2. Determine the time it takes for each task
3. For each task, determine the task that must happen before the current task can take place
4. Draw the network diagram
5. Compute early start and early finish times for each task
6. Compute the late start and late finish times for each task
7. Compute slack time
8. Determine the critical path

Below is an example of an Activity Network Diagram:



Where to find more information regarding the new seven tools of quality:

Foster Jr., S. Thomas. Managing Quality: An Integrative Approach. New Jersey:

Prentice Hall, 2001

http://www.goalqpc.org/nonflash/whatweteach/RESEARCH/

http://www.goalqpc.org/nonflash/whatweteach/RESEARCH/7mp.html

http://www.pekin.net/

http://www.skymark.com/resources/tools/

http://www.uvm.edu/~auditwww/tools/?Page=interrel.html

Quality Certifications

ISO 9000—ISO, the International Organization for Standardization, is the global federation of 130 organizations throughout the world that establishes national standards. ISO was established in 1947 and is based in Geneva, Switzerland. ISO seeks to promote the development of world standards in order to facilitate the international exchange of goods and services and to facilitate cooperation in intellectual, scientific and technological activity. Standards are rules, guidelines, or definitions of characteristics developed to ensure that materials, products, processes and services are fit for their purpose. For example, ISO defined the shape and content of the original 'standard' credit card so that all companies could develop equipment that would use the same type of card.

ISO 9000 is a family of standards developed by ISO that relate to quality management systems. ISO 9000 was first developed in 1987. The original standard was replaced with ISO 9001 in 1994, and was replaced in 2000 with “ISO 9000:2000.” ISO 9000 defines standard methods used by a company to ensure that its processes meet customer requirements. Said another way, ISO 9000 states what constitutes acceptable company practices in order for the company to be considered to have properly managed its processes for product quality. ISO 9000 does not, however, define specific product standards or ensure that a company's products are 'of quality.' Companies that use these standards arrange to be audited by independent auditing agencies to confirm proper usage of the standards, then publicize themselves as "ISO 9000 compliant" or "ISO 9000 certified." More and more, companies prefer to vend from ISO 9000 compliant suppliers.

Ford, DaimlerChrysler and General Motors have adapted and interpreted ISO 9000 to meet the particular needs of the automobile industry and refer to that interpretation as Quality System Requirements QS-9000 or simply QS-9000.[i] ISO also developed a '14000' family of standards relating to environmental management systems that work in similar fashion as ISO 9000. Numerous research efforts have failed to demonstrate any significant relationship between ISO 9000 implementation and improvement in quality practices or performance measures.

Professional Quality Organizations

ASQ (**The American Society for Quality**[ii], formerly the American Society for Quality Control) is the most widely recognized organization devoted to individuals that work as quality professionals in America. It is also one of the oldest such associations in the world. There are many other such organizations for quality professionals. **AQP**, **The Association for Quality and Participation**,[iii] a smaller professional organization, eventually merged with ASQ. **APQC**, **The American Productivity & Quality Center**[iv], is primarily designed to provide various quality-related services to its corporate members. Since the 1990s, various quality organizations have emerged in other countries, American states, and regions of the world. One example, is **EOQ**, **The European Organization for Quality**.

Quality Awards and Prizes

Malcolm Baldrige National Quality Award[v]—The Malcolm Baldrige National Quality Award is the United States National Award for Quality. The Award was created in 1987 in order to promote quality and the recognition and sharing of quality techniques in American industry, in response to the loss of American leadership in quality to foreign competitors. The Award is named for Malcolm Baldrige, who served as Secretary of Commerce in the Reagan administration until he passed away in 1987. The award is administered by the National Institute of Standards and Technology, an agency under the Department of Commerce. Alternate terms for the award are MBNQA, “the Baldrige” or 'the Baldy.'

Up to two Baldies can be awarded each year in each of five categories: manufacturing, service, small business, education and health care.[1] Applicants are judged for quality in seven areas:

leadership, strategic planning, customer and market focus, information and analysis, human resource development and management, process management and business results. Winners are determined by an independent Board of Examiners, comprised of quality expert volunteers from industry, universities and government, who review company applications and company data then conduct site visits in order to reach their decision. Baldy winners typically share their lessons and experience with other firms. Many states now have quality awards that are awarded on a basis similar to the Baldy.

Over its history, the point values for each of the seven judged areas have changed. Today, "business results" represents a disproportionately high portion of the points. This emphasis on "business results first" seems to be highly inconsistent with the process improvement perspective of quality philosophy.

Critics of the quality philosophy often point to several negative events surrounding the Baldrige Award. The Wallace Company, the first small business to win the Award, filed for bankruptcy within a year of receiving the Award. Around 1990, a Baldrige was awarded to IBM for design and development of the midrange AS/400 mid-size computer. Unfortunately, at introduction, IBM found the market for this larger computer had dwindled in favor of the smaller, personal computers that people typically use today. In 2003, Globe Metallurgical, one of the first companies to win the Award, also filed for bankruptcy. Further, a hypothetical stock portfolio of Baldrige Award winners "beat" the S&P 500 for nine years … until 2002. That year the portfolio lost 34% of its value, as compared to a 48% increase in the S&P 500.

The Deming Prize--The Deming Prize has been awarded annually since 1951 by JUSE, the Union of Japanese Scientists and Engineers[vi], an organization at the heart of quality management in Japan. The Deming Prize is the oldest quality prize in the world as well as the most prestigious quality prize in Japan. Deming prize winners are expected to meet remarkably high expectations in a large number of areas such as quality policies, employee involvement in the organization of quality, use of statistical quality techniques, systematic handling of standards, empowerment and training of employees, methods of customer satisfaction, quality assurance through process control, environmental protection, use of the PDSA cycle, quality circles, and tangible effects such as cost and profit as well as intangible effects.[vii] There are several categories for the Prize, including factories, small companies and individual citizens.

In 1989, Florida Power & Light (FPL) became the first American company to win the Deming Prize.[viii] FPL's quality initiative resulted in many performance improvements:

- Using root cause analysis, meter readers determined that one root cause of meter reading difficulties was due to aggressive dogs in home owner's yards. The readers began to carry binoculars for reading meters from a distance. The time required for meter reading was substantially reduced.
- Line repairmen performed root cause analysis to improve line downtime. They determined that a significant root cause of downed lines was fallen branches. This was because tree branch trimming was not scheduled based on the rate at which branches grow, but rather were being trimmed in rotational order. The repairmen developed a new trimming schedule based on the different growth rates of the trees, trimming branches of faster growing trees more often. Line downtime was reduced by almost half.

Unfortunately, when the next President took the helm of FPL, he found that quality principles (such as worker empowerment) conflicted with his traditional, "top-down" management style and so he dissolved most of the company's quality initiative.

Shingo Prize for Excellence in Manufacturing[ix]--The Shingo Prize for Excellence in Manufacturing recognizes companies for excellence in process improvement, quality and customer satisfaction through implementing lean and just-in-time methods, eliminating waste, achieving zero defects and continuously improving. The Shingo Prize is named for Shigeo Shingo, an industrial engineer who worked with Taiichi Ohno at Toyota as it developed the modern manufacturing practices often referred to as the Toyota Production System. The Shingo Prize has been awarded since 1989 and is administered by Utah State University. Individual states are now presenting Shingo Prizes; the North Carolina Shingo Prize is administered through the Industrial Extension Service of North Carolina State University.[x]

Kano's 'Delightful Quality' Model

In the late 1970s and early 1980s, while consulting Konica, Noritaki Kano delineated two major types of quality.[xi] Kano's work was almost certainly influenced by Herzberg's Two Factor Theory (or, alternatively, Motivation-Hygiene Theory), introduced in a 1959 book and popularized by a 1968 *Harvard Business Review* article:

> **Expected quality** delivers that which the customer expects and assumes (Example: casino customers expect clean restrooms.). If the customer DOES receive expected quality, it does little to affect his/her perceptions …he "expects and assumes" this type of quality. On the other hand, if a customer DOES NOT receive expected quality, he/she will be motivated to reject the product (Example: "Look at this filthy restroom … and there's no toilet paper. I'll never stay at this casino again!"). Expected quality, therefore, contributes much more to customer dissatisfaction than to customer satisfaction. Alternate terms for expected quality include assumed quality, stay-on quality, order qualifiers and must-be quality.

Delight quality delivers that which the customer DOES NOT expect or assume (Example: "Wow, I just love this casino … their restrooms have newspapers, chandeliers, fresh flowers, gold-plated fixtures, air that smells like fresh-cut roses, free mouthwash & cologne and an attendant to hand me towels!). If a customer DOES receive delight quality, he/she will be motivated to select the product … he did NOT expect this type of quality and so is delighted to receive it. On the other hand, if a customer DOES NOT receive delight quality, it does little to affect his/her perceptions …he/she didn't expect to receive it. Delight quality, therefore, contributes much more to customer satisfaction than to customer dissatisfaction. Alternate terms for delight quality include latent quality (ie, fulfilling latent needs), turn-on quality and 'wow' quality.

The only real difference between Herzberg's Theory and Kano's Model is that Herzberg is speaking of how ***workers*** are motivated/satisfied by the conditions of the work environment (ie, by the nature of the bathrooms at their workplace) while Kano is speaking of how ***customers*** are motivated/satisfied by the quality of the purchased product (ie, by the quality of the bathrooms during the casino 'experience' they purchased). Further, there just doesn't seem to be a whole lot of difference between Kano's Model and the concept of "order-winners versus order-qualifiers" that was "introduced" by Terry Hill of the very, very prestigious London Business School in his 1997 textbook *Operations Strategy*.

According to Kano's model, quality requirements must be fulfilled in order … the customer must be given expected quality first, then delight quality. It is of little use to try to delight restroom customers with newspapers and cologne … if there's no toilet paper. In that sense, Kano's model also bears a similarity to Maslow's Hierarchy of Needs Model, where an individual's basic physiological and safety needs (such as food, water, shelter and warmth) come before higher-level esteem & "belonging" needs (such as friendship, family and respect).[2]

Costs of Quality (COQ)

For most of the 20th Century, most Western managers assumed that it costs *more* to deliver a product of higher quality (than one of lower quality). At first blush, such a line of thinking looks reasonable. Wouldn't it seem that higher quality would require more carefully crafted materials, higher skilled labor, more talented product designers, greater process precision, more inspection and so forth, all of which would raise costs? It turns out, however, that the interaction between quality and cost is a good bit more complicated.

Joe Juran devoted the first forty-one pages of the first edition of his *Quality Control Handbook* (1951) to a discussion of the economics and costs of quality. It included statements such as:

- "In numerous instances a design is both better *and* cheaper than some other design … by simplifying the design, [using] fewer parts and [subjecting] them to fewer operations."
- "To improve quality … requires that fewer defects be produced … this means less scrap, fewer reworks, less sorting [out of defects], etc. … shop costs go way down."
- "The fewer defects made in the shop, the fewer there are to go on to the customer. In this way, customer complaints also go down."
- "It is possible to fail in business even though quality of product is good. However it is not possible to stay in business if quality is poor … unless one has a monopoly."
- "Quality makes sales." "The value of a good reputation for quality is unbelievably high. To the sales force, a good quality reputation is a matchless tool for competition." "Quality of product is thus a weapon of competition."
- " 'Gold in the mine' … what current costs would disappear if all defects disappeared? … material scrapped … [costs] to effect repairs on salvageable product … burden arising from excess production capacity necessitated by defectives … excess inspection costs … investigation of causes of defects … discounts on 'seconds' … customer complaints … delays and stoppages caused by defectives … customer goodwill."

In essence, Juran made the argument that high quality *reduces* cost, increases revenue and is a minimal requirement to stay in business for the long-term. He presented and updated this discussion at the beginning of the second edition (1962) and third edition (1974) of his Handbook as well.

In his 1951 book, Juran, on page 30, states there are four types of quality costs:

- The basic costs to make a product meet specification
- Inspection in its various forms; sampling, sorting
- Quality control
- Avoidable costs (costs 1, 2 and 3 are unavoidable); "gold in the mine"

In 1957, W. J. Masser of General Electric authored an article in which he modified Juran's types of quality into:

- Prevention costs—costs that keep defects from occurring in the first place, such as precision equipment
- Appraisal costs—costs from formal evaluations of product quality, such as inspection
- Failure costs—costs that arise from defects, such as scrap, rework, customer complaints, etc.

Quality guru Phillip Crosby cited Masser's three types in his 1979 book *Quality is Free* (at pages 123-124).

In the third edition of his Handbook (1974), Juran expanded Masser's list into four types:

- Prevention costs—costs that keep defects from occurring in the first place, such as precision equipment
- Appraisal costs—costs from formal evaluations of product quality, such as inspection
- Internal failure costs—costs of defects inside the company, such as scrapped materials
- External failure costs—costs of defects outside the company, such as a warranty claim

The American Society for Quality adopted this four "costs of quality" typology into its common practice.

Notice that, of the four types, only prevention costs can rise as quality improves. The other three types of costs will decrease as quality improves … and they will rise as quality declines. The list is robust. For example, it has been pointed out that environmental disaster costs (such as monies spent to clean up after the Exxon Valdez Alaskan oil spill) are essentially external failure costs resulting from poor quality.

In *Quality is Free*, Phillip Crosby synthesized and "boiled down" the existing understanding of the costs of quality into lay terms with easy-to-understand truisms that appealed to corporate executives such as:

- "It is always cheaper to do the job right the first time."
- "The cost of quality is the cost of doing things wrong. It is the scrap, rework, service after the service, warranty, inspection, tests and similar activities made … by problems."
- "Every penny you don't spend on doing things wrong, over or instead … becomes half a penny right on the bottom line."
- "Think of where your company could be if you completely eliminate failure costs."
- "It is much less expensive to prevent errors than to rework, scrap or service them. The expense of waste can run as much as 15 to 25 percent of sales, and in some companies it does."
- "There are millions of products produced every day that don't wind up in court."

Hopefully, 21st Century Western managers will better understand that quality is not too expensive to produce, but rather it's too expensive NOT to produce it !

Shingo is associated through his books with several well-known operational concepts including:

- **Zero-Quality Control (ZQC)**—ZQC is a method that advocates a combination of "100% source inspection" (an alternate term for "quality at the process") and poka-yoke in order to reduce defects and to reduce the need for statistical quality control methods such as control charts.

- **Non-Stock Production**—The intent of non-stock production methods is to, in so much as possible, make-to-order rather than make-to-stock. The concept of non-stock production is very closely related to the concepts of kanban and "pull" systems.

Customer-Oriented Quality Concepts

Voice of the Customer (VOC)--Companies are often cautioned to include "the voice of the customer" in its policies and planning, to pay close attention to customer needs and preferences. VOC is essentially an alternate term for the concept otherwise referred to as "customer-driven," "customer-focused," or "customer-centered."

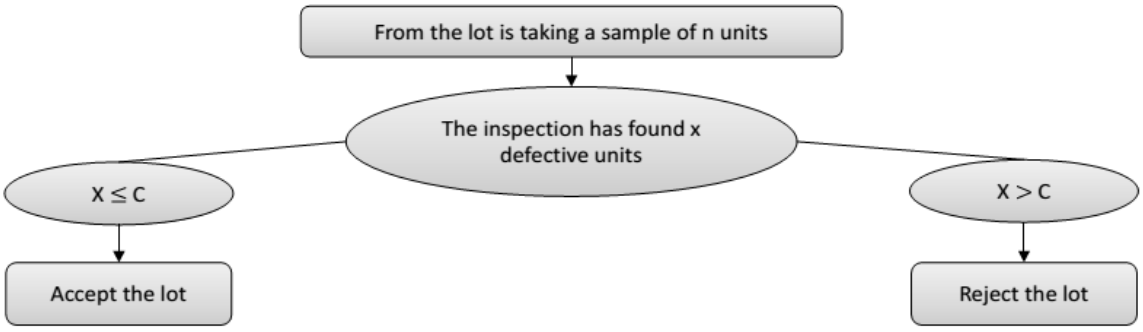
Customer Retention--Companies are recognizing that it is often less expensive to incur costs in order to retain a well understood and desirable customer than to incur costs associated with losing that customer as well as the costs associated with acquiring a replacement customer … who may not be as well understood or desirable. Under this perspective, for instance, a portion of advertising costs can be viewed as an expense incurred due to a lack of customer retention. Many companies are beginning attempts to measure their rate of customer retention.

SINGLE AND DOUBLE SAMPLING PLANS

Introduction An acceptance sampling is one of three major statistical areas which are used for quality control and improvement. An acceptance sampling is a form of testing which involves taking random samples of lots and measuring them against predetermined standards. Depending on kind of production, a lot or a batch can contain one or, in majority cases, more than one raw material, component or final product. The lot can be inspected immediately following production or before the product is shipped to the customer. From other side, the lots can be inspected as they are received from the supplier. In the first case it is conducted outgoing inspection and in the second case incoming inspection (Montgomery et al., 2011). If a supplier's process provides no defective units and no economic justification to make an inspection exists, the lot is accepted without inspection. But when defective units might result in a considerably high failure costs for the buyers, or when it is known that a supplier's processes do not meet a certain level of quality standards, than is a 100% inspection recommended (Montgomery, 2009). An acceptance sampling is between those two approaches to the lot inspection and is used only when the samples of lots are inspected. On this way, an acceptance sampling can, in the same time, save resources and ensure that the output of a process confirms to requirements. An acceptance sampling is most useful to conduct.

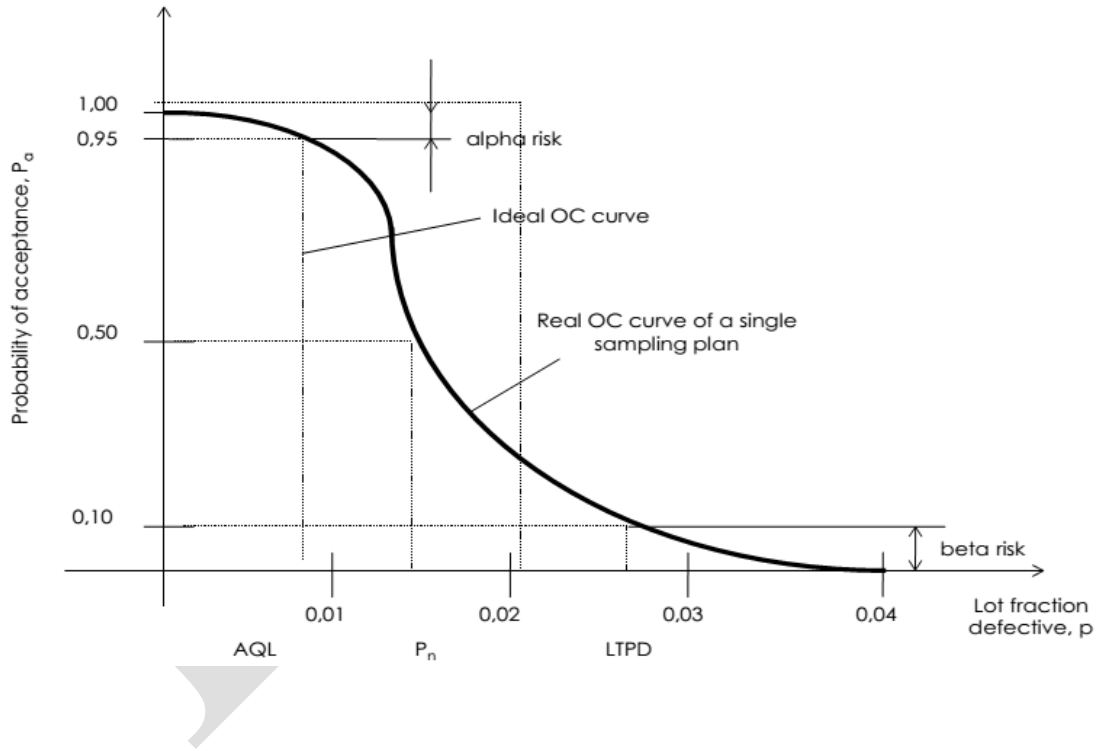
when product testing is destructive, very expensive, very time consuming or when product liability risks are significant (Starbird, 1994). Because of sampling, it could happen that the lot which has a satisfactory predetermined level of quality will be rejected and that the lot which does not have a satisfactory predetermined level of quality will be accepted. The first situation is known as supplier's risk or α, and the second situation is known as buyer's risk or ß (Wadsworth et al., 2002). There are a number of different ways to classify acceptance sampling plans. One of major classification is by variables and attributes (Montgomery et al., 2011). Variables are quality characteristics which are measured on a numerical scale, and attributes are quality characteristics which are expressed as binary result (e.g. a good-bad lot, go-no go). Acceptance sampling plans for attributes are easier to conduct than acceptance sampling plans for variables. Because of that, acceptance sampling plans for attributes are more common in practice and in this paper focus is on this type of acceptance sampling plans. There are different kinds of acceptance sampling plans for attributes: single, double, multiple, sequential and skip lot sampling plans (NIST/SEMATECH, 2003). Single and double acceptance sampling plans are most used in practice because they are simple to use, and so the main focus of this paper is on these kinds of acceptance sampling plans. The aim of this paper is to investigate how usage of different acceptance sampling plans could lead to different conclusions and decisions about accepting or rejecting a lot. Both single and double sampling plans in different cases are considered and compared. Parameters for the sampling plans are changed to simulate possible results and conclusions of inspections. The sampling plans are compared at the same level of probability of acceptance. Inference is based on a statistical test for the difference in proportions. When the test shows that the difference is statistically significant, the producer's quality manager could lower quality costs easily by choosing one of the observed sampling plans. The producer's quality manager may choose to use the sampling plan which shows statistically smaller lot fraction of defective units. On that way, the customer could be misleading about the lot. The paper is emphasizing the probability of this fraud and investigates circumstances under which is this possible to do. Chosen sampling plans Because of wide and often usage, in the paper are considered single and double sampling plans for attributes only. Single sampling plan The simplest types of sampling plans are those which involve a single sample. Such sampling plan is single sampling plan for attributes. This plan is determined with two parameters. The first parameter is the number of random chosen units from the observed lot or n, and the second is the number of defective units which can be tolerated or c. If the lot inspection finds more defective units in the lot than is allowed, the lot will be rejected. If the lot is rejected, there are some possible actions which can the customer undertake. The most rigorous action would be sending the lot back to the producer. But in most cases, next step action is inspecting all units in the lot and replacing defective units with good units. On the other hand, if the lot inspection finds x defective units in the lot, where the x is equal or smaller than c, the lot will be accepted.

The way of making this decision is shown in Figure 1



An operating characteristic (OC) curve is a graphical display of a sampling plan. It describes how well a sampling plan discriminates between good and bad lots (Dumičić, et al., 2006). Under good lots are meant lots which have satisfactory determined level of quality or lots in which the number of defective units is not greater than the maximum allowed number of defective units. Bad lots are such lots which should be rejected because they do not have expected level of quality. Each sampling plan has a unique OC curve. In Figure 2 is shown the general shape of the OC curve for a single sampling plan. The shape of the OC curve for a single sampling plan is determined with parameters n and c.

Figure 2
The Operating Characteristic (OC) Curve of a Single Sampling Plan



The OC curve shows the probability that a lot, submitted with a certain fraction defective, will be either accepted, or rejected (Montgomery et al., 2011). On the x-axis are shown proportions (or probabilities) of defective units which are made in a certain process. These proportions are denoted as p and represent the lot fraction of defective units. On the y-axis are shown probabilities of acceptance of a certain lot. These probabilities are denoted as Pa. In acceptance sampling units for an inspection are chosen randomly and without replacing from a finite population (a finite lot). So the proper way to compute probability of acceptance would be to use the hypergeometric distribution. But if the lot size is large relative to the sample size, the binomial distribution may be used. This approximation is quite satisfactory if the lot size is more than ten times the sample size (Wadsworth et al., 2002). In other words, if the lot size N is large enough to be declared as infinite, the distribution of the number of defectives x in a random sample of n units will be binomial with parameters n and p. In this case, the probability of observing exactly x defectives is:

$$P\{x \text{ defectives}\} = f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (1)$$

where n is the sample size or the number of chosen units for inspection, x is the number of founded defective units in the lot, p is lot fraction defective or assumed process probability to make a defective unit. From that, the probability of acceptance or probability that x is less than or equal to c is:

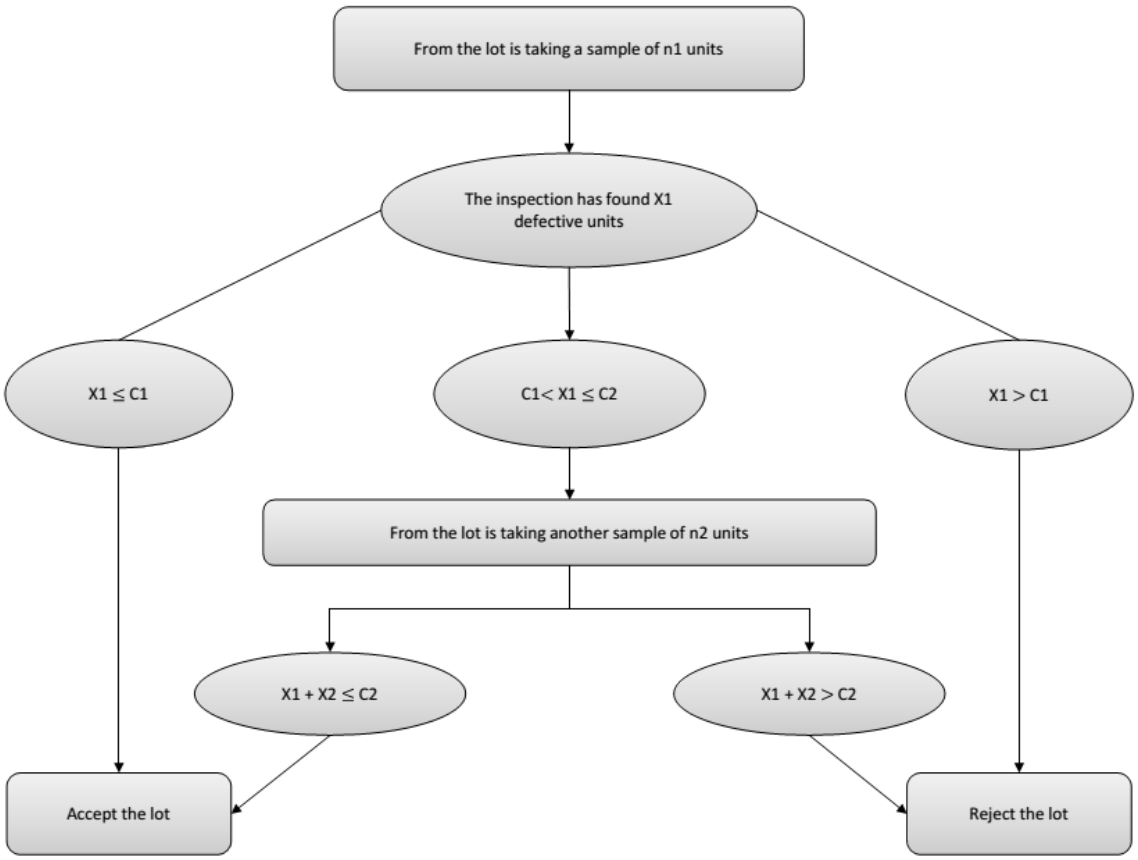
$$P_a = P\{x \le c\} = \sum_{x=0}^{c} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (2)$$

The OC curve contains three specific points which are shown in Figure 2. These points are the acceptable quality level (AQL), the lot tolerance percent defective (LTPD) and the neutral quality level (Pn). The AQL represents the poorest or minimum level of quality of supplier's process that the buyer would consider to be acceptable as a process average (Montgomery et al., 2011). The AQL is base for making a decision of accepting or rejecting the observed lot. Because of sampling, there always exists some probability of rejecting a lot. Even if the supplier's process ensures a smaller lot fraction defective than specified with AQL, this probability will exist. This probability is known as supplier's risk or a risk. There is always the probability that the buyer will accept a lot of poor quality, too. This probability is known as buyer's risk or ß risk. From the other side, the LTPD is the minimum quality level that the buyer is willing to accept. Between the AQL and LTPD points there is an area of indifference. So, there in the point Pn is the neutral quality level. In this point the probability of acceptance and rejecting a lot is equal and it has value of 0.50.

Double sampling plan

In contrast to a single sampling plan, a double sampling plan implies possibility of taking two independent samples of units from the lot. The second sample is formed only when it is necessary. In Figure 3 is shown the way of decision making in a double sampling plan.

Figure 3
Decision Making in a Double Sampling Plan


From the lot is taking a sample of n1 units
The inspection has found X1 defective units
X1 ≤ C1
C1< X1 ≤ C2
X1 > C1
From the lot is taking another sample of n2 units
X1 + X2 ≤ C2
X1 + X2 > C2
Accept the lot
Reject the lot

Like in a single sampling plan, first are taken some units from observed lot for inspection. On that way in a double sampling plan is formed the first sample of n1 units. Because of possibility that one more sample will be formed, the number of units in this sample is usually smaller than the number of units in the sample in a single sampling plan. If the sample size in a single sampling plan is equal to n, it can be written $n > n_1$. This difference in sample sizes is a big advantage for a double sampling plan. Because of smaller initial sample size, a double sampling plan needs less resources to conduct it than a single sampling plan. It has to be emphasized that this advantage comes to expression only if is needed to chose just one sample to make a decision. In a double sampling plan there are set two limits of maximum allowed number of defective units. The first limit is denoted as c1 and is known as the acceptance level for the first sample. The second limit is denoted as c2 and is known as the acceptance level for both samples. The first limit is always smaller than the second. Usually is the second limit double bigger than the first one. If the number of defective units which have been found in the first sample is smaller or equal to the first limit, the lot will be accepted. But if the number of defective units which have been found in the first sample is greater than the second limit, the lot will be rejected. In Figure 4 are shown OC curves for these situations. The curve A shows the probability of acceptance on the first sample and is calculated as

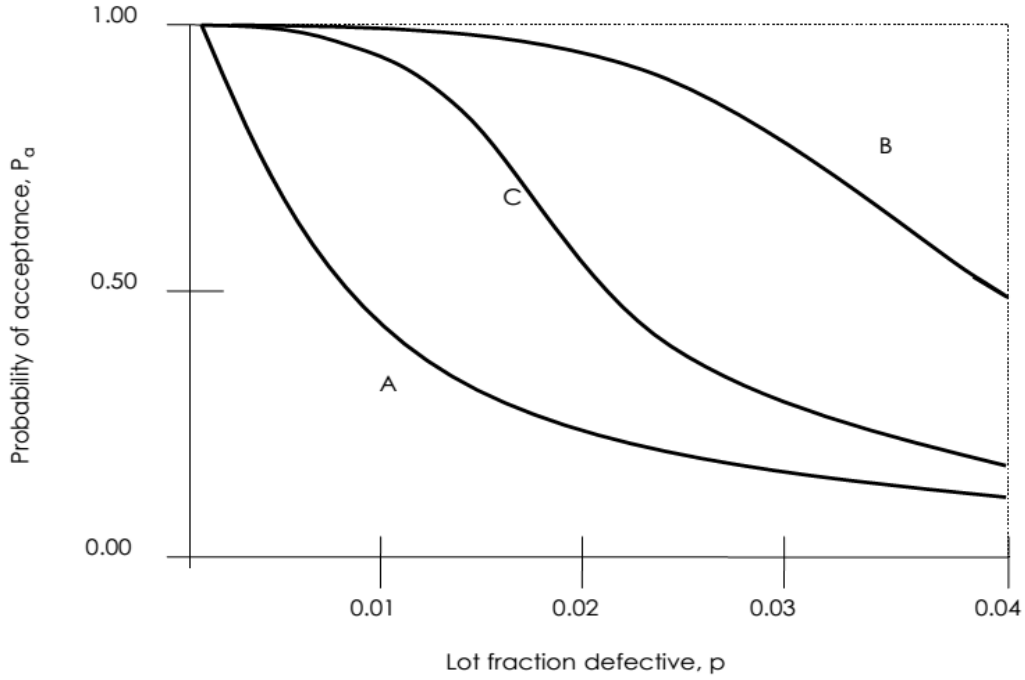
$$P_a^A = P_a^I = P\{x_1 \le c_1\} = \sum_{x_1=0}^{c_1} \frac{n_1!}{x_1!(n_1 - x_1)!} p^{x_1}(1-p)^{n_1 - x_1} \quad (3)$$

The curve B describes the probability of not rejecting the lot on the first sample and is calculated as

$$P_a^B = P\{x_1 \le c_2\} = \sum_{x_1=0}^{c_2} \frac{n_1!}{x_1!(n_1 - x_1)!} p^{x_1}(1-p)^{n_1 - x_1} \quad (4)$$

The difference between the two curves is the probability of taking a second sample.

Figure 4
The Operating Characteristic (OC) Curve of a Double Sampling Plan



If the number of defective units found in the first sample is bigger than the first limit simple but not bigger than the second limit, a decision cannot be made. In this case, the second sample of n2 units from the lot has to be formed. Usually the second sample size is equal to the first sample size or is twice as the first. The number of founded defective units in the second sample, which is denoted as x2, is added to the number of founded defective units in the first sample and then is compared to the second limit. If the overall number of defective units in both samples together is found to be equal to or smaller than the second limit, then the lot will be accepted. But if the overall number of defective units in both samples together is found to be bigger than the second limit, then the lot will be rejected. The OC curve which represents this situation is shown in Figure 4 and is denoted as curve C. The curve C is the final curve and describes the probability of acceptance for the double sampling plan. It can be calculated as sum probabilities of acceptance on the first and second samples or as

$$P_a = P_a^I + P_a^I = P\{x_1 \le c_1\} + \sum_{i=c_1+1}^{c_2} P\{x_1 = i\} P\{x_2 \le c_2 - i\} \quad (5)$$

POSSIBLE QUESTIONS

PART-B(SIX MARKS)

1. What are the seven tools of quality control?Are some more important than others? would you use these tools separately or together?
2. Discuss what is meant by quality control and quality improvement.
3. Discuss Deming's fourteen points for improvement of quality in detail?
4. Define quality and write about Quality Guru's?
5. Assess the progress of the benchmarking exercise to date, explaining the actions that have been undertaken and those that are still required.
6. Relationship between quality and statistical process control?
7. Define Leadership and explain the seven habits of quality leaders.
8. What assumptions are necessary for constructing an x bar chart?
9. What are the components of a quality control chart?
10. Define Leadership and explain the seven habits of quality leaders.

PART- C (TEN MARKS)

11. What assumptions are necessary for constructing an X bar chart?

[i] http://www.asq.org/standcert/9000.html
[ii] http://www.asq.org
[iii] http://www.asq.org/teamwork/
[iv] http://www.apqc.org
[v] http://www.quality.nist.gov
[vi] http://www.juse.or.jp/e-renmei/e-r-4.htm
[vii] _juse#crit
[viii] A list of all Deming Prize winners is found at: http://deming.eng.clemson.edu/pub/den/deming_prize3.htm
[ix] http://www.shingoprize.org
[x] http://www.ies.ncsu.edu/ncshingo/
[xi] Actually it was *five* types of quality, but I am simplifying the Kano Model for you here in this lecture note.