III B.Sc. Statistics

Subject name: Design of Experiment

Subject code :CST62

Unit :2

7/2

UNIT-II

Analysis of variance [ANOVA]

Introduction:

Analysis of variance is a powerful statistical tool for test of significance. The basic purpose of the analysis of variance is to test the homogenity of several means.

Definition:

Analysis of variance is the separation of variance ascaibable to one group & causes from the variance ascaibable to other group.

Assumption:

1) The observations are independent.
2) Parent population from which observations are taken is normal.
3) Various treatments and environment effective are additive in nature.

Applications:

ANOVA techniques is now frequently apply in testing the linearity of the fitted regression line or the significance of the correlation ratio i r.

Types of ANOVA:

1) One - way classification

2) Two - way classification

Cochran's theorem: (Statement)

Let $x_1, x_2, \ldots, x_n$ denote a random sample from normal population $N(0, \sigma^2)$ Let the sum of the squares of these values be written in the form,

$$\sum_{i=1}^{n} x_i^2 = Q_1 + Q_2 + \cdots + Q_k$$

where Q_j is a quadratic form in $x_1, x_2, \ldots, x_n$ with rank r_j, $(j = 1, 2, \ldots, k)$ then the random variables $Q_1, Q_2, \ldots, Q_k$ are mutually independent and Q_j/σ^2 is chi-square variate with r_j degrees of freedom if and only if $\sum_{j=1}^{k} r_j = n$

One - way classification:

Let us suppose that N observations x_{ij} $(i = 1, 2, \ldots, k)$ $(j = 1, 2, \ldots, n_i)$ of a random variable X are grouped on some basis into k classes of sizes $n_1, n_2, \ldots, n_k$ respectively

				mean	Total
x_{11}	x_{12}	. . .	x_{1n_1}	$\bar{x}_{1.}$	T_1.
x_{21}	x_{22}	. . .	x_{2n_2}	$\bar{x}_{2.}$	T_2.
.	.		.	.	.
.	.		.	.	.
.	.		.	.	.
x_{i1}	x_{i2}	. . .	x_{in_i}	$\bar{x}_{i.}$	T_i.
.	.				
.	.				
.	.				
x_{k1}	x_{k2}	. . .	x_{kn_k}	$\bar{x}_{k.}$	T_k.
					G

The total variation on the observation x_{ij} can be split into the following two components

1) The variation between the classes or the variation due to different basis of classification commonly known

2) The variation within the classes inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human endeavour.

The second type of variation is due to chance causes which are beyond the control of human hand.

a/2 Null hypothesis (H_0)

To test the equivality of the population means i.e all mean values are homogenity

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

Alternative hypothesis (H_1)

All mean values are not homogenity (Heterogenity)

$$H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k \neq \mu$$

Mathematical model:

The linear mathematical model is

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

$$= \mu_i + \varepsilon_{ij} + \mu - \mu \qquad \mu_i - \mu = \alpha_i$$

$$= \mu + (\mu_i - \mu) + \varepsilon_{ij}$$

$= \mu + \alpha i + \varepsilon ij$

Where,

$x ij \Rightarrow$ represent i^{th} row and j^{th} column

$\mu \Rightarrow$ general mean effect

$\alpha i \Rightarrow$ effect of i^{th} row

$\varepsilon ij \Rightarrow$ error effect due to chance

Statistical analysis

We consider

$$\sum_{i=1}^{k} \sum_{j=1}^{ni} (x ij - \bar{x}..)^2 = \sum_{i=1}^{k} \sum_{j=1}^{ni} (x ij - \bar{x} i. + \bar{x} i.$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{ni} \Big[(x ij - \bar{x} i.)^2 + (\bar{x} i. - \bar{x}..)^2$$

$$+ 2 (x ij - \bar{x} i.)(\bar{x} i. - \bar{x}..) \Big]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{ni} (x ij - \bar{x} i.)^2 + \sum_{i=1}^{k} \sum_{j=1}^{ni} (\bar{x} i. - \bar{x}$$

$$+ 2 \sum_{i=1}^{k} \sum_{j=1}^{ni} \Big[(x ij - \bar{x} i.)(\bar{x} i. - \bar{x}$$

Since the product term is vanished

$$2 \sum_{i=1}^{k} \sum_{j=1}^{ni} (x ij - \bar{x} i.)(\bar{x} i. - \bar{x}..)$$

The algebraic sum of deviation of the i^{th} row from the mean is zero.

Then we have

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{..}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i.}-\bar{x}_{..})^2$$

where

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{..})^2 \Rightarrow$ total sum of squares.

$=$ TSS or ST^2

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2 \Rightarrow$ error sum of square

$=$ SSE or SE^2

$\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x}_{..})^2$ = column sum of squares (or)

sum of square due to treatment

$=$ SSC (or) SC^2

TSS / Total sum of square = SSC + SSE

$ST^2 = SC^2 + SE^2$

Degrees of freedom:

Degrees of freedom for column $c-1$

Degrees of freedom for error $N-c$.

Degrees of freedom for total $N-1$

Mean sum of square

The sum of squares divided by its degrees of freedom is known as mean sum of squares.

1) Mean sum of squares for column

$$MSC = \frac{SSC}{c-1} = \frac{SC^2}{c-1}$$

2) Mean sum of squares for error

$$MSE = \frac{SSE}{N-c} = \frac{SE^2}{N-c}$$

ANOVA - Table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of square	Variance ratio
Between column	SSC (or) SC^2 $\sum n_i(\bar{x}_i - \bar{x}_{..})^2$	$\gamma = C-1$	MSSc (or) ΔC^2 $= \frac{SC^2}{C-1}$	$F_c = \frac{\Delta C^2}{\Delta E^2}$
Error	SSE (or) SE^2 $\sum(x_{ij} - \bar{x}_i)^2$	$N-C=?$	MSSE (or) ΔE^2 $= \frac{SE^2}{N-C}$	
Total	TSS (or) ST^2 $\sum\sum(x_{ij} - \bar{x}_{..})^2$	$N-1=?$		

If an observed value of $F >$ the tabulated value of F_α at specified level of significance that is 5% or 1%. the Ho is rejected.

Problem

1) Random samples of same size deluxe room air conditioners from three manufactures showed the following life expectation (month)

Manufacturer I 34 28 32 29 31

Manufacturer II 23 34 37 32

Manufacturer III 30 39 35 37 32 31

use the ANOVA at 5% & 10% and indicate whether the average life expectations of 3 different brands of room air conditioner differ significantly

Calculation:

Null hypothesis (H0):

The average life expectation of three brands of room air conditione are same

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

Alternative hypothesis (H1):

The average life expectation of three brands of room air conditioner are not same.

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$

Level of significance

$\alpha = 5\%$.

$\alpha = 0.05$

Test statistic:

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
34	23	30	1156	529	900
28	34	39	784	1156	1521
32	37	35	1024	1369	1225
29	32	37	841	1024	1369
31		32	961		1024
		31			961
$\Sigma x_1 =$ 154	$\Sigma x_2 =$ 126	$\Sigma x_3 =$ 204	$\Sigma x_1^2 =$ 4766	$\Sigma x_2^2 =$ 4078	$\Sigma x_3^2 =$ 7000

N = No. of observation

GT = Total No. of observation

(G.T)

Grand total

correction factor:

$$C.F = \frac{[G.T]^2}{N}$$

$$= \frac{[418]^2}{15}$$

$$= \frac{234256}{15}$$

$$\boxed{C.F = 15617}$$

Total sum of square:

$$TSS = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - C.F$$

$$= 4766 + 4078 + 7000 - 15617$$

$$= 15844 - 15617$$

$$TSS = 227$$

column sum of square:

$$SSC = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_2} + \frac{(\Sigma X_3)^2}{N_3} - C.F$$

$$= \frac{(154)^2}{5} + \frac{(126)^2}{4} + \frac{(204)^2}{6} - 15617$$

$$= \frac{23716}{5} + \frac{15876}{4} + \frac{41616}{6} - 15617$$

$$= 4743.2 + 3969 + 6936 - 15617$$

$$= 15648.2 - 15617$$

$$= 31.2$$

Error sum of square

$$SSE = SST - SSC$$

$$= 227 - 31.2$$

$$SSE = 195.8$$

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of square	F-test Variance ratio
Between column	$SSC = 31.2$	$C-1 = 3-1 = 2$	$MSS_C = \frac{SSC}{C-1} = \frac{31.2}{2} = 15.6$	$F_C = \frac{MSS_E}{MSS_C} = \frac{16.316}{15.6} = 1.04$
Error	$SSE = 195.8$	$N-C = 15-3 = 12$	$MSS_E = \frac{SSE}{N-C} = \frac{195.8}{12} = 16.316$	
Total	$TSS = 227$	$N-1 = 15-1 = 14$		

Degrees of freedom

$$\gamma = (N-C, C, 1)$$

$$= (15-3, 3-1)$$

$$= (12, 2)$$

Table value:

At 5% level of significance for (12,2) degrees of freedom the table value is $F_\alpha = 19.41$

conclusion :

calculated value < table value

$1.0415 < 19.41$

$|F| < F_\alpha$

Ho is accepted

Two-way classification

Let us consider the case when there are two factors which may affect the variate values x_{ij}

Example:

The yield of milk may be affected by differences in treatments that is rations as well as the differences in variety.

Let us now suppose that N cows are divided into h different groups, each group containing k cows.

Let us consider the effect of k treatments on the yield of milk.

						Mean
	x_{11}	x_{12}	...	x_{1j}	...	x_{1h}	$\bar{x}_{1.}$
	x_{21}	x_{22}	...	x_{2j}	...	x_{2h}	$\bar{x}_{2.}$
	.	.		.		.	.
	.	.		.		.	.
	.	.		.		.	.
	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ih}	$\bar{x}_{i.}$
	.	.		.		.	
	.	.		.		.	
	.	.		.		.	
	x_{k1}	x_{k2}	...	x_{kj}	...	x_{kh}	$\bar{x}_{k.}$
Mean	$\bar{x}_{.1}$	$\bar{x}_{.2}$	...	$\bar{x}_{.j}$	...	$\bar{x}_{.h}$	$\bar{x}_{..}$
Total	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.h}$	

Null hypothesis (Ho)

The treatment as well as varieties are homogeneous

$$Ho(t): \mu_{1.} = \mu_{2.} = \ldots = \mu_{k.} = \mu$$

$$Ho(v) = \mu_{.1} = \mu_{.2} = \ldots = \mu_{.h} = \mu$$

these equivalent

$$Ht: \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$$

$$Hv: \beta_1 = \beta_2 = \ldots = \beta_h = 0$$

Mathematical model:

Let us suppose that x_{ij} ($i=1$, $j = 1, 2, \ldots h$) are independent the mathematical model

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

x_{ij} ⇒ Effect of i^{th} row and j^{th} column

μ ⇒ General mean effect

α_i ⇒ effect due to i^{th} row

β_j ⇒ effect due to j^{th} column

ε_{ij} ⇒ Error effect due to chance

Statistical Analysis:

We can consider

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij}-\bar{x}..)^2 = \Sigma\Sigma\left[\underline{x_{ij}} - \underline{\bar{x}_{i.}} + \bar{x}_{i.} - \underline{\bar{x}_{.j}} + \bar{x}_{.j} - \bar{x}.. + \underline{\bar{x}..} - \bar{x}..\right]$$

$$= \Sigma\Sigma\left[(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}..) + (\bar{x}_{i.} - \bar{x}..) + (\bar{x}_{.j} - \bar{x}..)\right]^2$$

$$= \Sigma\Sigma(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}..)^2 + \Sigma\Sigma(\bar{x}_{i.} - \bar{x}..)^2 + \Sigma\Sigma(\bar{x}_{.j} - \bar{x}..)^2$$

$$+ 2\Sigma\Sigma(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}..)(\bar{x}_{i.} - \bar{x}..)$$

$$+ 2\Sigma\Sigma(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}..)(\bar{x}_{.j} - \bar{x}..)$$

$$+ 2\Sigma\Sigma(\bar{x}_{i.} - \bar{x}..)(\bar{x}_{.j} - \bar{x}..)$$

Since algebaic sum of deviation set of observation about their mean is zero that is all the product term is vanished. So, we have

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij}-\bar{x}_{..})^2 = \sum\sum(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}$$

$$+\bar{x}_{..})^2+h\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x}_{..}$$

$$+K\sum_{j=1}^{h}(\bar{x}_{.j}-\bar{x}_{..})^2$$

$$TSS = SSR + SSC + SSE$$

$$ST^2 = SR^2 + SC^2 + SE2$$

Where,

* $\sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij}-\bar{x}_{..})^2$ is the total sum of squares
= TSS or ST^2

* $h\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x}_{..})^2$ is row sum of square
= SSR or SR^2

* $K\sum_{j=1}^{h}(\bar{x}_{.j}-\bar{x}_{..})^2$ is column sum of sq
= SSC or SC^2

* $\sum_{i=1}^{k}\sum_{j=1}^{h}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$ is error sum of square
= SSE or SE^2

Degrees of freedom

For row $\gamma = r - 1$

For column $\gamma = c - 1$

For error $\gamma = (r-1)(c-1)$

For total $\gamma = N - 1$ (or) $rc - 1$

Mean sum of square

The sum of square divided by its degrees of freedom is known as mean sum of square.

* Mean sum of square for row

$$SR^2 = \frac{SSR}{r-1} \text{ (or) } \frac{SR^2}{r-1}$$

* Mean sum of square for column

$$SC^2 = \frac{SSC}{c-1} \text{ (or) } \frac{SC^2}{c-1}$$

* Mean sum of square for error

$$SE^2 = \frac{SSE}{(r-1)(c-1)} \text{ (or) } \frac{SE^2}{(r-1)(c-1)}$$

Test statistic

For row

$$FR = \frac{SR^2}{SE^2}$$

For column

$$FC = \frac{SC^2}{SE^2}$$

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	Variance ratio
Between rows	$SSR = h\Sigma(\bar{x}_i - \bar{x}_{..})^2$	$(r-1)$	$SR^2 = \frac{SR^2}{r-1}$	$F_R = \frac{MSSR}{MSSE}$
Between columns	$SSC = K\Sigma(\bar{x}_{.j} - \bar{x}_{..})^2$	$(C-1)$	$SC^2 = \frac{SC^2}{C-1}$	$F_C = \frac{MSSC}{MSSE}$
Error	$SSE = \Sigma\Sigma(x_{ij} - \bar{x}_{i.} + \bar{x}_{.j} - \bar{x}_{..})^2$	$(r-1)(C-1)$	$SE^2 = \frac{SE^2}{(r-1)(C-1)}$	
Total	$TSS = \sum_{i=1}^{K}\sum_{j=1}^{h}(x_{ij} - \bar{x}_{..})^2$	$N-1$		

If an observed value of F > the tabulated value of F_α at specified level of significance that is 5% or 1% the Ho is rejected.

Multiple Range Test [MRT]

1) Student - Newman Keull's test:

The Newman Keull's method is a step-wise multiple comparision. This procedure used to indentify sample means that are significantly different from each other. The procedure for Newman Keull's test are

i) The employes step-wise approach when comparing sample means.

ii) Prior to any mean comparision, all sample means are rank ordered in ascending or decending ordered, thereby producing an ordered range (P) of sample means.

iii) A comparision is made between the largest and smallest sample means within the largest range.

iv) Assuming that the largest range is H means (P=H). A significant difference between the largest and smallest mean as revealed by the "Newman Keuls" method could result in reflection of the null hypothesis for that specific range of means.

v) The next largest comparision of 2 sample means would be made within a smaller range 3 means (P=3).

vi) continue this process until a final comparision is made.

vii) If there is no significant difference between the 2 sample means, all the null hypothesis within that range would be retained no further comparision within smaller ranges are necessary.

18/2

Duncan's Multiple Range Test [DMRT]

Duncan's Multiple Range test is a post hoc test to measure specific differences between pairs of means.

This test is commonly used in agronomy and other agricultural research. The result of the test is a set of subsets of means, where in each subset means have been found not to be significantly different from one other.

This test is summarized the way in finding several significant differences with increasing values (decending order) which depending on the extend of the distance between the treatment means after arranged.

Treatment	Main yield (kg)	Rank
T_2	2678	1
T_3	2552	2
T_4	2128	3
T_1	2127	4
T_5	1496	5
T_6	1681	6
T_7	1316	7

$$SE = \sqrt{\frac{2S^2}{r}}$$

$$= 217.68$$

(t-1) value of the shortest significant range are computed

$$R_p = \frac{(r_p)(SE)}{\sqrt{2}} \quad \text{for } p = 2, 3, \ldots$$

from the largest mean subtract R_p for largest p, then declare as significantly different from the largest mean.

For the remaining treatment whose values are larger, then the difference R_p. Continue this process till all the treatment of T.

Present the results by using this $(T_1, T_2, \ldots T_n)$ to indicate which treatment pairs are significantly different from each other.

Tukey's Range Test [TRT]

Tukey's honest significance difference test can be used on raw data to find means that are significantly different from each other.

This test compares the means of every treatment to the means of every other treatment.

This test is based on a formula very similar to that of the t-test. Tukey's test is essentially a t-test except that is corrects for experiment wise error rate.

$$q_s = \frac{Y_A - Y_B}{SE}$$

where,

$Y_A \rightarrow$ Is a larger of the two means being compared.

$Y_B \rightarrow$ Is the smaller of the two means being compared.

$SE \rightarrow$ Standard error.

This q_s value can be compared to a q value from the range distribution.

If the q_s value is larger than the q critical value, the two means are said to be significantly different.

$$q = \frac{(\bar{Y}_{max} - \bar{Y}_{min})}{\frac{s}{\sqrt{\frac{2}{n}}}}$$