Unit - I

Analysis of Variance :- The analysis of variance is a powerful statistical tool for tests of significance. The test of significance based on t-distⁿ is an adequate procedure only for testing the significance of the difference betⁿ two sample means. In a situation when we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same popⁿ. i.e. they have the same mean.

eg :- Five fertilizers are applied to four plots each of wheat and yields of wheat on each of the plot is given. Now we are interested in finding out whether the effect of these fertilizers on the yields is significantly different or in other words, whether the samples have come from the same normal popⁿ.

The answer to this problem is provided by the technique of analysis of variance. Thus basic purpose of the analysis of variance is to test the homogeneity of several means.

The term "Analysis of Variance" was introduced by prof. R. A. Fisher in 1930's to deal with problem in the analysis of agronomical data.

Cochran Thm :- used to justify results relating to the prob. dist^{n} of statistics that are used in ANOVA.

Page
Date

Variation is inherent in nature. The total variation in any set of numerical data is due to a no. of causes which may be classified as
i) Assignable causes, and
ii) Chance causes.
The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced seperately

Definition :-

According to Prof. R.A. Fisher, Analysis of Variance (ANOVA) is the "seperation of variance ascribable to one group of causes from the variance ascribable to other group.

to credit or assign it to a cause

The ANOVA consists in the estimation of the amount of variation due to each of the indept. factors (causes) seperately and then comparing this estimates due to assignable factors (causes) with the estimates due to chance factor (causes)

Assumption :- For the validity of the F-test in ANOVA, the following assumption are made :-
i) The obser^{n} are indept.
ii) Parent pop^{n} from which observation are taken as is normal and
iii) Various treatment and environmental effects are additive in nature.

Cochran's Theorem :- Let $x_1, x_2, \dots x_n$ denote a r.s. from normal pop^{n} $N(0, \sigma^2)$. Let the sum of the squares of these values are

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

Where Q_j is a quadratic form of $x_1, x_2, \ldots, x_n$ with rank (a_j) r_j; $j = 1, 2, \ldots, k$; then the r.v.s $Q_1, Q_2, \ldots, Q_k$ are mutually independent and Q_j/σ^2 is a χ^2-variate with r_j degrees of freedom (d.f.) where $\sum_{j=1}^{k} r_j = n$.

ANOVA for 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. $(N = \sum_{i=1}^{k} n_i)$ as exhibited below

				Means	Total
x_{11}	x_{12}	$\ldots$	x_{1n_1}	$\bar{x}_{1.}$	$T_{1.}$
x_{21}	x_{22}	$\ldots$	x_{2n_2}	$\bar{x}_{2.}$	$T_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
x_{k1}	x_{k2}	$\ldots$	x_{kn_k}	$\bar{x}_{k.}$	$T_{k.}$
					G

The total variation in the observations x_{ij} can be split into the following two components:

i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

ii) The variations within the classes i.e. the inherent variation of the r.v. 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 and the second type of variation is due to chance causes which are beyond the control of human hand.

The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the due to chance causes i.e. variability within the seperate classes.

Eg: In particular let us consider the effect of k different rations on the yield of milk of N cows (of same breed & stock) These N cows are divided into k classes of sizes $n_1, n_2, \ldots, n_k$ respectively.

$$N = \sum_{i=1}^{k} n_i$$

Here the sources of variation are
(i) Effect of the rations (treatment) t_i $i=1,2,\ldots k$
(ii) Error ε produced by numerous causes. They are not detected and identified & produce a variation of random nature.

Mathematical Model :- Let the linear model will be

$$\begin{aligned} x_{ij} &= u_i + \varepsilon_{ij} \\ &= u + (u_i - u) + \varepsilon_{ij} \\ &= u + \alpha_i + \varepsilon_{ij} \end{aligned} \quad \left.\begin{array}{l} \text{where } (i=1,2,\ldots,k \\ j=1,2,\ldots,n_i) \end{array}\right.$$

— (1)

(i) x_{ij} is the yield from the j^{th} cow, $(j=1,2,\ldots,n_i)$ fed on the i^{th} rations $(i=1,2,\ldots k)$

(ii) u is the general mean effect given by

$$\mu = \sum_{i=1}^{k} (n_i \mu_i)/N \qquad -(2)$$

where μ_i is the fixed effect due to the i^{th} ration, i.e. if there were no treatment differences and no chance causes then the yield of each cow will be μ.

(iii) α_i is the effect of the i^{th} ration (treatment) i.e. given by

$$\alpha_i = \mu - \mu_i \cdot (i = 1, 2, \ldots k) \qquad -(3)$$

i.e. the i^{th} treatment (rations) increase or decrease the yield by an amount α_i. On using (2), we get

$$\sum_{i=1}^{k} n_i \alpha_i = \sum_{i} n_i (\mu_i - \mu) = \sum n_i \mu_i - \mu \sum n_i$$

$$= N \cdot \mu - \mu \cdot N = 0$$

(iv) ε_{ij} is the error effect due to chance.

Assumption in the model

(i) All the obser^{n}: x_{ij} are indept.

(ii) Different effects are additive in nature

(iii) ε_{ij} are i.i.d. $N(0, \sigma_e^2)$

under the (iii) assumption, the model (1) becomes

$$E(x_{ij}) = \mu_i = \mu + \alpha_i \quad ; \quad \begin{matrix} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n_i \end{matrix}$$

Null hypothesis of ~~One~~ for ANOVA for One-way classification

Here we want to test the equality of pop^{n}. means i.e. the homogeneity of different rations. Hence null hypothesis is given by

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

which from (3) reduces to

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Statistical Analysis of the Model :— Let us write

$\bar{x}_{i.}$ = mean of the i^{th} class.

$$= \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i} \quad ; \quad (i = 1, 2, \cdots, k)$$

$\bar{x}_{..}$ = overall mean $= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$

$$= \frac{1}{N} \sum_{i=1}^{k} n_i \bar{x}_{i.}$$

The parameters μ and α_i in (1) are estimated by the principle of least squares on minimising the error (residual) sum of squares given by

$$E = \sum_i \sum_j \varepsilon_{ij}^2 = \sum_i \sum_j (x_{ij} - \mu - \alpha_i)^2 \quad \text{—(*)}$$

The normal eq^{n} for estimating μ and α_i are

$$\frac{\partial E}{\partial \mu} = -2 \sum_i \sum_j (x_{ij} - \mu - \alpha_i) = 0 \quad \text{—(**)}$$

and $$\frac{\partial E}{\partial \alpha_i} = -2 \sum_{j=1}^{n_i} (x_{ij} - \mu - \alpha_i) = 0 \quad \text{—(***)}$$

from (*), we get

$$\sum_i \sum_j x_{ij} - N\mu - \sum n_i \alpha_i = 0$$

$$\hat{\mu} = \frac{1}{N} \sum\sum x_{ij} = \bar{x}_{..} \qquad [\because \sum n_i \alpha_i = 0]$$

& from (**), we get

$$\sum_j x_{ij} - n_i \hat{\mu} - n_i \hat{\alpha} = 0$$

$$\Rightarrow \hat{\alpha} = \frac{1}{n_i} \sum x_{ij} - \hat{\mu} = \bar{x}_{i.} - \hat{\mu}$$

$$\hat{\alpha} = \bar{x}_{i.} - \bar{x}_{..}$$

Hence substituting the value of μ & α in (1) the model becomes

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

$$=$$

We introduce the error component ε_{ij} so that both the sides are equal. This is the deviation within the class which is due to Randomisation. Transposing $\bar{x}_{..}$ to the left squaring both sides and summing over i & j we get

$$\sum_i \sum_j [x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}_{..}]^2$$

$$\sum_i^K \sum_j^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_i^K \sum_j^{n_i} \left[(\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.}) \right]^2$$

$$= \sum_i^K \sum_j^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_i^K \sum_j^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

$$+ 2 \sum_i^K (\bar{x}_{i.} - \bar{x}_{..}) \sum_j^{n_i} (x_{ij} - \bar{x}_{..})$$

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

$$+ 2 \sum_i \sum_j \left[(\bar{x}_{i.} - \bar{x}_{..}) \sum_j (x_{ij} - \bar{x}_{i.}) \right]$$

But $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.}) = 0$ [algebraic sum of the deviations of the observations from their mean is zero]

$$\therefore \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \quad \text{---(1)}$$

$S_T^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{..})$ is known as total sum of squares (T.S.S.)

$S_E^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2$ is called within sum of squares or error sum of squares (S.S.E.) and

$S_t^2 = \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2$ is called S.S. due to treatments (S.S.T.)

Then T.S.S. = S.S.E. + S.S.T.

Degrees of freedom for various S.S. :-

S_T^2, the total S.S. will carry $(N-1)$ d.f. (one d.f. lost because of the linear constraint)

$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..}) = 0$

similarly the treatment sum of squares S_t^2 will have $(k-1)$ d.f.

& S_E^2 the error S.S. will have $(N-k)$ d.f.

Hence we see that the

$$N-1 = (N-k) + (k-1)$$

Mean Sum of Squares (M.S.S.)

Defn The sum of squares divided by its degrees of freedom gives the corresponding variance or the mean sum of squares (M.S.S.). Thus

$\frac{S_t^2}{(k-1)} = \frac{S.S.T.}{(k-1)} = S_t^2$ is M.S.S. due to treatment

& $\frac{S_E^2}{N-k} = \frac{S.S.E.}{(N-k)} = S_E^2$ is M.S.S. due to error.

Hence the test statistic for H_0 provided by the variance ratio

$$F = \frac{S_t^2}{S_E^2} \quad \text{(XXX)}$$

If H_0 is true the F- should take the value 1, otherwise it should be greater than unity. In order to find out if an observed value of F is significantly greater than unity. we have

to obtain the sampling dist^n of the F-statistic defined as

$$F = S_t^2 / S_E^2$$

Under H_0 by Cochran's thm., $\frac{S_t^2}{\sigma_e^2}$ & $\frac{S_E^2}{\sigma_e^2}$ are independently distributed as χ^2 variates with $(k-1)$ and $(N-k)$ d.f. respectively.

Hence the statistic

$$F = \left[\frac{S_t^2}{\sigma_e^2}\frac{1}{k-1}\right] \div \left[\frac{S_E^2}{\sigma_e^2}\cdot\frac{1}{(N-k)}\right] = \frac{S_t^2}{S_E^2}$$

follows Snedecor's F (central) dist^n with $[(k-1), (N-k)]$ d.f.

Thus if an observed value of F obtained from (xxx) is greater than tabulated value of F for $(k-1, N-k)$ d.f. at specified level of significance (usually 5% or 1%) then H_0 is refuted at that level otherwise H_0 may be retained.

The above statistical analysis is very elegantly presented in the following table

ANOVA Table for ONE-WAY CLASSIFIED DATA

Sources of Variation	Sum of Squares	d.f.	Mean Sum of Squares	Variance Ratio
Treatment (rations)	S_t^2	$k-1$	$s_t^2 = \frac{S_t^2}{k-1}$	$\frac{s_t^2}{s_E^2} = F_{k-1, N-k}$
Error	S_E^2	$N-k$	$s_E^2 = \frac{S_E^2}{N-k}$	
Total	S_T^2	$N-1$		