Lecture No 43:

- Analysis of Variance
- Experimental Design

Earlier, we compared two-population means by using a two-sample t-test.

However, we are often required to compare more than two population means simultaneously.

We might be tempted to apply the two-sample t-test to all possible pairwise comparisons of means. For example, if we wish to compare 4 population means, there will be $\binom{4}{2} = 6$ separate pairs, and to test the null hypothesis that all four population means are equal, we would require six two-sample t-tests.

Similarly, to test the null hypothesis that 10 population means are equal, we would need

$$\binom{10}{2} = 45$$

separate two-sample t-tests.

This procedure of running multiple two-sample t-tests for comparing means would obviously be tedious and time-consuming. Thus a series of two-sample t-tests is not an appropriate procedure to test the equality of several means simultaneously. Evidently, we require a simpler procedure for carrying out this kind of a test.

One such procedure is the Analysis of Variance, introduced by Sir R.A. Fisher (1890-1962) in 1923:

Analysis of Variance (ANOVA):

It is a procedure which enables us to test the hypothesis of equality of several population means

(i.e.

$H0 : \mu1 = \mu2 = \mu3 = \ldots\ldots = \mu k$

against

HA: not all the means are equal)

The concept of Analysis of Variance is closely related with the concept of Experimental Design:

EXPERIMENTAL DESIGN:

By an experimental design, we mean a plan used to collect the data relevant to the problem under study in such a way as to provide a basis for valid and objective inference about the stated problem.

The plan usually includes:

- the selection of treatments whose effects are to be studied,
- the specification of the experimental layout, and
- the assignment of treatments to the experimental units.

All these steps are accomplished before any experiment is performed. Experimental Design is a very vast area. In this course, we will be presenting only a very basic introduction of this area. There are two types of designs: **systematic and randomized designs.**

In this course, we will be discussing only the randomized designs, and, in this regard, it should be noted that for the randomized designs, the analysis of the collected data is carried out through the technique known as Analysis of Variance.

Two of the very basic randomized designs are:

i) The Completely Randomized (CR) Design,

and

ii) The Randomized Complete Block (RCB) Design

and we will consider these one by one.

We begin with the simplest design i.e. the Completely Randomized (CR) Design:

THE COMPLETELY RANDOMIZED DESIGN (CR DESIGN):

A completely randomized (CR) design, which is the simplest type of the basic designs, may be defined as a design in which the treatments are assigned to experimental units completely at random, i.e. the randomization is done without any restrictions.

This design is applicable in that situation where the entire experimental material is homogeneous (i.e. all the experimental units can be regarded as being similar to each other).

We illustrate the concept of the Completely Randomized (CR) Design (pertaining to the case when each treatment is repeated equal number of times) with the help of the following example:

EXAMPLE:

An experiment was conducted to compare the yields of three varieties of potatoes. Each variety as assigned at random to equal-size plots, four times.

The yields were as follow:

Variety		
A	**B**	**C**
23	**18**	**16**
26	**28**	**25**
20	**17**	**12**
17	**21**	**14**

Test the hypothesis that the three varieties of potatoes are not different in the yielding capabilities.

SOLUTION:

The first thing to note is that this is an example of the Completely Randomized (CR) Design. We are assuming that all twelve of the plots (i.e. farms) available to us for this experiment are homogeneous (i.e. similar) with regard to the fertility of the soil, the weather conditions, etc., and hence, we are assigning the four varieties to the twelve plots totally at random. Now, in order to test the hypothesis that the mean yields of the three varieties of potato are equal, we carry out the six-step hypothesis-testing procedure, as given below:

Hypothesis-Testing Procedure:

i) H0 : $\mu A = \mu B = \mu C$
HA : Not all the three means are equal

ii) Level of Significance:
$\alpha = 0.05$

iii) Test Statistic:

$$F = \frac{MS\ Treatments}{MS\ Error}$$

which, if H0 is true, has an F distribution with $\upsilon 1 = k-1 = 3 - 1 = 2$ and $\upsilon 2 = n-k = 12 - 3 = 9$ degree of freedom

iv) Computations:

The computation of the test statistic presented above involves quite a few steps, including the formation of what is known as the ANOVA Table.

First of all, let us consider what is meant by the ANOVA Table (i.e. the Analysis of Variance Table).

In the case of the Completely Randomized (CR) Design, the ANOVA Table is a table of the type given below:

ANOVA Table in the case of the Completely Randomized (CR) Design.

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between treatments	k-1	SST	MST	MST/MSE
Within treatments (Error)	n-k	SSE	MSE	--
Total	n-1	TSS	--	--

Let us try to understand this table step by step:

The very first column is headed 'Source of Variation', and under this heading, we have three distinct sources of variation:

'Total' stands for the overall variation in the twelve values that we have in our data-set.

Variety		
A	**B**	**C**
23	**18**	**16**
26	**28**	**25**
20	**17**	**12**
17	**21**	**14**

As you can see, the values in our data-set are 23, 26, 20, 17, 18, 28, and so on. Evidently, there is a variation in these values, and the term 'Total' in the lowest row of the ANOVA Table stands for this overall variation.
The term 'Variation between Treatments' stands for the variability that exists between the three varieties of potato that we have sown in the plots.
(In this example, the term 'treatments' stands for the three varieties of potato that we are trying to compare.)
(The term 'variation between treatments' points to the fact that:
It is possible that the three varieties or, at least two of the varieties are significantly different from each other with regard to their yielding capabilities. This variability between the varieties can be measured by measuring the differences between the mean yields of the three varieties.)
The third source of variation is 'variation within treatments'. This points to the fact that even if only one particular variety of potato is sown more than once, we do not get the same yield every time.

Variety		
A	**B**	**C**
23	**18**	**16**
26	**28**	**25**
20	**17**	**12**
17	**21**	**14**

In this example, variety A was sown four times, and the yields were 23, 26, 20, and 17 --- all different from one another! Similar is the case for variety B as well as variety C.The variability in the yields of variety A can be called 'variation within variety A'.

Similarly, the variability in the yields of variety B can be called 'variation within variety B'.Also, the variability in the yields of variety C can be called 'variation within variety C'. We can say that the term 'variability within treatments' stands for the combined effect of the above-mentioned three variations. The 'variation within treatments' is also known as the 'error variation'.

This is so because we can argue that if we are sowing the same variety in four plots which are very similar to each other, then we should have obtained the same yield from each plot!

If it is not coming out to be the same every time, we can regard this as some kind of an 'error'.

The second, third and fourth columns of the ANOVA Table are entitled 'degrees of freedom', 'Sum of Squares' and 'Mean Square'.

ANOVA Table in the case of the Completely Randomized (CR) Design.

Source of Variation	d.f.	Sum of Squares	Mean Square	F
Between treatments	k-1	SST	MST	MST/MSE
Within treatments (Error)	n-k	SSE	MSE	--
Total	n-1	TSS	--	--

The point to understand is that the sources of variation corresponding to treatments and error will be measured by computing quantities that are called Mean Squares, and 'Mean Square' can be defined as:

$$Mean\ Square = \frac{Sum\ of\ Squares}{Degrees\ of\ Freedom}$$

Corresponding to these two sources of variation, we have the following two equations:

$$1)\ 'MS\ Treatment' = \frac{'SS\ Treatment'}{d.f.}$$

AND

$$2)\ 'MS\ Error' = \frac{'SS\ Error'}{d.f.}$$

It has been mathematically proved that, with reference to Analysis of Variance pertaining to the Completely Randomized (CR) Design, the degrees of freedom corresponding to the Treatment Sum of Squares are k-1, and the degrees of freedom corresponding to the Error Sum of Squares are n-k.

Hence, the above two equations can be written as:

$$1)\ 'MS\ Treatment' = \frac{'SS\ Treatment'}{k-1}$$

AND

$$2)\ 'MS\ Error' = \frac{'SS\ Error'}{n-k}$$

How do we compute the various sums of squares?

The three sums of squares occurring in the third column of the above ANOVA Table are given by:

$$1)\ Total\ SS = TSS = \sum_i \sum_j X_{ij}^2 - C.F.$$

$$2)\ SS\ Treatment = SST = \frac{\sum_j T_{.j}^2}{r} - C.F.$$

where C.F. stands for 'Correction Factor', and is given by

$$C.F. = \frac{T_{..}^2}{n}$$

and r denotes the number of data-values per column (i.e. the number of rows).

(It should be noted that this example pertains to that case of the Completely Randomized (CR) Design where each treatment is being repeated equal number of times, and the above formulae pertain to this particular situation.With reference to the CR Design, it should be noted that, in some situations, the various treatments are not repeated an equal number of times.

For example, with reference to the twelve plots (farms) that we have been considering above, we could have sown variety A in five of the plots, variety B in three plots, and variety C in four plots.

Going back to the formulae of various sums of squares, the sum of squares for error is given by

$$3)\ SS\ Error = Total\ SS - SS\ Treatment$$

i.e.

$$SSE = TSS - SST$$

It is interesting to note that,

Total SS = SS Treatment + SS Error

In a similar way, we have the equation:

Total d.f. = d.f. for Treatment + d.f. for Error

It can be shown that the degrees of freedom pertaining to 'Total' are n - 1.

Now,

n-1 = (k-1) + (n-k)

i.e.

Total d.f. = d.f. for Treatment + d.f. for Error

The notations and terminology given in the above equations relate to the following table:

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{\cdot j}$	86	84	67	237	4953
$T_{\cdot j}^2$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

The entries in the body of the table i.e. 23, 26, 20, 17, and so on are the yields of the three varieties of potato that we had sown in the twelve farms.

The entries written in brackets next to the above-mentioned data-values are the squares of those values.

For example:

529 is the square of 23,
676 is the square of 26,
400 is the square of 20,
and so on.

Adding all these squares, we obtain :

$$\sum_i \sum_j X_{ij}^2 = 4953$$

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{\cdot j}$	86	84	67	237	4953
$T_{\cdot j}^2$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

The notation T.j stands for the total of the jth column.(You must already be aware that, in general, the rows of a bivariate table are denoted by the letter ‘i’, whereas the columns of a bivariate table are denoted by the letter ‘j’.

In other words, we talk about the ‘ith row’, and the ‘jth column’ of a bivariate table.)The ‘dot’ in the notation T.j indicates the fact that summation has been carried out over i (i.e. over the rows).

In this example, the total of the values in the first column is 86, the total of the values in the second column is 84, and the total of the values in the third column is 67.

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{.j}$	86	84	67	237	4953
$T_{.j^2}$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

Hence, ΣT.j is equal to 237.

ΣT.j is also denoted by T..

i.e.

T.. = ΣT.j

The 'double dot' in the notation T.. indicates that summation has been carried out over i as well as over j.

The row below T.j is that of T.j2, and squaring the three values of T.j, we obtain the quantities 7396, 7056 and 4489.

Adding these, we obtain ΣT.j2 = 18941.

	Variety			Total	$\sum_j X_{ij}^2$
	A	B	C		
	23 (529) 26 (676) 20 (400) 17 (289)	18 (324) 28 (784) 17 (289) 21 (196)	16 (256) 25 (625) 12 (144) 14 (196)	-- -- -- --	1109 2085 833 926
$T_{.j}$	86	84	67	237	4953
$T_{.j^2}$	7396	7056	4489	18941	↑ | Check ←
$\sum_i X_{ij}^2$	1894	1838	1221	4953	

Now that we have obtained all the required quantities, we are ready to compute SS Total, SS Treatment, and SS Error: We have

$$C.F. = \frac{T_{..}^2}{n} = \frac{(237)^2}{12} = 4680.75$$

Hence, the total sum of squares is given by

$$\begin{aligned} TSS &= \sum_i \sum_j X_{ij}^2 - C.F. \\ &= 4953 - 4680.75 \\ &= 272.25 \end{aligned}$$

Also, we have

$$SS\ Treatment = SST = \frac{\sum_j T_{.j}^2}{r} - C.F.$$

$$= \frac{18941}{4} - 4680.75$$

$$= 54.50$$

And, hence:

SS Error = SSE = TSS - SST

= 272.25 - 54.50 = 217.75

ALSO

In this example, we have n = 12, and k = 3, hence:

n-1=11,

k- 1 = 2,

and

n - k = 9.

Substituting the above sums of squares and degree of freedom in the ANOVA table, we obtain:

ANOVA TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50		
Error	9	217.75		
Total	11	272.25		

Now, the mean squares for treatments and for error are very easily found by dividing the sums of squares by the corresponding degrees of freedom.

Hence, we have

ANOVA TABLE;

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50	27.25	
Error	9	217.75	24.19	
Total	11	272.25	--	

As indicated earlier, the test-statistic appropriate for testing the null hypothesis

H0 : $\mu A = \mu B = \mu C$

versus

HA : Not all the three means are equal is:

$$F = \frac{MS\ Treatments}{MS\ Error}$$

which, if H0 is true, has an F distribution with υ1 = k-1 = 3 – 1 = 2 and υ2 = n-k = 12 – 3 = 9 degree of freedom Hence, it is obvious that F will be found by dividing the first entry of the fourth column of our ANOVA Table by the second entry of the same column i.e.

$$F = \frac{\text{MS Treatment}}{\text{MS Error}} = \frac{27.25}{24.19} = 1.13$$

We insert this computed value of F in the last column of our ANOVA table, and thus obtain:
ANOVA TABLE

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between treatments (i.e. Between varieties)	2	54.50	27.25	1.13
Error	9	217.75	24.19	--
Total	11	272.25	--	--

The fifth step of the hypothesis - testing procedure is to determine the critical region.
With reference to the Analysis of Variance procedure, it can be shown that it is appropriate to establish the critical region in such a way that our test is a right-tailed test. In other words, the critical region is given by:
Critical Region:

$$F > F\alpha\ (k - 1, n - k)$$

In this example:
The critical region is F > F0.05 (2,9) = 4.26
vi) Conclusion:

Since the computed value of F = 1.13 does not fall in the critical region, so we accept our null hypothesis and may conclude that, on the average, there is no difference among the yielding capabilities of the three varieties of potatoes.

In this course, we will not be discussing the details of the mathematical points underlying One-Way Analysis of Variance that is applicable in the case of the Completely Randomized (CR) Design. One important point that the students should note is that the ANOVA technique being presented here is valid under the following assumptions:
1. The k populations (whose means are to be compared) are normally distributed;
2. All k populations have equal variances i.e. σ12 = σ22 = … = σk2. (This property is called homoscedasticity.)
3. The k samples have been drawn randomly and independently from the respective populations.
Next, we begin the discussion of the Randomized Complete Block (RCB) Design:

The Randomized Complete Block Design (RCB Design):

A randomized complete block (RCB) design is the one in which:
i) The experimental material (which is not homogeneous overall) is divided into groups or blocks in such a manner that the experimental units within a particular block are relatively homogeneous.
ii) Each block contains a complete set of treatments, i.e., it constitutes a replication of treatments.
And
iii) The treatments are allocated at random to the experimental units within each block, which means the randomization is restricted.(A new randomization is made for every block.)The object of this type of arrangement is to bring the variability of the experimental material under control.

In simple words, the situation is as follows:

We have experimental material which is not homogeneous overall. For example, with reference to the example that we have been considering above, suppose that the plots which are closer to a canal are the most fertile ones, the ones a little further away are a little less fertile, and the ones still further away are the least fertile.

In such a situation, we divide the experimental material into groups or blocks which are relatively homogeneous. The randomized complete block design is perhaps the most widely used experimental design. Two-way analysis of variance is applicable in case of the randomized complete block (RCB) design.
We illustrate this concept with the help of an example:

EXAMPLE:

In a feeding experiment of some animals, four types of rations were given to the animals that were in five groups of four each.

The following results were obtained:

Groups	Rations			
	A	B	C	D
I	32.3	33.3	30.8	29.3
II	34.0	33.0	34.3	26.0
III	34.3	36.3	35.3	29.8
IV	35.0	36.8	32.3	28.0
V	36.5	34.5	35.8	28.8

The values in the above table represents the gains in weights in pounds.
Perform an analysis of variance and state your conclusions.
In the next lecture, we will discuss this example in detail, and will analyze the given data to carry out the following test:
H0 : $\mu A = \mu B = \mu C = \mu D$
HA : Not all the treatment-means are equal