R.M.K
GROUP OF ENGINEERING INSTITUTIONS

R.M.K
GROUP OF INSTITUTIONS

R.M.K
GROUP OF
INSTITUTIONS

R.M.K GROUP OF INSTITUTIONS

Please read this disclaimer before proceeding:

This document is confidential and intended solely for the educational purpose of RMK Group of Educational Institutions. If you have received this document through email in error, please notify the system manager. This document contains proprietary information and is intended only to the respective group / learning community as intended. If you are not the addressee you should not disseminate, distribute or copy through e-mail. Please notify the sender immediately by e-mail if you have received this document by mistake and delete this document from your system. If you are not the intended recipient you are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

R.M.K GROUP OF INSTITUTIONS

22MA401

PROBABILITY AND STATISTICS + LAB

DEPARTMENT	**Artificial Intelligence and Data Science**
BATCH/YEAR	**2022-2026/ II**
CREATED BY	**Dr. V. Banu Priya & Dr. S. Vidhya**
DATE	**26.02.2024**

R.M.K GROUP OF INSTITUTIONS

Table of Contents

COURSE OBJECTIVE

S. No	TOPIC
1	To Provide the necessary basic concepts of random variables and to introduce some standard distributions.
2	To introduce the basic concepts of two dimensional random variables
3	To test the hypothesis for small and large samples.
4	To introduce the concepts of Analysis of Variances.
5	To understand the concept of statistical quality control

PREREQUISITES

S. No.	TOPICS	COURSE NAME WITH CODE
1	Basic Probability	Higher Secondary level
2	Basic Statistics	

Syllabus

22MA401	PROBABILITY AND STATISTICS	L T P C
	(Theory Course with Laboratory Component)	3 2 0 4

UNIT I ONE DIMENSIONAL RANDOM VARIABLES **15**

Basic probability definitions- Independent events- Conditional probability (revisit) - Random variable - Discrete and continuous random variables – Moments – Moment generating functions – Binomial, Poisson, Geometric, Uniform, Exponential and Normal distributions.

Experiments using R Programming:

1. Finding conditional probability.
2. Finding mean, variance and standard deviation.

UNIT II TWO DIMENSIONAL RANDOM VARIABLES **15**

Joint distributions – Marginal and conditional distributions – Covariance – Correlation and linear regression – Transformation of random variables.

Experiments using R Programming:

1. Finding marginal density functions for discrete random variables
2. Calculating correlation and regression

UNIT III TESTING OF HYPOTHESIS **15**

Sampling distributions - Estimation of parameters - Statistical hypothesis - Large sample tests based on Normal distribution for single mean and difference of means -Tests based on t and F distributions for mean and variance – Chisquare - Contingency table (test for independence) - Goodness of fit.

Experiments using R Programming:

1. Testing of hypothesis for given data using Z – test.
2. Testing of hypothesis for given data using t – test.

UNIT IV DESIGN OF EXPERIMENTS **15**

One way and Two way classifications - Completely randomized design – Randomized block design – Latin square design.

Experiments using R Programming:

1. Perform one- way ANOVA test for the given data.
2. Perform two-way ANOVA test for the given data.

UNIT V STATISTICAL QUALITY CONTROL **15**

Control charts for measurements (X and R charts) – Control charts for attributes (p, c and np charts) – Tolerance limits.

Experiments using R Programming:

1. Interpret the results for $\bar{X}$-Chart for variable data
2. Interpret the results for R-Chart for variable data

TOTAL: 75 PERIODS

COURSE OUTCOMES

Course Outcomes	Description	Knowledge Level
CO1	Understand the fundamental knowledge of modern probability theory and standard distributions.	K1, K2
CO2	Categorize the probability models and function of random variables based on one and two dimensional random variables.	K2
CO3	Employ the concept of testing the hypothesis in real life problems.	K3
CO4	Implement the analysis of variance for real life problems.	K3
CO5	Apply the statistical quality control in engineering and management problems.	K3

CCO-PO/CO-PSO Mapping

CO's	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	2	1	-	-	-	-	-	1	-	-	-
CO2	3	2	1	-	-	-	-	-	1	-	-	-
CO3	3	2	1	-	-	-	-	-	1	-	-	-
CO4	3	2	1	-	-	-	-	-	1	-	-	-
CO5	3	2	1	-	-	-	-	-	1	-	-	-

CO's	PSO1	PSO2	PSO3
CO1	-	-	-
CO2	-	-	-
CO3	-	-	-
CO4	-	-	-
CO5	-	-	-

LECTURE PLAN

S.No.	Topics to be Recovered	No. of periods	Proposed Date	Actual Date	CO	Knowledge level	Mode of Delivery
1	Introduction	1	01-03-24		CO4	K2	PPT,Chalk ,Board
2	One way classifications	1	02-03-24		CO4	K3	PPT,Chalk ,Board
3	Solving Problems	1	05-03-24		CO4	K3	PPT,Chalk ,Board
4	Two way classifications	1	07-03-24		CO4	K3	PPT,Chalk ,Board
5	Solving Problems	2	08-03-24		CO4	K3	PPT,Chalk ,Board
6	Completely randomized design	1	11-03-24		CO4	K3	PPT,Chalk ,Board
7	Solving Problems	2	12-03-24		CO4	K3	PPT,Chalk ,Board
8	Randomized block design	1	14-03-24		CO4	K3	PPT,Chalk ,Board
9	Solving Problems	1	15-03-24		CO4	K3	PPT,Chalk ,Board
10	Latin square design	1	18-03-24		CO4	K3	PPT,Chalk ,Board
11	Solving Problems	2	19-03-24		CO4	K3	PPT,Chalk ,Board
12	Solving Problems	2	20-03-24		CO4	K3	PPT,Chalk ,Board

ACTIVITY BASED LEARNING

Activity based learning enhances students' critical thinking and collaborative skills. Experiential learning being the core, various activities such as quiz competitions, group discussion, etc. are conducted for all the five units to enhance the learning abilities of students. The students are the center of the activities, where student's opinions are valued, questions are encouraged, and discussions are done. These activities empower the students to explore and learn by themselves.

Sl.No.	TOPIC	Activity	Link
1	Design of Experiments	Practice quiz in Design of Experiments	https://quizizz.com/admin/quiz/5febc0bf9adef8001e1db866/anova

12

LECTURE NOTES

UNIT IV DESIGN OF EXPERIMENTS

Introduction:

An experiment is conducted with an objective or to verify certain hypothesis. The sequence of steps taken to ensure a scientific analysis leading to valid inferences about the hypothesis is called "**Design of Experiment**". It had its origin from agricultural research and it is credited to Prof. R.A. Fisher. For example, to verify the claim that a particular manure causes increase in the yield of paddy, we may conduct an agricultural experiment. In this experiment the quantity of manure used and the quantity of yield are two variables involved directly. These variables are called **experimental variables**. There may be other variables such as the fertility of soil, the amount of rainfall, the inherent quality of seed etc., which also affect the yield. These are called **extraneous variables** as far as the hypothesis is concerned.

The prime objective of design of experiment is to control the extraneous variables so that the results could be attributed only to the experimental variables.

Independent experimental variables are called factors. Factors may be quantitative or qualitative. Quantitative factor takes a real number as value. For example, amount of fertilizer, kilograms of products etc., Factors that are not quantitative are called qualitative.

Basic Principles of Design of Experiment:

There are three basic principles in designing of experiments. They are

Randomization

Replication

Local Control

1. Randomization

Random assignment of treatment to the experimental units is the most effective way of eliminating any unknown bias in the experiment. For example, if we select plots for use of manure at random, the plots are **experimental units** and application of manure is **treatment**. The plots-selected for not using the manure are called **control units** or **control group**. The control group provide a standard for comparison.

2. Replication

Replication is the process of repeating the same treatment on more than one of the experimental units. Two identically treated plots (or units) will not give identical results. The differences are attributed to uncontrollable random causes. Such differences are called experimental errors. As the number of replication increases this error is reduced. So, replication is necessary to increase the accuracy of estimates of the treatment effects.

3. Local Control

Another way of controlling the effects of extraneous variables is by employing the principle of local control. It consists of techniques of grouping, blocking and balancing of the experimental units.

Grouping means combining sets of homogeneous experimental units into groups so that different groups may be subjected to different treatments. Each group can have different number of experimental units.

Blocking means assigning the same number of plots or experimental units to different groups, called blocks. The plots in the same block are relatively similar or homogeneous. We may use at random different manures to different plots in a block.

Balancing means act of equalizing total effect of the extraneous variables on all the elements in the control group and in the experimental group.

Basic designs of experiments:

Depending on the number of extraneous variables whose effects are to be controlled, various design procedures are developed in the study of experimental design, we shall consider here three important designs.

1. Completely Randomized Design (CRD)
2. Randomized Block Design (RBD)
3. Latin Square Design (LSD)

1. Completely Randomized Design

In a completely randomized design the treatments are given to the experimental units by a procedure of random allocation. It is used when the units are homogeneous.

Suppose there are 5 manures and 20 plots we shall give a random allocation of treatments as below. Write the numbers of the plots 1, 2, 3, ….20 in identical cards and shuffle well. Name the manures or treatments as A, B, C, D, E.

Choose 4 cards at random. The plots bearing these numbers may be given manure A. Shuffle the remaining cards well and choose another 4 cards at random. The plots having these numbers may be given manure B and so on. This is called completely randomized design. In this design there is only one factor namely "treatment".

Merits:

C.R.D results in the maximum use of the experimental units, since all the experimental material can be used.

The design is very flexible. Any number of treatments can be used and different treatments can be used unequal number of times without unduly complicating the statistical analysis in most of the cases.

The statistical analysis remains simple, if some or all the observations for any treatment are rejected or lost or missing for some purely random accidental reasons. We merely carry out the standard analysis on the available data. Moreover, the loss of information due to missing data is smaller in comparison with any other design.

It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small experiments, i.e., for experiments with small number of treatments.

Demerits:

In certain circumstances, the design suffers from the disadvantage of being inherently less informative than other more sophisticated layouts. This usually happens, if the experimental material is not homogeneous. Since, randomization is not restricted in any direction to ensure that the units receiving one treatment are similar to those receiving the other treatment, the whole variations among the experimental units is included in the residual variance. This makes the design less efficient and results in less sensitivity in detecting significant effects. As such C.R.D is seldom used in field experimentation, where due to the fertility gradient of the soil, the whole experimental material, viz., field is not homogeneous and it is better to use more efficient designs like Randomized Block Design (R.B.D) or Latin Square Design (L.S.D) etc.,

Applications:

Completely randomized design is more useful in laboratory technique and methodological studies, (eg.) in physics, chemistry or cookery, in chemical and biological experiments, in some green house studies, etc., where either the experimental material in homogeneous or the intrinsic variability between units can be reduced.

C.R.D is also recommended in situations where an appreciable fraction of units is likely to be destroyed or fail to respond.

2. Randomized Block Design

Suppose we want to test the effect of $r$fertilisers on the yield of paddy. We divide the plots into h blocks, each block is relatively homogeneous and each block contains r plots. Within each block the plots are selected at random and the r treatments (i.e. fertilizers) are given. Thus in each block only one plot receives one fertilizer. This is repeated for all the h blocks. This design is called randomized block design.

The basic idea in this design is to compare all treatment effects within a block of experimental units, reducing the errors due to extraneous factors by the process of randomization. The data collected from experiments with randomized block design form a two-way classification, classified according to two factors, blocks and treatments.

Merits and Demerits of Randomized Block Design

Merits:

It has a simple layout.

The design controls the variability in the experimental units and gives the treatments equivalence to show their effects.

The analysis of the design is simple and straight forward as in the case of two-way classification of analysis of variance.

The analysis is possible, even in the case of missing observations.

Demerits:

The design is not suitable for large number of treatments, since in this case the block size is large and hence homogeneity of units may not be possible.

Unequal number of replications for equal treatment is not possible.

The shape of the experimental material should be rectangular.

It controls the variability in one direction only.

The analysis of this decision is not as simple as a completely randomized design.

3. Latin Square Design

In this design the experimental units are divided according to two factors and arranged in rows and columns as a $n \times n$ square, known as Latin square. Let there be n treatments each repeated n times so that each row and each column receives only one treatment. The treatment to a plot is decided randomly. Latin square design is used in a wide variety of fields.

Latin square model is effective only if one of the factors has substantial influence on the other. If it is not so, it is not an improvement of the randomized block design.

Data from the Latin square experiments form a three-way classification according to the factors rows, columns and treatments.

Merits and Demerits of Latin Square Design

Merits:

Latin square design controls variability in two directions of the experimental material.

The analysis of the design is simple and straight forward and is a three way classification of analysis of variance.

Demerits:

The process of randomization is not as simple as in RBD.

The number of treatments should be equal to the number of rows and number of columns.

The experimental area should be in the form of a square.

It is suitable only in the case of smaller number of treatments (preferably less than 10).

A 2x2 Latin square is not possible.

Analysis of Variance (ANOVA)

In sampling theory we have seen how to test the significance of difference between two sample means, assuming the samples were drawn from populations having the same variance. But in practice in many situations we have to test the differences between more than two sample means. The technique known as **analysis of variance** (abbreviated as ANOVA) will enable us to test the significance of the differences among more than two sample means.

For example, if we want to compare the mileage achieved by five different brands of petrol we can use ANOVA.

Literally, analysis of variance is a technique that analyse variances. But by doing so, it provides us with a test for the significance of the difference among means.

Basic Concepts:

In order to use analysis of variance we make the following assumptions.

the samples are drawn from normal populations.

the samples are independently drawn from these populations.

all the populations have the same variance.

> Note: In case we are not in a position to make these assumptions in a particular problem the analysis of variance technique should not be used. In such cases we can use non-parametric tests.

Analysis of variance is a technique of partitioning the total sum of squared deviations of all sample values from the grand mean into two parts (i) sum of squares between samples (SSB) and (ii) sum of squares within samples (SSW).

SSB is simply the variance between samples. This variation between samples is due to **assignable** causes.

SSW is variance within the samples. This variation is due to various chance causes which could not be controlled. This is called residual random variation (or error).

Under the null hypothesis there is no difference between means of the populations, we apply F-Test to see any significant difference between the two variances exist or not.

We consider the following types of ANOVA.

One-way classification for CRD.

Two-way classification for RBD.

Three-way classification for LSD.

One-way classification (CRD)

In one-way classification the observations or experimental units are classified according to one factor of interest. For example, the yields of several plots of land may be classified according to the type of fertilizers used. Here the factor is treatment namely the type of fertilizers.

Suppose we have r independent random samples of sizes $n_1, n_2, \ldots n_r$ from r normal populations whose means are $\mu_1, \mu_2, \ldots \mu_r$ so that $\sum n_i = N$. Then

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

$$H_1: not\ all\ are\ equal$$

One-way classification ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Between columns	SSC	$c-1$	$MSC = \frac{SSC}{c-1}$	$F = \frac{MSC}{MSE}$ or $F = \frac{MSE}{MSC}$
Within columns (or Error)	SSE	$N-c$	$MSE = \frac{SSC}{N-c}$	
Total	TSS	$N-1$		

Since value of F should be greater than 1, we take the larger of MSC or MSE to the numerator.

N = Total number of observations

T = Total value of all the observations

c = number of columns

Correction Factor (CF) = $\frac{T^2}{N}$

$$TSS = \sum x_i^2 - \frac{T^2}{N}$$

R.M.K
GROUP OF
INSTITUTIONS

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 6$	$c - 1 = 2$	$MSC = \frac{SSC}{c-1} = 3$	$F = \frac{MSE}{MSC} = \frac{4.86}{3} = 1.62$	$F(7,2) = 19.35$
Within columns (or Error)	$SSE = 34$	$N - c = 7$	$MSE = \frac{SSC}{N-c} = 4.86$		
Total	$TSS = 40$	$N - 1 = 9$			

Conclusion: $F_{cal} < F_{tab}$. Hence H_0 is accepted.

Example 2: There are three main brands of a certain powder. A set of 120 sample values is examined and found to be allocated among four groups (A, B, C and D) and three brands (I, II, III) as shown here under: Is there any significant difference in brands preference? Answer at 5% level.

Brands	Groups			
	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	8	19	11	13

Solution:

H_0: *There is no significant difference in brands*

H_1: *There is significant difference in brands*

$$MSC = \frac{SSC}{c-1} = 40.1 \qquad MSE = \frac{SSC}{N-c} = 20.06$$

$$F = \frac{MSC}{MSE} = \frac{40.1}{20.06} = 1.999$$

Groups	Brands			Total	x_1^2	x_2^2	x_3^2
	I x_1	II x_2	III x_3				
A	0	5	8	**13**	0	25	64
B	4	8	19	**31**	16	64	361
C	8	13	11	**32**	64	169	121
D	15	6	13	**34**	225	36	169
Total	**27**	**32**	**51**	***110***	**305**	**294**	**715**

- Here N=12, c=3 and T=110

$$CF = \frac{T^2}{N} = 1008.3$$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 305 + 294 + 715 - 1008.3 = 305.7$$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N} = 1088.5 - 1008.3 = 80.2$$

$$SSE = TSS - SSC = 305.7 - 80.2 = 225.5$$

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 80.2$	$c - 1 = 2$	$MSC = \frac{SSC}{c-1} = 40.1$	$F = \frac{MSC}{MSE} = \frac{40.1}{20.06} = 1.999$	$F(2,9) = 4.26$
Within columns (or Error)	$SSE = 225.5$	$N - c = 9$	$MSE = \frac{SSC}{N-c} = 20.06$		
Total	$TSS = 305.7$	$N - 1 = 11$			

Conclusion: $F_{cal} < F_{tab}$. Hence H_0 is accepted.

Example 3: The following table shows the lives in hours of four brands of electric lamps.

Brand A	1610	1610	1650	1680	1700	1720	1800	
Brand B	1580	1640	1640	1700	1750			
Brand C	1460	1550	1600	1620	1640	1660	1740	1820
Brand D	1510	1520	1530	1570	1600	1680		

Perform an analysis of variance test the homogeneity of the mean lives of the four brands of lamps.

Solution:

$$H_0: There\ is\ no\ significant\ difference\ between\ the\ four\ brands$$

$$H_1: There\ is\ significant\ difference\ between\ the\ four\ brands$$

Subtract each element with 1600 and then divide by 10, we get

x_1 A	x_2 B	x_3 C	x_4 D	Total	x_1^2	x_2^2	x_3^2	x_4^2
1	-2	-14	-9	-24	1	4	196	81
1	4	-5	-8	-8	1	16	25	64
5	4	0	-7	2	25	16	0	49
8	10	2	-3	17	64	100	4	9
10	15	4	0	29	100	225	16	0
12	-	6	8	26	144	-	36	64
20	-	14	-	34	400	-	196	-
-	-	22	-	22	-	-	484	-
57	31	29	-19	98	735	361	957	267

Here N=26, c=4 and T=98

$$CF = \frac{T^2}{N} = 369.39$$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 735 + 361 + 957 + 267 - 369.39 = 1950.61$$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N} = 464.14 + 192.2 + 105.13 + 60.17 - 369.39 = 452.25$$

$$SSE = TSS - SSC = 1950.61 - 452.25 = 1498.36$$

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 452.25$	$c - 1 = 3$	$MSC = \frac{SSC}{c-1} = 150.75$	$F = \frac{MSC}{MSE} = \frac{150.75}{68.11} = 2.21$	$F(3, 22) = 3.05$
Within columns (or Error)	$SSE = 1498.36$	$N - c = 22$	$MSE = \frac{SSC}{N-c} = 68.11$		
Total	$TSS = 1950.61$	$N - 1 = 25$			

Conclusion: $F_{cal} < F_{tab}$. Hence H_0 is accepted.

Exercise:

1. The following are the numbers of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory:

Technician I (X_1)	Technician II (X_2)	Technician III (X_3)	Technician IV (X_4)
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$, whether the differences among the 4 sample means, can be attributed to chance.

Two-way classification (RBD)

two-way classification of analysis of variance, we consider one classification along column-wise and the other along row-wise. For example, the yield of a crop in several plots of land may be classified according to different varieties of seeds and different varieties of fertilizers. So, seeds and fertilizers are the two factors.

H_0: *There is no significant difference between rows and columns*

H_1: *There is a significant difference between rows and columns*

Two-way classification ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Between columns	SSC	$c-1$	$MSC = \frac{SSC}{c-1}$	$F_c = \frac{MSC}{MSE}$
Between rows	SSR	$r-1$	$MSR = \frac{SSR}{r-1}$	$F_R = \frac{MSR}{MSE}$
Residual (Error)	SSE	$(c-1)(r-1)$	$MSE = \frac{SSE}{(c-1)(r-1)}$	
Total	TSS	$rc-1$		

F_c *and* F_R should be calculated in such a way that $F_c > 1$ *and* $F_R > 1$.

N = Total number of observations

T = Total value of all the observations

c = number of columns; r = number of rows

Correction Factor (CF) = $\frac{T^2}{N}$

$TSS = \sum x_i^2 - \frac{T^2}{N}$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N}$$

$$SSR = \left[\frac{(\sum y_1)^2}{m_1} + \frac{(\sum y_2)^2}{m_2} + \cdots\right] - \frac{T^2}{N}$$

$$SSE = TSS - SSC - SSR$$

Find the table value of F for $df\,(v_1 = c - 1, v_2 = r - 1)$ at $\alpha\%$ level of significance.

Conclusion: *If* $F_{cal} < F_{tab}$, we accept H_0, otherwise reject H_0.

Example 1: The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.

Workers	Machine Type			
	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Test whether the five men differ with respect to mean productivity and test whether the mean productivity is the same for the four different machine types.

Solution:

H_0: *There is no significant difference in productivity of workers and the machines*

H_1: *There is a significant difference in productivity of workers and the machines*

We shall subtract 40 from each value.

Workers		Machine Type								
		A x_1	B x_2	C x_3	D x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
	y_1	4	-2	7	-4	5	16	4	49	16
	y_2	6	0	12	3	21	36	0	144	9
	y_3	-6	-4	4	-8	-14	36	16	16	64
	y_4	3	-2	6	-7	0	9	4	36	49
	y_5	-2	2	9	-1	8	4	4	81	1
Total		5	-6	38	-17	20	101	28	326	139

Here N=20, c=4, r=5 and T=20

$$CF = \frac{T^2}{N} = 20$$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 101 + 28 + 326 + 139 - 20 = 574$$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N} = \left[\frac{(5)^2}{5} + \frac{(-6)^2}{5} + \frac{(38)^2}{5} + \frac{(-17)^2}{5}\right] - 20 = 338.8$$

$$SSR = \left[\frac{(-4)^2}{4} + \frac{(3)^2}{4} + \frac{(-8)^2}{4} + \frac{(-7)^2}{4} + \frac{(-1)^2}{4}\right] - 20 = 161.5$$

$$SSE = TSS - SSC - SSR = 574 - 338.8 - 161.5 = 73.7$$

NOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 338.8$	$c - 1 = 3$	$MSC = \frac{338.8}{3} = 112.93$	$F_c = \frac{112.93}{6.14} = 18.39$	$F_c(3,12) = 3.49$
Between rows	$SSR = 161.5$	$r - 1 = 4$	$MSR = \frac{161.5}{4} = 40.38$	$F_R = \frac{40.38}{6.14} = 6.57$	$F_R(4,12) = 3.26$
Residual (Error)	$SSE = 83.7$	$(c-1)(r-1) = 12$	$MSE = \frac{83.7}{12} = 6.14$		
Total	$TSS = 584$	$rc - 1 = 19$			

For Column: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

For Row: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Hence there is significant difference between the productivity of men as well as machines.

Example 2: An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1	Engine 2	Engine 3	Total
Detergent A	45	43	51	139
Detergent B	47	46	52	145
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Total	182	176	207	565

Perform the ANOVA and test at 0.01 level of significance whether there are differences in the detergents or in the engines.

Solution:

H_0: *There is no significant difference between the engines and between the detergei*

H_1: *There is a significant difference between the engines and between the detergent*

We shall subtract 50 from each value.

	x_1	x_1	x_1	Total	x_1^2	x_2^2	x_3^2
y_1	-5	-7	1	-11	25	49	1
y_2	-3	-4	2	-5	9	16	4
y_3	-2	0	5	3	4	0	25
y_4	-8	-13	-1	-22	64	169	1
Total	-18	-24	7	-35	102	234	31

Here N=12, c=3, r=4 and T=-35

$$CF = \frac{T^2}{N} = 102.08$$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 102 + 234 + 31 - 102.08 = 264.92$$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N} = \left[\frac{(-18)^2}{4} + \frac{(-24)^2}{4} + \frac{(7)^2}{4}\right] - 102.08 = 135.17$$

$$SSR = \left[\frac{(-11)^2}{3} + \frac{(-5)^2}{3} + \frac{(3)^2}{3} + \frac{(-22)^2}{3}\right] - 102.08 = 110.92$$

$$SSE = TSS - SSC - SSR = 264.92 - 135.17 - 110.92 = 18.83$$

NOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=0.01
Between columns	$SSC = 135.17$	$c - 1 = 2$	$MSC = \frac{135.17}{2} = 67.59$	$F_c = \frac{67.59}{3.14} = 21.52$	$F_c(2,6) = 10.92$
Between rows	$SSR = 110.92$	$r - 1 = 3$	$MSR = \frac{110.92}{3} = 36.97$	$F_R = \frac{36.97}{3.14} = 11.77$	$F_R(3,6) = 9.78$
Residual (Error)	$SSE = 18.83$	$(c-1)(r-1) = 6$	$MSE = \frac{18.83}{6} = 3.14$		
Total	$TSS = 264.92$	$rc - 1 = 11$			

Conclusion:

For Column: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Hence there is significant difference between the engines

For Row: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Hence there is significant difference between the detergents

Example 3: The varieties of a crop are tested in a randomized block design with four replications, the layout being as given below. The yields are given in kilograms. Analysis for significance.

C48	A51	B52	A49
A47	B49	C52	C51
B49	C53	A49	B50

Solution:

Given 3 varieties of crops and four replications in blocks.

We shall subtract 50 from each value.

H_0: *There is no significant difference between the varieties and between the blocks*

H_1: *There is a significant difference between the varieties and between the blocks*

Variety	Replication blocks								
	1 x_1	2 x_2	3 x_3	4 x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
A y_1	1	-1	-3	-1	-4	1	1	9	1
B y_2	2	-1	-1	0	0	4	1	1	0
C y_3	-2	2	1	3	4	4	4	1	9
Total	1	0	-3	2	0	9	6	11	10

Here N=12, c=4, r=3 and T=0

$$CF = \frac{T^2}{N} = 0$$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 9 + 6 + 11 + 10 - 0 = 36$$

$$SSC = \left[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots\right] - \frac{T^2}{N} = \left[\frac{(1)^2}{3} + \frac{(0)^2}{3} + \frac{(-3)^2}{3} + \frac{(2)^2}{3}\right] - 0 = 4.67$$

$$SSR = \left[\frac{(-4)^2}{4} + \frac{(0)^2}{4} + \frac{(4)^2}{4}\right] - 0 = 8$$

$$SSE = TSS - SSC - SSR = 36 - 4.67 - 8 = 23.33$$

NOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 4.67$	$c - 1 = 3$	$MSC = \frac{4.67}{3} = 1.56$	$F_c = \frac{3.89}{1.56} = 2.49$	$F_c(6,3) = 8.94$
Between rows	$SSR = 8$	$r - 1 = 2$	$MSR = \frac{8}{2} = 4$	$F_R = \frac{4}{3.89} = 1.03$	$F_R(2,6) = 19.33$
Residual (Error)	$SSE = 23.33$	$(c-1)(r-1) = 6$	$MSE = \frac{23.33}{6} = 3.89$		
Total	$TSS = 36$	$rc - 1 = 11$			

Conclusion:

For Column: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

For Row: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

Hence there is no significant difference between the blocks and between the varieties of crops.

Exercise:

1. A company appoints 4 salesmen A, B, C and D and observes their sales in 3 seasons – Summer, Winter and Monsoon. The figures (in lakhs of Rs.) are given in the following table.

Season	Salesmen			
	A	B	C	D
Summer	45	40	38	37
Winter	43	41	45	38
Monsoon	39	39	41	41

Carry out an analysis of variance.

2. Three varieties A, B, and C of a crop are tested in a randomized block design with 4 replications. The plot yields in pounds are as follows:

A	6	C	5	A	8	B	9
C	8	A	4	B	6	C	9
B	7	B	6	C	10	A	6

Analyse the experimental yield and state your conclusion.

Three-way classification (LSD)

We have seen data from a Latin square experiment results in a three-way classification say (i) variety of seeds (ii) types of spacing (or plots) and (iii) different manure treatment.

H_0: *There is no significant difference between rows, between columns, between treatments*

H_1: *There is a significant difference between rows, between columns, between treatments*

Three-way classification ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio
Between columns	SSC	$n-1$	$MSC = \frac{SSC}{n-1}$	$F_c = \frac{MSC}{MSE}$
Between rows	SSR	$n-1$	$MSR = \frac{SSR}{n-1}$	$F_R = \frac{MSR}{MSE}$
Between letters	SSK	$n-1$	$MSK = \frac{SSK}{n-1}$	$F_K = \frac{MSK}{MSE}$
Residual (Error)	SSE	$(n-1)(n-2)$	$MSE = \frac{SSE}{(n-1)(n-2)}$	
Total	TSS	k^2-1		

F_c, F_R *and* F_K should be calculated in such a way that $F_c > 1, F_R > 1$ *and* $F_K > 1$.

N = Total number of observations

T = Total value of all the observations

Correction Factor (CF) = $\frac{T^2}{N}$

$$TSS = \sum x_i^2 - \frac{T^2}{N}$$

$$SSC = \frac{1}{n}\left[\left(\sum x_1\right)^2 + \left(\sum x_2\right)^2 + \cdots + \left(\sum x_n\right)^2\right] - \frac{T^2}{N}$$

$$SSR = \frac{1}{n}\left[\left(\sum y_1\right)^2 + \left(\sum y_2\right)^2 + \cdots + \left(\sum y_n\right)^2\right] - \frac{T^2}{N}$$

$$SSK = \frac{1}{n}\left[\left(\sum z_1\right)^2 + \left(\sum z_2\right)^2 + \cdots + \left(\sum z_n\right)^2\right] - \frac{T^2}{N}$$

$$SSE = TSS - SSC - SSR - SSK$$

Find the table value of F for $df(v_1 = c - 1, v_2 = r - 1)$ at $\alpha\%$ level of significance.

Conclusion: $If\ F_{cal} < F_{tab}$, we accept H_0, otherwise reject H_0.

Example 1: A farmer wishes to test the effects of four different fertilizers A, B, C, D on the yield of wheat. In order to eliminate sources of error due to variability in soil fertility, he uses the fertilizers, in a Lattin square arrangement as indicated in the following table, where the numbers indicate yields in bushels per unit area.

A 18	C 21	D 25	B 11
D 22	B 12	A 15	C 19
B 15	A 20	C 23	D 24
C 22	D 21	B 10	A 17

Perform an analysis of variance to determine, if there is a significant difference between the fertilizers at $\alpha = 0.05$ level of significance.

Solution:

H_0: *There is no significant difference between rows, between columns, between treatments*

H_1: *There is a significant difference between rows, between columns, between treatments*

We

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	-2	1	5	-9	-5	4	1	25	81
y_2	2	-8	-5	-1	-12	4	64	25	1
y_3	-5	0	3	4	2	25	0	9	16
y_4	2	1	-10	-3	-10	4	1	100	9
Total	-3	-6	-7	-9	-25	37	66	159	107

N = 16, T = -25, $CF = \frac{T^2}{N} = 39.06$

$TSS = \sum x_i^2 - \frac{T^2}{N} = 37 + 66 + 159 + 107 - 39.06 = 329.94$

$SSC = \frac{1}{n}[(\sum x_1)^2 + (\sum x_2)^2 + \cdots + (\sum x_n)^2] - \frac{T^2}{N}$

$= \frac{1}{4}[(-3)^2 + (-6)^2 + (-7)^2 + (-9)^2] - 39.06 = 4.69$

$SSR = \frac{1}{n}[(\sum y_1)^2 + (\sum y_2)^2 + \cdots + (\sum y_n)^2] - \frac{T^2}{N}$

$= \frac{1}{4}[(-9)^2 + (-1)^2 + (4)^2 + (-3)^2] - 39.06 = 29.19$

find SSK:

A	B	C	D
-2	-9	1	5
-5	-8	-1	2
0	-5	3	4
-3	-10	2	1
-10	**-32**	**5**	**12**

$$SSK = \frac{1}{n}[(\sum z_1)^2 + (\sum z_2)^2 + \cdots + (\sum z_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(-10)^2 + (-32)^2 + (5)^2 + (12)^2] - 39.06 = 284.19$$

$$SSE = TSS - SSC - SSR - SSK = 329.94 - 4.69 - 29.19 - 284.19 = 11.87$$

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 4.29$	$n - 1 = 3$	$MSC = \frac{4.29}{3} = 1.56$	$F_c = \frac{MSE}{MSC} = 1.26$	$F_c(6,3) = 8.94$
Between rows	$SSR = 29.19$	$n - 1 = 3$	$MSR = \frac{29.19}{3} = 9.73$	$F_R = \frac{MSR}{MSE} = 4.91$	$F_R(3,6) = 4.76$
Between letters	$SSK = 284.19$	$n - 1 = 3$	$MSK = \frac{284.19}{3} = 94.73$	$F_K = \frac{MSK}{MSE} = 47.8$	$F_K(3,6) = 4.76$
Residual (Error)	$SSE = 11.87$	$(n-1)(n-2) = 6$	$MSE = \frac{11.87}{6} = 1.98$		
Total	$TSS = 329.94$	$n^2 - 1 = 15$			

Conclusion:

Between columns: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

Hence there is no significant difference between columns.

Between rows: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Between letters: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Hence, we conclude that there is a significant difference between rows and between letters.

Example 2: Analyse the variance in the latin square of yields in (kgs) of paddy where P, Q, R, S denote the different methods of cultivation.

S 122	P 121	R 123	Q 122
Q 124	R 123	P 122	S 125
P 120	Q 119	S 120	R 121
R 122	S 123	Q 121	P 122

Examine whether the different methods of cultivation have given significantly different yields.

Solution:

$$H_0: There\ is\ no\ significant\ difference\ between\ rows, between\ columns,$$
$$and\ between\ the\ methods\ of\ cultivation.$$
$$H_1: There\ is\ a\ significant\ difference\ between\ rows, \quad between\ columns,$$
$$and\ between\ the\ methods\ of\ cultivation.$$

We shall subtract 120 from each value.

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	2	1	3	2	8	4	1	9	4
y_2	4	3	2	5	14	16	9	4	25
y_3	0	-1	0	1	0	0	1	0	1
y_4	2	3	1	2	8	4	9	1	4
Total	8	6	6	10	30	24	20	14	34

N = 16, T = 30, $CF = \frac{T^2}{N} = 56.25$

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 24 + 20 + 14 + 34 - 56.25 = 35.75$$

$$SSC = \frac{1}{n}[(\sum x_1)^2 + (\sum x_2)^2 + \cdots + (\sum x_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(8)^2 + (6)^2 + (6)^2 + (10)^2] - 56.25 = 2.75$$

$$SSR = \frac{1}{n}[(\sum y_1)^2 + (\sum y_2)^2 + \cdots + (\sum y_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(8)^2 + (14)^2 + (0)^2 + (8)^2] - 56.25 = 24.75$$

To find SSK:

P	Q	R	S
1	2	3	2
2	4	3	5
0	-1	1	0
2	1	2	3
5	**6**	**9**	**10**

$$SSK = \frac{1}{n}\left[\left(\sum z_1\right)^2 + \left(\sum z_2\right)^2 + \cdots + \left(\sum z_n\right)^2\right] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(5)^2 + (6)^2 + (9)^2 + (10)^2] - 56.25 = 4.25$$

$$SSE = TSS - SSC - SSR - SSK = 35.75 - 2.75 - 24.75 = 4$$

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 2.75$	$n-1=3$	$MSC = \frac{2.75}{3} = 0.92$	$F_c = \frac{MSC}{MSE} = 1.37$	$F_c(3,6) = 4.76$
Between rows	$SSR = 24.75$	$n-1=3$	$MSR = \frac{24.75}{3} = 8.25$	$F_R = \frac{MSR}{MSE} = 12.31$	$F_R(3,6) = 4.76$
Between letters	$SSK = 4.25$	$n-1=3$	$MSK = \frac{4.25}{3} = 1.42$	$F_K = \frac{MSK}{MSE} = 2.12$	$F_K(3,6) = 4.76$
Residual (Error)	$SSE = 4$	$(n-1)(n-2) = 6$	$MSE = \frac{4}{6} = 0.67$		
Total	$TSS = 35.75$	$n^2-1=15$			

Between columns: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

Hence there is no significant difference between columns of plots in yield.

Between rows: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

So, there is significant differences between rows of plots in yield.

Between letters: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

So, there is no significant difference between the different methods of cultivation in the yield of paddy.

Example 3: A variable trial was conducted on wheat with 4 varieties in a Latin square design. The plan of the experiment and the per plot yield are given below:

C	25	B	23	A	20	D	20
A	19	D	19	C	21	B	18
B	19	A	14	D	17	C	20
D	17	C	20	B	21	A	15

Analyze data and interpret the result.

Solution:

H_0: *There is no significant difference between rows, between columns, between treatments*

H_1: *There is a significant difference between rows, between columns, between treatments*

We shall subtract 20 from each value.

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	5	3	0	0	8	25	9	0	0
y_2	-1	-1	1	-2	-3	1	1	1	4
y_3	-1	-6	-3	0	-10	1	36	9	0
y_4	-3	0	1	-5	-7	9	0	1	25
Total	0	-4	-1	-7	-12	36	46	11	29

$$TSS = \sum x_i^2 - \frac{T^2}{N} = 36 + 46 + 11 + 29 - 9 = 113$$

$$SSC = \frac{1}{n}[(\sum x_1)^2 + (\sum x_2)^2 + \cdots + (\sum x_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(0)^2 + (-4)^2 + (-1)^2 + (-7)^2] - 9 = 7.5$$

$$SSR = \frac{1}{n}[(\sum y_1)^2 + (\sum y_2)^2 + \cdots + (\sum y_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(8)^2 + (-3)^2 + (-10)^2 + (-7)^2] - 9 = 46.5$$

To find SSK:

A	B	C	D
0	3	5	0
-1	-2	1	-1
-6	-1	0	-3
-5	1	0	-3
-12	**1**	**6**	**-7**

$$SSK = \frac{1}{n}[(\sum z_1)^2 + (\sum z_2)^2 + \cdots + (\sum z_n)^2] - \frac{T^2}{N}$$

$$= \frac{1}{4}[(-12)^2 + (1)^2 + (6)^2 + (-7)^2] - 9 = 48.5$$

$$SSE = TSS - SSC - SSR - SSK = 113 - 7.5 - 46.5 - 48.5 = 10.5$$

NOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	Variance ratio	F-Table α=5%
Between columns	$SSC = 7.5$	$n-1=3$	$MSC = \frac{7.5}{3} = 2.5$	$F_c = \frac{MSC}{MSE} = 1.43$	$F_c(3,6) = 4.76$
Between rows	$SSR = 46.5$	$n-1=3$	$MSR = \frac{46.5}{3} = 15.5$	$F_R = \frac{MSR}{MSE} = 8.86$	$F_R(3,6) = 4.76$
Between letters	$SSK = 48.5$	$n-1=3$	$MSK = \frac{48.5}{3} = 16.17$	$F_K = \frac{MSK}{MSE} = 9.24$	$F_K(3,6) = 4.76$
Residual (Error)	$SSE = 10.5$	$(n-1)(n-2)=6$	$MSE = \frac{10.5}{6} = 1.75$		
Total	$TSS = 113$	$n^2-1=15$			

Conclusion:

Between columns: $F_{cal} < F_{tab}$. $\therefore H_0$ is accepted.

So, there is no significant difference between columns.

Between rows: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Between letters: $F_{cal} > F_{tab}$. $\therefore H_0$ is rejected.

Hence, we conclude that there is a significant difference between rows and between letters.

Exercise:

1. The following is a Latin square of a design, when 4 varieties of seeds are being tested. Set up the analysis of variance table and state your conclusion. You may carry out suitable change of origin and scale.

A	105	B	95	C	125	D	115
C	115	D	125	A	105	B	105
D	115	C	95	B	105	A	115
B	95	A	135	D	95	C	115

2. Analyse the following results of Latin square experiment.

Row/ Column	1	2	3	4
1	A (12)	D (20)	C (16)	B (10)
2	D (18)	A (14)	B (11)	C (14)
3	B (12)	C (15)	D (19)	A (13)
4	C (16)	B (11)	A (15)	D (20)

PRACTICE QUIZ: UNIT IV DESIGN OF EXPERIMENTS

- To determine whether the test statistic of ANOVA is statistically significant, it can be compared to a critical value. What two pieces of information are needed to determine the critical value?

    a) sample size, number of groups

    b) mean, sample standard deviation

    c) expected frequency, obtained frequency

    d) MSTR, MSE

2. Which of the following is an assumption of one-way ANOVA comparing samples from three or more experimental treatments?

    a) All the response variables within the k populations follow a normal distribution.

    b) The samples associated with each population are randomly selected and are independent from all other samples.

    c) The response variables within each of the k populations have equal variances.

    d) Any number of treatments can be used.

3. When the k population means are truly different from each other, it is likely that the average error deviation:

    a) is relatively large compared to the average treatment deviations

    b) is relatively small compared to the average treatment deviations

    c) is about equal to the average treatment deviation

    d) differ significantly between at least two of the populations

- When conducting a one-way ANOVA, the _____ the between-treatment variability is when compared to the within-treatment variability, the _____ the value of F_{DATA} will be tend to be.

    a) smaller, larger

    b) smaller, smaller

    c) larger, larger

    d) larger, more random

5. You obtained a significant test statistic when comparing three treatments in a one-way ANOVA. In words, how would you interpret the alternative hypothesis H_A?

    a) All three treatments have different effects on the mean response.

    b) Exactly two of the three treatments have the same effect on the mean response.

    c) At least two treatments are different from each other in terms of their effect on the mean response.

    d) At most two treatments are different from each other in terms of their effect on the mean response.

6. In a study, subjects are randomly assigned to one of three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are compared. The appropriate statistical test for comparing these means is:

    a) the correlation coefficient

    b) chi square

    c) the t-test

    d) the analysis of variance

- You carried out an ANOVA on a preliminary sample of data. You then collected additional data from the same groups; the difference being that the sample sizes for each group were increased by a factor of 10, and the within-group variability has decreased substantially. Which of the following statements is NOT correct?

    a) The degrees of freedom associated with the error term has increased

    b) The degrees of freedom associated with the treatment term has increased

    c) SSE has decreased

    d) F_{DATA} has changed

8. If F_{DATA} follows an F distribution with $df_1 = 4 \; and \; df_2 = 5$, what is the boundary value of F where $P(F_{DATA} < F) = 0.95$?

    a) 0.05

    b) 5.1922

    c) 6.2561

    d) 15.5291

9. Suppose the critical region for a certain test of the null hypothesis is of the form $F > 9.48773$ and the computed value of F from the data is 1.86. Then:

    a) H_0 should be rejected.

    b) The significance level is given by the area to the left of 9.48773 under the appropriate F-distribution.

    c) The significance level is given by the area to the right of 9.48773 under the appropriate F-distribution.

    d) The hypothesis test is two-tailed

10. Assuming no bias, the total variation in a response variable is due to error (unexplained variation) plus differences due to treatments (known variation). If known variation is large compared to unexplained variation, which of the following conclusions is the best?

a) There is no evidence for a difference in response due to treatments.

b) There is evidence for a difference in response due to treatments.

c) There is significant evidence for a difference in response due to treatments

d) The treatments are not comparable.

11. What would happen if instead of using an ANOVA to compare 10 groups, you performed multiple $t - tests$?

a) Nothing, there is no difference between using an ANOVA and using a $t - test.$

b) Nothing serious, except that making multiple comparisons with a *t*-test requires more computation than doing a single ANOVA.

c) Sir Ronald Fischer would be turning over in his grave; he put all that work into developing ANOVA, and you use multiple $t - tests$

d) Making multiple comparisons with a *t*-test increases the probability of making a Type I error.

12. An investigator randomly assigns 30 college students into three equal size study groups (early- morning, afternoon, late-night) to determine if the period of the day at which people study has an effect on their retention. The students live in a controlled environment for one week, on the third day of the experimental treatment is administered (study of predetermined material). On the seventh day the investigator tests for retention. In computing his ANOVA table, he sees that his MS within groups is larger than his MS between groups. What does this result indicate?

a) An error in the calculations was made.

b) There was more than the expected amount of variability between groups.

c) There was more variability between subjects within the same group than there was between groups.

d) There should have been additional controls in the experiment.

13. If the sample means for each of k treatment groups were identical (yes, this is extremely unlikely), what would be the observed value of the ANOVA test statistic?

    a) 1.0

    b) 0.0

    c) $A\ value\ between\ 0.0\ and\ 1.0$

    d) $A\ negative\ value$

14. What is the function of a post-test in ANOVA?

    a) Determine if any statistically significant group differences have occurred.

    b) Describe those groups that have reliable differences between group means.

    c) Set the critical value for the $F-test$ (or chi-square).

    d) Determine the $F-Value.$

15. Assume that there is no overlap between the box and whisker plots for three drug treatments where each drug was administered to 35 individuals. The box plots for these data:

    a) provide no evidence for, or against, the null hypothesis of ANOVA

    b) represent evidence for the null hypothesis of ANOVA

    c) represent evidence against the null hypothesis of ANOVA

    d) can be very misleading, you should not be looking at box plots in this setting

16. Which of the following statement is True?

    a) The analysis of variance helps us to test the equality of two or more sample variance.

    b) Total sum of squares is given by sum of squares due to treatment and error sum of squares.

    c) Another assumption of analysis of variance is that the distribution of the dependent variable in each population has the same mean.

    d) Analysis of variance cannot be used when there are samples of unequal size.

17. Which of the following statement is False?

a) One of the assumptions of analysis of variance is that the dependent variable is normally distributed in each of the populations being compared.

b) Analysis of variance technique originated in Agrarian research.

c) Another assumption of analysis of variance is that the distribution of the dependent variable in each population has the same mean.

d) Total variation $SST\ (or)\ TSS\ equals\ SSC + SSE$ for one-way classification.

18. Sum of the squares between the samples is given by

a) $(X_1 + X)^2 + (X_2 - X)^2 + \cdots.$

b) $(X_1 - X)^2 + (X_2 - X)^2 + \cdots.$

c) $(X_1 + X)^2 + (X_2 + X)^2 + \cdots.$

d) $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots.$

19. Which of the following statement is False?

a) For Latin Square, the treatments are grouped into replicates in two ways, once in rows and once in columns.

b) Latin Square Designs are used in industrial and medical experimentation also.

c) Latin Square arrangement is suitable only in the special cases where the land exhibits marked trends in fertility.

d) The Latin Square Design is superior to the Randomized Block Design.

20. In $4x4$ Latin square, the total of such possibilities are

a) 8

b) 24

c) 576

d) 200

Answers:

1	2	3	4	5	6	7	8	9	10
a	d	b	b	c	d	b	b	c	b
11	12	13	14	15	16	17	18	19	20
d	c	b	b	c	b	c	d	d	c

ASSIGNMENTS: UNIT IV

LEVEL-1

- The following are the numbers of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory:

Technician I	Technician II	Technician III	Technician IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$, whether the difference among the 4 sample means, can be attributed to chance.

2. Four varieties A, B, C, D of a fertilizer are tested in a randomized block design with 4 replications. The plot yields in pounds are as follows:

Column / Row	1	2	3	4
1	A (12)	D (20)	C (16)	B (10)
2	D (18)	A (14)	B (11)	C (14)
3	B (12)	C (15)	D (19)	A (13)
4	C (16)	B (11)	A (15)	D (20)

Analyse the experimental yield.

3. In an agricultural experiment conducted under Latin square design, the following observations are made. A, B, C, D represent 4 varieties of paddy, the rows represent 4 different fertilizers and the columns represent 4 different years. The figures in the table are yield in kgs per plot. Using a 0.05 LOS, test the hypothesis H_0: There is no difference in the average yield of 4 varieties of paddy.

Fertilizer Treatment	2000	2001	2002	2003
I	A 70	B 75	C 68	D 81
II	D 66	A 59	B 55	C 63
III	C 59	D 66	A 39	B 42
IV	B 41	C 57	D 39	A 55

LEVEL-2

1. The following data shows the lives in hours of four brands of electric lamps:

Brand

A: 1610 1610 1650 1680 1700 1720 1800

B: 1580 1640 1640 1700 1750

C: 1460 1550 1600 1620 1640 1660 1740 1820

D: 1510 1520 1530 1570 1600 1680

Perform an analysis of variance and test the homogeneity of the mean lives of the four brands of lamps.

2. A **completely randomized design** experiment with 10 plots and 3 treatments, gave the following results:

Plot No	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Analyze the results for treatment effects.

3. Three varieties of a crop are tested in a randomized block design with four replications, the layout being as given below: The yields are given in kilograms.

Analyze for significance

C48 A51 B52 A49

A47 B49 C52 C51

B49 C53 A49 B50

LEVEL-3

1. Four experiments determine the moisture content of samples of a powder, each observer taking a sample from each of six consignments. The assessments are given below:

Observer	Consignment					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

2. The following data resulted from an experiment to compare three burners BI,B2,B3. A Latin square design was used as the tests were made on 3 engines and were spread over 3 days.

	Engine 1	Engine 2	Engine 3
Day1	B1-16	B2-17	B3-20
Day2	B2-16	B3-21	B1-15
Day3	B3-15	B1-12	B2-13

Test the hypothesis that there is no difference between the burners.

Example 3: A variable trial was conducted on wheat with 4 varieties in a Latin square design. The plan of the experiment and the per plot yield are given below:

C	25	B	23	A	20	D	20
A	19	D	19	C	21	B	18
B	19	A	14	D	17	C	20
D	17	C	20	B	21	A	15

Analyse data and interpret the result.

LEVEL-4

- Analyse the variance in the latin square of yields in (kgs) of paddy where P, Q, R, S denote the different methods of cultivation.

S 122	P 121	R 123	Q 122
Q 124	R 123	P 122	S 125
P 120	Q 119	S 120	R 121
R 122	S 123	Q 121	P 122

Examine whether the different methods of cultivation have given significantly different yields.

2. A farmer wishes to test the effects of four different fertilizers A, B, C, D on the yield of wheat. In order to eliminate sources of error due to variability in soil fertility, he uses the fertilizers, in a Lattin square arrangement as indicated in the following table, where the numbers indicate yields in bushels per unit area.

A 18	C 21	D 25	B 11
D 22	B 12	A 15	C 19
B 15	A 20	C 23	D 24
C 22	D 21	B 10	A 17

Perform an analysis of variance to determine, if there is a significant difference between the fertilizers at $\alpha = 0.05$ level of significance.

3. A company appoints 4 salesmen A, B, C and D and observes their sales in 3 seasons – Summer, Winter and Monsoon. The figures (in lakhs of Rs.) are given in the following table.

Season	Salesmen		
	B	C	D
Summer	40	38	37
Winter	41	45	38
Monsoon	39	41	41

Carry out an analysis of variance.

LEVEL-5

1. An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1		Engine 3	Total
Detergent A	45	43	51	139
Detergent B	47	46	52	145
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Total	182	176	207	565

Perform the ANOVA and test at 0.01 level of significance whether there are differences in the detergents or in the engines.

2. The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.

Workers	Machine Type			
	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Test whether the five men differ with respect to mean productivity and test whether the mean productivity is the same for the four different machine types.

The following are the numbers of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory:

Technician I (X_1)	Technician II (X_2)	Technician III (X_3)	Technician IV (X_4)
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$, whether the differences among the 4 sample means, can be attributed to chance.

PART A QUESTIONS AND ANSWERS: UNIT IV

1. What do you understand by "Design of an experiment"? (K3, CO4)

Solution:

The design of an experiment may be defined as "the logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be well defined.

2. What are the basic principles of the design of experiments? (K3, CO4)

Solution:

The basic principles of the design of experiments are:

(i) Replication

(ii) Randomization

(iii) Local control

3. What are the basic design of experiments? (K3, CO4)

Solution:

The basic design of experiments are:

(i) Completely Randomized Design

(ii) Randomized Block Design

(iii) Latin Square Design

4. What do you mean by analysis of variance? (K3, CO4)

Solution:

Analysis of variance is a technique that analyses variances. It separates the variance ascribable to one group of causes from the variance ascribable to other groups.

5. When do you apply analysis of variance? (K3, CO4)

Solution:

When we have to test the differences between means of more than two samples we use analysis of variance.

What are the assumptions in using ANOVA? **(K3, CO4)**

Solution:

(i) The samples are drawn from normal populations.

(ii) The samples are drawn independently from these populations.

(iii) All the populations have the same variance.

7. Define the term completely randomized design. **(K1, CO4)**

Solution:

In a completely randomized design the treatments are given to the experimental units by a procedure of random allocation. It is used when the units are homogeneous.

8. State any two difference between CRD and RBD. **(K2, CO4)**

Solution:

(i) Completely randomized design analysis results in one-way classification, whereas randomized block design analysis results in two-way classification.

(ii) Experimental errors are large in CRD compared to RBD is more popular.

9. What are the advantages of a Latin square design? **(K3, CO4)**

Solution:

(i) Latin square design controls variation in two directions of the experimental material as rows and columns resulting in the reduction of experimental error.

(ii) The analysis of the design results in a three-way classification of analysis of variance.

(iii) The analysis remains relatively simple even with missing data.

10. Why a 2x2 Latin square design is not possible? **(K3, CO4)**

Solution:

Consider, a nxn Latin square design, then the degrees of freedom for SSE is $= (n-1)(n-2)$. For n=2, d.f of SSE=0 and hence, MSE is not defined.

$\therefore$ comparisons are not possible. Hence, a 2x2 Latin square design is not possible.

PART B QUESTIONS: UNIT IV

1. An experiment was performed to judge the effect of four different fuels and three different types of launchers on the range of a certain rocket. Test, on the basis of following range in miles, whether there is a significant effect due to differences in fuels and, whether there is a significant effect due to differences in launchers. Use 0.01 level of significance. (K3, CO4)

	Fuel 1	Fuel 2	Fuel 3	Fuel 4
Launcher X	45	47	48	42
Launcher Y	43	46	50	37
Launcher Z	51	52	55	49

2. A part of investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at 3 different positions on the roof. The forces required to shear each of these bolts (coded values) are as follows: (K3, CO4)

Position 1:	90	82	79	98	83	91	
Position 2:	105	89	93	104	89	95	86
Position 3:	83	89	80	94			

Perform an analysis of variance to test at the 0.05 level of significance whether the difference among the sample means at the 3 positions are significant.

3. The table shown below gives the samples got from the normal population with equal variances. Test the hypothesis that the sample mean is equal at 5% level of significance. (K3, CO4)

A	8	10	12	8	7
B	12	11	14	9	4
C	18	16	12	8	6
D	16	15	13	12	9

4. A Latin square experiment given below are the yields in quintals per acre on the paddy crop carried out for testing the effect of five fertilizers A, B, C, D, E. Analyze the data for variations. (K3, CO4)

B 25	**A 18**	**E 27**	**D 30**	**C 27**
A 19	D 31	C 29	E 26	B 23
C 28	B 22	D 33	A 18	E 27
E 28	C 26	A 20	B 25	D 33
D 32	E 25	B 23	C 28	A 20

SUPPORTIVE ONLINE CERTIFICATION COURSES

The following NPTEL and Coursera courses are the supportive online certification courses for the Unit Testing of Hypothesis

https://onlinecourses.nptel.ac.in/noc24_cs20/preview

https://onlinecourses.nptel.ac.in/noc24_mg14/preview

https://nptel.ac.in/content/storage2/courses/103106120/LectureNotes/Lec3_4.pdf

https://www.coursera.org/lecture/basic-statistics/7-01-hypotheses-N1Klj

REAL TIME APPLICATIONS

View the lecture on YouTube:

1. Introduction, Terms and Concepts of DOE with practical examples

   https://www.youtube.com/watch?v=aWhIlCOImXg

2. Introduction to Design of Experiments

   https://www.youtube.com/watch?v=pTAUa6qXV6E

3. Applications of Design of Experiments

   https://www.youtube.com/watch?v=cIXYKynq1-o

MINI PROJECT - UNIT – IV

Write a Python programme for the following problems and compute the result.

Level 1

1. A completely randomized design experiment with 10 plots and 3 treatments gave the following results:

Plot No:	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
yield	5	4	3	7	5	1	3	4	1	7

Analyze the results for treatment effects.

Level 2

2. The following Latin square of a design when 4 varieties of seeds are being tested. Set up the analysis of variance table and state your conclusion.

A 105	B 95	C 125	D 115
C 115	D 125	A 105	B 105
D 115	C 95	B 105	A 115
B 95	A 135	D 95	C 115

MINI PROJECT - UNIT – IV

Write a Python programme for the following problems and compute the result.

Level 3

The following are the numbers of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory:

Technician I (X_1)	Technician II (X_2)	Technician III (X_3)	Technician IV (X_4)
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$, whether the differences among the 4 sample means, can be attributed to chance.

Level 4

A company appoints 4 salesmen A, B, C and D and observes their sales in 3 seasons – Summer, Winter and Monsoon. The figures (in lakhs of Rs.) are given in the following table.

Season	Salesmen			
	A	B	C	D
Summer	4 5	4 0	3 8	3 7
Winter	4 3	4 1	4 5	3 8
Monsoon	3 9	3 9	4 1	4 1

MINI PROJECT - UNIT – IV

Write a Python programme for the following problems and compute the result.

Level 5

An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1	Engine 2	Engine 3	Total
Detergent A	45	43	51	139
Detergent B	47	46	52	145
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Total	182	176	207	565

Perform the ANOVA and test at 0.01 level of significance whether there are differences in the detergents or in the engines.

ASSESSMENT SCHEDULE

S. NO.	ASSESSMENT DETAILS	PROPOSED DATE
1	MCQ TEST – I	15.03.2024
2	MCQ TEST – II	16.03.2024
3	CYCLE TEST – II	18.03.2024
4	INTERNAL ASSESSMENT TEST – II	11.04.2024

PRESCRIBED TEXT BOOKS & REFERENCE BOOKS

TEXT BOOKS:

1. Grewal. B.S. and Grewal. J.S., “Numerical Methods in Engineering and Science ", 10th Edition, Khanna Publishers, New Delhi, 2015.

2. Johnson, R.A., Miller, I and Freund J., “Miller and Freund’s Probability and Statistics for

Engineers", Pearson Education, Asia, 8th Edition, 2015.

REFERENCES:

1. Burden, R.L and Faires, J.D, "Numerical Analysis”, 9th Edition, Cengage Learning, 2016.

2. Devore. J.L., "Probability and Statistics for Engineering and the Sciences”, Cengage Learning, New Delhi, 8th Edition, 2014.

3. Gerald. C.F. and Wheatley. P.O. "Applied Numerical Analysis” Pearson Education, Asia, New Delhi, 2006.

4. Spiegel. M.R., Schiller. J. and Srinivasan. R.A., "Schaum’s Outlines on Probability and Statistics ", Tata McGraw Hill Edition, 2004.

5. Walpole. R.E., Myers. R.H., Myers. S.L. and Ye. K., “Probability and Statistics for Engineers and Scientists", 8th Edition, Pearson Education, Asia, 2007.

Thank you