MODULE - V

Design of Experiments & ANOVA

Introduction

Design of Experiments (DOE) is a systematic approach for planning and conducting experiments to understand relationships between factors and responses. Analysis of Variance (ANOVA) is a statistical technique for assessing differences in means between multiple groups. DOE helps optimize processes, while ANOVA quantifies the significance of these optimizations. It's a powerful combination in scientific research, manufacturing, and quality control to make informed decisions. DOE designs experiments, and ANOVA analyzes results, enhancing decision-making across diverse fields.

Topic Learning Objectives:

Experimental Design and Hypothesis Testing: Gain proficiency in selecting appropriate experimental designs and using ANOVA for hypothesis testing to efficiently investigate the impact of factors on a response variable.

Factor Analysis and Data Interpretation: Develop the skills to identify significant factors, understand their interactions, and optimize processes or systems. Learn how to interpret and derive meaningful insights from experimental data.

Practical Application: Apply DOE and ANOVA techniques in real-world scenarios, such as quality improvement, process optimization, and scientific research, to make data-driven decisions and enhance outcomes.

5.1.2 Upon Completion of this module, students will be able to:

Effective Experimental Planning: Develop the ability to plan and execute experiments systematically, considering factors, levels, and experimental designs for efficient data collection.

Hypothesis Testing Proficiency: Gain expertise in formulating hypotheses and applying ANOVA to determine statistically significant differences among groups or treatments and understand the impact of factors.

Data Analysis and Interpretation: Learn to interpret and draw meaningful insights from experimental data, making informed decisions and improvements in quality, processes, and research based on empirical evidence.

Design and Analysis of Experiments

Planning an experiment to obtain appropriate data and drawing inference out of the data with respect to any problem under investigation is known as design and analysis of experiments. This might range anywhere from the formulations of the objectives of the experiment in clear terms to the final stage of the drafting reports incorporating the important findings of the enquiry. The structuring of the dependent and independent variables, the choice of their levels in the experiment, the type of experimental material to be used, the method of the manipulation of the variables on the experimental material, the method of recording and tabulation of data, the mode of analysis of the material, the method of drawing sound and valid inference etc. are all intermediary details that go with the design and analysis of an experiment.

Principles of Experimentation

Almost all experiments involve the three basic principles, viz., randomization, replication and local control. These three principles are, in a way, complementary to each other in trying to increase the accuracy of the experiment and to provide a valid test of significance, retaining at the same time the distinctive features of their roles in any experiment.

Randomization: Assigning the treatments or factors to be tested to the experimental units according to definite laws or probability is technically known as randomization.

This provides a basis for making a valid estimate of random fluctuations which is so essential in testing of significance of genuine differences.

Replication: Replication is the repetition of experiment under identical conditions but in the context of experimental designs, it refers to the number of distinct experimental units under the same treatment. Replication, with randomization, will provide a basis for estimating the error variance. In the absence of randomization, any amount of replication may not lead to a true estimate of error. The greater the number of replications, greater is the precision in the experiment.

Local control: Local control means the control of all factors except the ones about which we are investigating. Local control, like replication is yet another device to reduce or control the variation due to extraneous factors and increase the precision of the experiment.

Note: In short, it may be mentioned that while randomization is a method of eliminating a systematic error (*i.e.*, bias) in allocation thereby leaving only random error component of variation, the other two *viz*., replication and local control try to keep this random error as low as possible. All the three however are essential for making a valid estimate of error variance and to provide a valid test of significance.

Completely randomized design

A completely randomized design (CRD) is one where the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. For the CRD, any difference among experimental units receiving the same treatment is considered as experimental error. Hence, CRD is appropriate only for experiments with homogeneous experimental units, such as laboratory experiments, where environmental effects are relatively easy to control. For field experiments, where there is generally large variation among experimental plots in such environmental factors as soil, the CRD is rarely used.

Randomized complete block design

The randomized complete block design (RCBD) is one of the most widely used experimental designs in forestry research. The design is especially suited for field experiments where the number of treatments is not large and there exists a conspicuous factor based on which homogenous sets of experimental units can be identified. The primary distinguishing feature of the RCBD is the presence of blocks of equal size, each of which contains all the treatments.

Analysis of variance (ANOVA)

Analysis of variance (abbreviated as ANOVA) is an extremely useful technique concerning researches in the fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines. This technique is used when multiple sample cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either z-test or the t-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

The ANOVA technique is important in the context of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on. In such circumstances one generally does not want to consider all possible combinations of two populations at a time for that would require a great number of tests before we would be able to arrive at a decision. This would also consume lot of time and money, and even then, certain relationships may be left unidentified (particularly the interaction effects). Therefore, one quite often utilizes the ANOVA technique and through it investigates the differences among the means of all the populations simultaneously.

The basic principle of ANOVA

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population, it is assumed that the values of (X_{ij}) differ from the mean of this population only because of random effects i.e., there are influences on (X_{ij}) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the j^{th} population and the grand mean is attributable to what is called a ‘specific factor’ or what is technically described as treatment effect. Thus, while using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied. In short, we have to make two estimates of population variance viz., one based on between samples variance and the other based on within samples variance. Then the said two estimates of population variance are compared with F-test, wherein we work out.

$$F = \frac{Estimate\ of\ population\ variance\ based\ on\ between\ samples\ variance}{Estimate\ of\ population\ variance\ based\ on\ within\ samples\ variance}$$

This value of F is to be compared to the F-limit for given degrees of freedom. If the F value we work out is equal or exceeds* the F-limit value (see the F table) we may say that there are significant differences between the sample means.

ANOVA TECHNIQUE

Under one-way (or single factor) ANOVA, we consider only one factor and then observe that the reason for said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor. The technique involves the following steps:

(i) Obtain the mean of each sample i.e., obtain $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_k$ when there are k samples.

ii) Work out the mean of the sample means as follows:

$$\bar{\bar{X}} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3+\cdots+\bar{X}_k}{Number\ of\ samples(k)}$$

iii) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the number of items in the corresponding sample, and then obtain their total. This is known as the sum of squares for variance between the samples (or SS between). Symbolically, this can be written:

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

iv) Divide the result of the step (iii) by the degrees of freedom between the samples to obtain variance or mean square (MS) between samples.

$$MS\ between = \frac{SS\ between}{(k-1)}$$

Where $k-1$ represents degrees of freedom (d.f) between samples.

v) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain them total. This total is known as the sum of squares for variance within samples (or SS within).

$$SS\ within = \sum (X_{1i} - \bar{X}_1)^2 + \sum (X_{2i} - \bar{X}_2)^2 + \ldots + \sum (X_{ki} - \bar{X}_k)^2$$

vi) Divide the result of step (v) by the degrees of freedom within samples to obtain the variance or mean square (MS) within samples.

$$MS\ within = \frac{SS\ within}{(n-k)}$$

Where $(n-k)$ represents degrees of freedom within samples,

n=total number of items in all the samples i.e., $n_1 + n_2 + \cdots + n_k$

k= number of samples.

vii) For a check, the sum of squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means.

$$SS\ for\ total\ variance = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2, i = 1,2,3..\ \&\ j = 1,2,3\ \ldots$$

This should be equal to total of the result of the (iii) and (v) steps explained above i.e.,

$$SS\ for\ total\ variance = SS\ between + SS\ within$$

The degrees of freedom for total variance will be equal to the number of items in all samples minus one i.e., (n – 1). The degrees of freedom for between and within must add up to the degrees of freedom for total variance i.e., (n – 1) = (k – 1) + (n – k)
This fact explains the additive property of the ANOVA technique.
(viii) Finally, F-ratio may be worked out as under:

$$F - ratio = \frac{MS\ between}{MS\ within}$$

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations. For this purpose, we look into the table*, giving the values of F for given degrees of freedom at different levels of significance. If the worked-out value of F, as stated above, is less than the table value of F, the difference is taken as insignificant i.e., due to chance and the null-hypothesis of no difference between sample means stands. In case the calculated value of F happens to be either equal or more than its table value, the difference is considered as significant (which means the samples could not have come from the same universe) and accordingly the conclusion may be drawn. The higher the calculated value of F is above the table value, the more definite and surer one can be about his conclusions.

Note:

i) It should be remembered that ANOVA test is always a one-tailed test, since a low calculated value of F from the sample data would mean that the fit of the sample means to the null hypothesis (viz., $\bar{X}_1 = \bar{X}_2 = \cdots = \bar{X}_k$) is a very good fit.

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean Square (MS) (This is SS divided by d.f.) and is an estimation of variance to be used in F-ratio*	*F-ratio*
Between samples or categories	$n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$	$(k-1)$	$\frac{SS \text{ between}}{(k-1)}$	$\frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \ldots + \Sigma\left(X_{ki} - \bar{X}_k\right)^2$ $i = 1, 2, 3, \ldots$	$(n-k)$	$\frac{SS \text{ within}}{(n-k)}$	
Total	$\Sigma\left(X_{ij} - \bar{\bar{X}}\right)^2$ $i = 1, 2, \ldots$ $j = 1, 2, \ldots$	$(n-1)$		

Table 1: Analysis of Variance for one-way ANOVA Technique (there are k sample having n items)

Problems

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

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

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

$$\overline{X_1} = \frac{6+7+3+8}{4} = 6$$

$$\overline{X_2} = \frac{5+5+3+7}{4} = 5$$

$$\overline{X_3} = \frac{5+4+3+4}{4} = 4$$

Mean of Samples means, $\bar{\bar{X}} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3}{k} = \frac{6+5+4}{3} = 5$

Now we work out SS between and SS within samples:

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

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

SS within = $\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2 + \sum(X_{3i} - \bar{X}_3)^2$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$

$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$

$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$

$$= \{0 + 1 + 9 + 4\} + \{0 + 0 + 4 + 4\} + \{1 + 0 + 1 + 0\}$$

$$= 14 + 8 + 2 = 24$$

$$SS\ for\ total\ variance = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (3-5)^2$$

$$+ (7-5)^2 + (5-5)^2 + (4-5)^2 + (3-5)^2 + (4-5)^2$$

$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1 = 32$$

Alternatively, it (SS for total variance) can also be worked out as,

SS for total = SS between + SS within=8+24=32

We can now set up the ANOVA table for this problem

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

The above table shows that the calculated value of F is 1.5 which is less than the table value of 4.26 at 5% level with d.f. being v1 = 2 and v2 = 9 and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

2. Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data:

Food I	**Food II**	**Food III**
8	4	11
12	5	8
19	4	7
8	6	13
6	9	7
11	7	9

Solution: Using the same procedure as explained in problem1 we set null hypothesis as

H0: $\mu_1 = \mu_2 = \mu_3$

H1: The means are not equal

The other computed values as follows

Since, $\bar{X}_1 = 5, \bar{X}_2 = 9, \bar{X}_3 = 10$

Total mean = $\bar{X}$= 8

SSB = 6(5 – 8)2 + 6(9 – 8)2 + 6(10 – 8)2 = 84

SSE = 68

MSB = SSB/df1 = 42

MSE = SSE/df2 = 4.53

f = MSB/MSE = 42/4.53 = 9.33

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

Since f > F, the null hypothesis stands rejected

3. Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided randomly into four groups of five flashlights each. Then each group of flashlights uses a different brand of battery. The lifetimes of the batteries, to the nearest hour, are as follows.

Brand A	Brand B	Brand C	Brand D
42	28	24	20
30	36	36	32
39	31	28	38
28	32	28	28
29	27	33	25

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

Solution: H0: μ1= μ2=μ3

H1: The means are not equal
Significance level $\alpha = 0.05$
SS total = 560.2
SS between samples = 68.2
SS within samples = SS total -SS between samples=492.0

Source of variation	SS	d.f	MS	F-ratio	5% F-limit(from the F-table)
Between sample	68.2	3	22.7333	0.7393	3.24
Within sample	492.0	16	30.75		
Total	560.2	19			

At the $\alpha = 0.05$ level of significance, there is not enough evidence to conclude that the mean lifetimes of the brands of batteries differ, thus it is failed to reject the null hypothesis.

4. Data on Scholastic Aptitude Test (SAT) scores are published by the College Entrance Examination board in National College-Bound Senior. SAT scores for randomly selected students from each of four high-school rank categories are displayed in the following table.

Top Tenth	Second Tenth	Secon-fifth	Third fifth
528	514	649	372
586	457	506	440
680	521	556	495
718	370	413	321
	532	470	424
			330

Construct the one-way ANOVA table for the data. Compute SSC and SSE using the defining formulas.

Solution: $\overline{X_1} = 628.0, \overline{X_2} = 478.8, \overline{X_3} = 518.8, \overline{X_4} = 397.0, \bar{\bar{X}} = 494.1$

SS between samples (SSC) = 132508.2

SS Within samples (SSE) = 95877.6 and SS total = SSC+SSE=228385.8

Source of variation	SS	d.f	MS	F-ratio
Between sample	132508.2	3	44169.40	7.37
Within sample	95877.6	16	5992.35	
Total	560.2	19		

Exercise

1. Manufacturers of golf balls always seem to be claiming that their ball goes the farthest. A writer for a sports magazine decided to conduct an impartial test. She randomly selected 20 golf professionals and then randomly assigned four golfers to each of five brands. Each golfer drove the assigned brand of ball. The driving distances, in yards, are displayed in the following table.

Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
286	279	270	284	281
276	277	262	271	293
281	284	277	269	276
274	288	280	275	292

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. Do the data provide sufficient evidence to conclude that a difference exists in mean weekly earnings among nonsupervisory workers in the five industries? Perform the required hypothesis test using $\alpha = 0.05$.

2. The U.S. Bureau of Prisons publishes data in Statistical Report on the times served by prisoners released from federal institutions for the first time. Independent random samples of released prisoners for five different offense categories yielded the following information on time served, in months. At the 1% significance level, do the data provide sufficient evidence to conclude that a difference exists in mean time served by prisoners among the five offense groups?

Major	n_i	X_i	s_i
Counterfeiting	15	14.5	4.5
Drug Laws	17	18.4	3.8
Firearms	12	18.2	4.5
Forgery	10	15.6	3.6
Fraud	11	11.5	4.7

TWO-WAY ANOVA

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

(a) ANOVA technique in context of two-way design when repeated values are not there: As we do not have repeated values, we cannot directly compute the sum of squares within samples as we had done in the case of one-way ANOVA. Therefore, we have to calculate this residual or error variation by subtraction, once we have calculated (just on the same lines as we did in the case of one-way ANOVA) the sum of squares for total variance and for variance between varieties of one treatment as also for variance between varieties of the other treatment.
The various steps involved are as follows:

(i) Take the total of the values of individual items in all the samples and call it T.

(ii) Work out the correction factor as under correction factor = $\frac{T^2}{n}$

(iii) Find out the square of all the item values (or their coded values as the case may be) one by one and then take its total. Subtract the correction factor from this total to obtain the sum of squares of deviations for total variance. Total SS = $\sum X_{ij}^2 - \frac{T^2}{n}$

(v) Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns or (SS between columns).

(vi) Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the number of items in the corresponding row and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (or SS between rows).

(vii) Sum of squares of deviations for residual or error variance can be worked out by subtracting the result of the sum of 5^{th} and 6^{th} steps from the result of 4^{th} step stated above. In other words, Total SS – (SS between columns + SS between rows) = SS for residual or error variance.

(viii) Degrees of freedom (d.f.) can be worked out as under:

d.f. for total variance = (c . r – 1)

d.f. for variance between columns = (c – 1)

d.f. for variance between rows = (r – 1)

d.f. for residual variance = (c – 1) (r-1) where c = number or columns, r = number of rows

ix) ANOVA table cab be setup in the usual fashion as shown below

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean square (MS)*	*F-ratio*
Between columns treatment	$\sum\frac{(T_j)^2}{n_j}-\frac{(T)^2}{n}$	$(c-1)$	$\frac{SS \text{ between columns}}{(c-1)}$	$\frac{MS \text{ between columns}}{MS \text{ residual}}$
Between rows treatment	$\sum\frac{(T_i)^2}{n_i}-\frac{(T)^2}{n}$	$(r-1)$	$\frac{SS \text{ between rows}}{(r-1)}$	$\frac{MS \text{ between rows}}{MS \text{ residual}}$
Residual or error	Total *SS* – (*SS* between columns + *SS* between rows)	$(c-1)(r-1)$	$\frac{SS \text{ residual}}{(c-1)(r-1)}$	
Total	$\sum X_{ij}^2-\frac{(T)^2}{n}$	$(c.r-1)$		

Table 2: Analysis of variance for Two-way ANOVA

In the table c = number of columns, r = number of rows and SS residual = Total SS – (SS between columns + SS between rows).

Thus, MS residual or the residual variance provides the basis for the F-ratios concerning variation between columns treatment and between rows treatment. MS residual is always due to the fluctuations of sampling, and hence serves as the basis for the significance test. Both the F-ratios are compared with their corresponding table values, for given degrees of freedom at a specified level of significance, as usual and if it is found that the calculated F-ratio concerning variation between columns is equal to or greater than its table value, then the difference among columns means is considered significant. Similarly, the F-ratio concerning variation between rows can be interpreted·

(b) ANOVA technique in context of two-way design when repeated values are not there:

In case of a two-way design with repeated measurements for all of the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure we can calculate the sum of squares and degrees of freedom in the same way as we had worked out the sum of squares for variance within samples in the case of one-way ANOVA. Total SS, SS between columns and SS between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as 'interaction variation' (Interaction is the measure of inter relationship among the two different classifications). After making all these computations, ANOVA table can be set up for drawing inferences.

Problems

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

Per Acre Production Data of Wheat

(in metric tonnes)

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

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

Solution: As the given problem is a two-way design of experiment without repeated values, we shall adopt all the above stated steps.

Step (i) T=60, n=12, $\therefore \, correction\ factor = \frac{T^2}{n} = \frac{60\times60}{12} = 300$

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

Step (iii) SS between columns treatment $= \left[\frac{24\times24}{4} + \frac{20\times20}{4} + \frac{16\times16}{4}\right] - \left[\frac{60\times60}{12}\right] = 144 + 100 + 64 - 300 = 8$

Step (iv) SS between rows treatment $= \left[\frac{16\times16}{3} + \frac{16\times16}{3} + \frac{9\times9}{3} + \frac{19\times19}{3}\right] - \left[\frac{60\times60}{12}\right] = 85.33 + 85.33 + 27.00 + 120.33 - 300 = 18$

Step (v) SS residual or error = Total SS-(SS between columns + SS between rows) = 33-(8+18) = 6.

Setting up the ANOVA table

Source of variation	SS	d.f	MS	F-ratio	5% F-limit(from the F-table)
Between columns (i.e., between varieties of seeds)	8	(3-1)=2	8/2=4.00	4/1=4	F(2,6)=5.14
Between rows(i.e., between varieties of fertilizers)	18	(4-1)=3	18/3=6	6/1=6	F(3,6)=4.76
Residual or error	6	$(3-1)\times(4-1) = 6$	6/6=1		
Total	32	$(3\times4)-1 = 11$			

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

2. Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:

Amount of Blood Pressure Reduction in Millimeters of Mercury

	Drug		
	X	Y	Z
Group of People A	14	10	11
	15	9	11
B	12	7	10
	11	8	11
C	10	11	8
	11	11	7

Do the drugs act differently? Are the different groups of people affected differently? Is the interaction term significant? Justify your answer at 5% level of significance.

Solution: As the given problem is a two-way design of experiment with repeated values, we shall adopt all the above stated steps.

Step (i) T=187, n=18,

$correction\ \ factor = \frac{187\times187}{18} = 1942.7$

Step (ii) $Total\ SS = \left(14^2 + 15^2 + 12^2 + 11^2 + 10^2 + 9^2 + 7^2 + 8^2 + 11^2 + 11^2 + 11^2 + 11^2 + 10^2 + 11^2 + 8^2 + 7^2\right) - \frac{(187^2)}{18} = 2019 - 1942.72 = 76.28$

Step (iii) SS between columns (i.e., between drugs) = $\left[\frac{73\times73}{6} + \frac{56\times56}{6} + \frac{58\times58}{6}\right] - \left[\frac{187^2}{18}\right] = 888.16 + 522.66 + 560.67 - 1942.72 = 28.77$

Step (iv) SS between rows (i.e., between people) = $\left[\frac{70\times70}{6} + \frac{59\times59}{6} + \frac{58\times58}{6}\right] - \left[\frac{187^2}{18}\right] = 816.67 + 580.16 + 560.67 - 1942.72 = 14.78$

Step (v) SS within samples=$(14 - 14.5)^2 + (15 - 14.5)^2 + (10 - 9.5)^2 + (9 - 9.5)^2 + (11 - 11)^2 + (11 - 11)^2 + (12 - 11.5)^2 + (7 - 7.5)^2 + (8 - 7.5)^2 + (10 - 10.5)^2 + (11 - 10.5)^2 + (10 - 10.5)^2 + (11 - 10.5)^2 + (11 - 11)^2 + (11 - 11)^2 + (8 - 7.5)^2 + (7 - 7.5)^2 = 3.50$

Step(vi) SS for interaction variation = Total SS-(SS between columns + SS between rows) = 29.33

Setting up the ANOVA table

Source of variation	SS	d.f	MS	F-ratio	5% F-limit(from the F-table)
Between columns (i.e., between drugs)	28.77	3-1=2	28.77/2=14.385	14.385/0.389=36.9	F(2,9)=4.26
Between rows(i.e., between people)	14.78	3-1=2	14.78/2=7.390	7.390/0.389=19.0	F(2,9)=4.26

Interaction	29.33	4	29.33/4	7.308/0.389	F(4,9)=3.63
Within samples(error)	3.50	18-9=9	3.50/9=0.389		
Total	76.28	18-1=17			

(NOTE: These figures are left-over figures and have been obtained by subtracting from the column total the total of all other value in the said column. Thus, interaction SS = (76.28) – (28.77 + 14.78 + 3.50) = 29.23 and interaction degrees of freedom = (17) – (2 + 2 + 9) = 4).

The above table shows that all the three F-ratios are significant of 5% level which means that the drugs act differently, different groups of people are affected differently and the interaction term is significant. In fact, if the interaction term happens to be significant, it is pointless to talk about the differences between various treatments i.e., differences between drugs or differences between groups of people in the given case.

3. The following data show the number of worms quarantined from the GI areas of four groups of muskrats in a carbon tetrachloride anthelmintic study. Conduct a two-way ANOVA test·

I	II	III	IV
338	412	124	389
324	387	353	432
268	400	469	255
147	233	222	133
309	212	111	265

Solution: Using the same procedure as explained in problem1

Source of variation	Sum of Squares	Degrees of freedom	Mean Square	Ratio F
Between the groups	62111.6	8	9078.067	F = MST / MSE = 9.4062 / 3.66 **F = 2.57**
Within the groups	98787.8	16	4567.89	
Total	167771.4	24		

Since F = MST / MSE

= 9.4062 / 3.66 **F = 2.57**

Exercise:

1. The following data represents the number of units of tablet production (in thousands) per day by five different technicians by using four different types of machines.

Workers	A	B	C	D
P	54	48	57	46
Q	56	50	62	53
R	44	46	54	42
S	53	48	56	44
T	48	52	59	48

a) Test whether the mean productivity of different machines is same?
b) Test whether the 5 technicians differ with respect to mean productivity?

ANOVA IN LATIN-SQUARE DESIGN

Latin-square design is an experimental design used frequently in agricultural research. In such a design the treatments are so allocated among the plots that no treatment occurs, more than once in any one row or any one column. The ANOVA technique in case of Latin-square design remains more or less the same as we have already stated in case of a two-way design, except the fact that the variance is split into four parts as under:
(i) variance between columns;
(ii) variance between rows;
(iii) variance between varieties;
(iv) residual variance.
All these above stated variances are worked out as under:

Variance between columns or *MS* between columns	$= \dfrac{\sum \dfrac{(T_j)^2}{n_j} - \dfrac{(T)^2}{n}}{(c-1)} = \dfrac{SS \text{ between columns}}{\text{d.f.}}$
Variance between rows or *MS* between rows	$= \dfrac{\sum \dfrac{(T_i)^2}{n_i} - \dfrac{(T)^2}{n}}{(r-1)} = \dfrac{SS \text{ between rows}}{\text{d.f.}}$
Variance between varieties or *MS* between varieties	$= \dfrac{\sum \dfrac{(T_v)^2}{n_v} - \dfrac{(T)^2}{n}}{(v-1)} = \dfrac{SS \text{ between varieties}}{\text{d.f.}}$
Residual or error variance or *MS* residual	$= \dfrac{\text{Total } SS - (SS \text{ between columns} + SS \text{ between rows} + SS \text{ between varieties})}{(c-1)(c-2)^{*}}$
where total	$SS = \sum (x_{ij})^2 - \dfrac{(T)^2}{n}$ c = number of columns r = number of rows v = number of varieties

Note:

1. CODING METHOD: Coding method is furtherance of the short-cut method. This is based on an important property of F-ratio that its value does not change if all the n item values are either multiplied or divided by a common figure or if a common figure is either added or subtracted from each of the given n item values. Through this method big figures are reduced in magnitude by division or subtraction and computation work is simplified without any disturbance on the F-ratio. This method should be used specially when given figures are big or otherwise inconvenient. Once the given figures are converted with the help of some common figure, then all the steps of the short-cut method stated above can be adopted for obtaining and interpreting F-ratio.

2. In place of c we can as well write r or v since in Latin-square design c = r = v

Problems

1. Analyze and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat viz., A, B, C and D under a Latin-square design

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

Solution: Using the coding method, we subtract 20 from the figures given in each of the small squares and obtain the coded figures as under:

		Columns				Row totals
		1	2	3	4	
	1	*C* 5	*B* 3	*A* 0	*D* 0	8
Rows	2	*A* –1	*D* –1	*C* 1	*B* –2	–2
	3	*B* –1	*A* –6	*D* –3	*C* 0	–10
	4	*D* –3	*C* 0	*B* 1	*A* –5	–7
Column totals		0	–4	–1	–7	*T* = –12

Squaring these coded figures in various columns and rows we have:

		Squares of coded figures				Sum of squares
		Columns				
		1	2	3	4	
Rows	1	*C* 25	*B* 9	*A* 0	*D* 0	34
	2	*A* 1	*D* 1	*C* 1	*B* 4	7
	3	*B* 1	*A* 36	*D* 9	*C* 0	46
	4	*D* 9	*C* 0	*B* 1	*A* 25	35
Sum of squares		36	46	11	29	*T* = 122

Step (i) T=60, n=12, $\therefore\ correction\ factor = \frac{T^2}{n} = \frac{(-12)\times(-12)}{16} = 9$

Step (ii) $Total\ SS = \sum(X_{ij})^2 - \frac{T^2}{n} = 122 - 9 = 113$

Step (iii) SS between columns = $\sum\frac{(T_j)^2}{n_j} - \frac{T^2}{n} = \frac{66}{4} - 9 = 7.5$

Step (iv) SS between rows treatment = $\sum\frac{(T_i)^2}{n_i} - \frac{T^2}{n} = \frac{222}{4} - 9 = 46.5$

Step (v) For finding SS for variance between varieties, we would first rearrange the coded data in the following form:

Varieties of wheat	Yield in different parts of field				Total (T_v)
	I	II	III	IV	
A	–1	–6	0	–5	–12
B	–1	3	1	–2	1
C	5	0	1	0	6
D	–3	–1	–3	0	–7

Now we can work out SS for variance between varieties as under:

SS for variance between varieties=$\sum\frac{T_v^2}{n_v} - \frac{T^2}{n} = \left\{\frac{(-12)^2}{4} + \frac{1^2}{4} + \frac{6^2}{4} + \frac{(-7)^2}{4}\right\} - 9 = 48.5$

Sum of square of residual variance will work out to, 113-(7.5+46.5+48.5) =10.50

d.f. for variance between columns = (c – 1) = (4 – 1) = 3

d.f. for variance between rows = (r – 1) = (4 – 1) = 3

d.f. for variance between varieties = (v – 1) = (4 – 1) = 3

d.f. for total variance = (n – 1) = (16 – 1) = 15

d.f. for residual variance = (c – 1) (c – 2) = (4 – 1) (4 – 2) = 6

ANOVA table in Latin-Square design can now be set up as shown below

Source of variation	SS	d.f	MS	F-ratio	5% F-limit(from the F-table)
Between columns	7.50	3	7.50/3=2.50	2.50/1.75=1.43	F(3,6)=4.76
Between rows	46.50	3	46.50/3=15.50	15.50/1.75=8.85	F(3,6)=4.76
Between varieties	48.50	3	48.50=16.17	16.17/1.75=9.24	F(3,6)=4.76
Residual or error	10.50	6	10.50/6=1.75		
Total	113.00	15			

The above table shows that variance between rows and variance between varieties are significant and not due to chance factor at 5% level of significance as the calculated values of the said two variances are 8.85 and 9.24 respectively which are greater than the table value of 4.76. But variance between columns is insignificant and is due to chance because the calculated value of 1.43 is less than the table value of 4.76.

2. Below are given the plan and yield in kgs/plot of a 5x5 Latin square experiment on the wheat crop carried out for testing the effects of five, manorial treatments A, B, C, D, and E. 'A' denotes control.

B 15	A 8	E 17	D 20	C 17	R1 = 77
A 9	D 21	C 19	E 16	B 13	R2 = 78
C 18	B 12	D 23	A 8	E 17	R3 = 78
E 18	C 16	A 10	B 15	D 23	R4 = 82
D 22	E 15	B 13	C 18	A 10	R5 = 78

$C_1 = 82, C_2 = 72, C_3 = 82, C_4 = 77, C_5 = 80$; $GT = 393$

Analyze the data and state your conclusions.

Solution:

Step (i) $\therefore\ correction\ factor = \frac{T^2}{n} = 6177.96$

Step (ii) $Total\ SS = \sum(X_{ij})^2 - \frac{T^2}{n} = 483.04$

Step (iii) SS between columns (SSC) = $\sum \frac{(T_j)^2}{n_j} - \frac{T^2}{n} = 14.24$

Step (iv) SS between rows treatment (SSR) = $\sum \frac{(T_i)^2}{n_i} - \frac{T^2}{n} = 3.04$

Step (v) To get SS due to treatments, first find the totals for each treatment using the given data as follows:

Treatment (A)	B	C	D	E
8	15	17	20	17
9	13	19	21	16
8	12	18	23	17
10	15	16	23	18
10	13	18	22	15
$T_1 = 45$ 83	$T_2 = 68$	$T_3 = 88$	$T_4 = 109$	$T_5 =$

SS for variance between treatments=$\sum \frac{T_v^2}{n_v} - \frac{T^2}{n} = 454.64$

Sum of square of residual variance(error)= TSS-SSR-SSC-SST=11.12

ANOVA table in Latin-Square design can now be set up as shown below

Source of variation	SS	d.f	MS	F-ratio	5% F-limit(from the F-table)
Between columns	4	3.04	0.76		3.26
Between rows	4	14.24	3.56		5.41
Between varieties	4	454.24	113.66	123.34	
Residual or error	12	11.12	0.92		
Total	24	484.04			

The observed highly significant value of the variance ratio indicates that there are significant differences between the treatment means. S.E. of the difference between the treatment means is given by $SED = \sqrt{2 \times \frac{EMS}{r}} = \sqrt{2 \times \frac{0.92}{5}} = 0.61$ and Critical difference=SED * t 5% at df=1.33

Summary of results

Treatment means will be calculated from the original table on treatment totals.

Treatments	A	B	C	D	E	CD 5%
Mean yield in Kgs/plot	9.0	13.6	17.6	21.8	16.6	1.33

The treatment has been compared by setting them in the descending order of their yields.

Treatments	D	C	E	B	A	CD 5%

Mean yield in Kgs/plot	21.8	17.6	16.6	13.6	9.0	1.33

The treatment 'D' is the best of all. The treatments 'C' and 'E' do not differ significantly each other. The yield obtained by applying every one of the manorial treatments is significantly higher that obtained without applying any manure.

Analysis of co-variance (ANOCOVA)

The object of experimental design in general happens to be to ensure that the results observed may be attributed to the treatment variable and to no other causal circumstances. For instance, the researcher studying one independent variable, X, may wish to control the influence of some uncontrolled variable (sometimes called the covariate or the concomitant variables), Z, which is known to be correlated with the dependent variable, Y, then he should use the technique of analysis of covariance for a valid evaluation of the outcome of the experiment. "In psychology and education primary interest in the analysis of covariance rests in its use as a procedure for the statistical control of an uncontrolled variable.

While applying the **ANOCOVA** technique, the influence of uncontrolled variable is usually removed by simple linear regression method and the residual sums of squares are used to provide variance estimates which in turn are used to make tests of significance. In other words, covariance analysis consists in subtracting from each individual score (Yi) that portion of it Yi´ that is predictable from uncontrolled variable (Zi) and then computing the usual analysis of variance on the resulting (Y – Y´)'s, of course making the due adjustment to the degrees of freedom because of the fact that estimation using regression method required loss of degrees of freedom.

ASSUMPTIONS IN ANOCOVA

The ANOCOVA technique requires one to assume that there is some sort of relationship between the dependent variable and the uncontrolled variable. We also assume that this form of relationship is the same in the various treatment groups. Other assumptions are:
(i) Various treatment groups are selected at random from the population.
(ii) The groups are homogeneous in variability.
(iii) The regression is linear and is same from group to group.

Problems

1. The following are paired observations for three experimental groups:

Group I		*Group II*		*Group III*	
X	*Y*	*X*	*Y*	*X*	*Y*
7	2	15	8	30	15
6	5	24	12	35	16
9	7	25	15	32	20
15	9	19	18	38	24
12	10	31	19	40	30

Y is the covariate (or concomitant) variable. Calculate the adjusted total, within groups and between groups, sums of squares on X and test the significance of differences between the adjusted means on X by using the appropriate F-ratio. Also calculate the adjusted means on X.

Solution: We apply the technique of analysis of covariance and work out the related measures as

	Group I		Group II		Group III	
	X	Y	X	Y	X	Y
	7	2	15	8	30	15
	6	5	24	12	35	16
	9	7	25	15	32	20
	15	9	19	18	38	24
	12	10	31	19	40	30
Total	49	33	114	72	175	105
Mean	9.80	6.60	22.80	14.40	35.00	21.00

Correction factor for $X = \frac{(\sum X)^2}{N} = 7616.27$ and

$\sum Y = 33 + 72 + 105 = 210$ and correction factor for Y= $\frac{(\sum Y)^2}{N} = 2940$

$\sum X^2 = 9476, \sum Y^2 = 3734, \sum XY = 5838$ and correction factor for XY= $\frac{\sum X \sum Y}{N} = 4732$

Hence, total SS for X = X^2 - correction factor for X= 9476 – 7616.27 = 1859.73

SS between for X = $\left\{\frac{49^2}{5} + \frac{114^2}{5} + \frac{175^2}{5}\right\} - \{correction\ factor\ for\ X\}$=480.2+2599.2+6125-7671.27=1588.13

SS within X = (total SS for X) - (SS between for X) = 1859.73-1588.13 = 271.60

Similarly, we work out the following values in respect of Y

Total SS for Y= Y^2 - correction factor for Y =3734-2940=794

SS between for Y =$\left\{\frac{33^2}{5} + \frac{72^2}{5} + \frac{105^2}{5}\right\}$ −correction factor for Y=519.6

SS within for Y = (total SS for Y) – (SS between for Y) = (794) – (519.6) = 274.4

Then, we work out the following values in respect of both X and Y

Total sum of product of XY $= \sum XY - correction\ factor\ for\ XY = 5838 - 4732 = 1106$

SS between for XY = $\left\{\frac{49\times33}{5} + \frac{114\times72}{5} + \frac{175\times105}{5}\right\} - correction\ factor\ for\ XY$

= (323.4 + 1641.6 + 3675) – (4732) = 908

SS within for XY = (Total sum of product) – (SS between for XY) = (1106) – (908) = 198

ANOVA table for X, Y and XY can now be set up as shown below

Source	d.f.	SS for X	SS for Y	Sum of product XY
Between groups	2	1588.13	519.60	908
Within groups	12	E_{XX} 271.60	E_{YY} 274.40	E_{XY} 198
Total	14	T_{XX} 1859.73	T_{YY} 794.00	T_{XY} 1106

Adjusted total SS= $T_{XX} - \frac{T_{XY}^2}{T_{YY}} = 1859.73 - \frac{1106^2}{794} = 1859.73 - 1540.60 = 319.13$

Adjusted SS within group= $E_{XX} - \frac{E_{XY}^2}{E_{YY}} = 271.60 - \frac{198^2}{274.40} = 128.73$

Adjusted SS between groups= (adjusted total SS) – (Adjusted SS within group) = (319.13 – 128.73) = 190.40

ANOVA table for adjusted X

Source	*d.f.*	*SS*	*MS*	*F-ratio*
Between groups	2	190.40	95.2	8.14
Within group	11	128.73	11.7	
Total	13	319.13		

At 5% level, the table value of F for v1 = 2 and v2 = 11 is 3.98 and at 1% level the table value of F is 7.21. Both these values are less than the calculated value (i.e., calculated value of 8.14 is greater than table values) and accordingly we infer that F-ratio is significant at both levels which means the difference in group means is significant.

Adjusted means on X will be worked out as follows:

Regression coefficient for X on Y i.e.,$b = \frac{Sum\ of\ product\ within\ group}{Sum\ of\ squares\ within\ groups\ for\ Y} = \frac{198}{274.40} = 0.7216$

Deviation of initial group means from general mean (= 14) in case of Y		*Final means of groups in X (unadjusted)*
Group I	–7.40	9.80
Group II	0.40	22.80
Group III	7.00	35.00

Adjusted means of groups in X = (Final mean) – b (deviation of initial mean from general mean in case of Y). Hence,

Adjusted mean for Group I = (9.80) – 0.7216 (–7.4) = 15.14

Adjusted mean for Group II = (22.80) – 0.7216 (0.40) = 22.51

Adjusted mean for Group III = (35.00) – 0.7216 (7.00) = 29.95

Video Links:

1. Design of experiments
2. Basic of ANOVA
3. Problems on one-way ANOVA
4. Problems on two-way ANOVA
5. Problems on two-way ANOVA
6. Problems on ANOVA for Latin-Square design
7. Problems on ANOCOVA

Disclaimer: The content provided is prepared by department of Mathematics for the specified syllabus by using reference books mentioned in the syllabus. This material is specifically for the use of RVITM students and for education purpose only.