ANALYSIS OF VARIANCE (ANOVA)

INTRODUCTION

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between 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 population I.e., the means are equal.

For example, three types of fertilisers are applied to five plots each and their yields on each of the plot is given as follows

plots	Yield of wheat in tons		
	fertiliser A	Fertiliser B	fertiliser C
1	20	18	25
2	21	20	25
3	23	17	25
4	16	15	25
5	20	25	25
Mean	100/5=20	95/5=19	125/5=25

We have to study if the effect of these fertilisers on the yield is significantly different, or in other words the samples are from the same population. The answer to this is provided by the technique of analysis of variance.

The basic purpose of analysis of variance is to test the homogeneity of several means.

The term ‘Analysis of variance’ was introduced by Prof. R.A. Fisher in 1920’s.

Variation is inherent in nature,

The total variation in any set of numerical data is due to a number of causes which may be calculated as i) assignable causes and ii) chance causes

The variation due to assignable causes can be detected and measured whereas the chance causes is beyond the control of human and cannot be traced.

Examples of assignable causes of variation

Inappropriate procedures, substandard raw materials, measurement errors, temperature etc.,

DEFINITION:

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

The ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to assignable factors(causes), with the estimate due to chance factor (causes). The later being known as experimental error.

ASSUMPTIONS FOR ANOVA TEST

ANOVA test is based on the test statistics F (variance Ratio)

For the validity of the F-test in ANOVA, the following assumptions are made:

i) The observations are independent,
ii) Parent population from which observations are taken is normal, and
iii) Various treatment and environmental effects are additive in nature.

IMPORTANCE:

ANOVA technique enables us to compare several population means simultaneously and thus results in lot of savings in time and money

The origin of ANOVA technique lies in agricultural experiments but it finds its applications in almost all types of design of experiments in various diverse fields such as in industry, education, psychology, business etc.,

The ANOVA technique is not designed to test the equality of several population variances. It's objective is to test the equality of several population means or the homogeneity of several independent sample means.

In addition to testing homogeneity of several sample means, the ANOVA technique is now frequently applied in testing the linearity of the fitted regression line or the significance of the correlation ratio.

MATHEMATICAL MODEL

FIXED EFFECT MODEL AND RANDOM EFFECT MODELS

A fixed effects model is a statistical model in which the model parameters are fixed or non-random quantities.

In random effects model in which all or some of the model parameters are random variables.

Fixed effects model:

Suppose the k-levels of the factor (treatments) under consideration are the only levels of interest and all these are included in the experiment by the investigator or out of a large number of classes, the k classes (treatments) in the model have been specifically chosen by the experimenter. In such a case α_i's the effect of the i^{th} treatment [$\alpha_i = (\mu_i - \mu)$] are fixed constants (unknown) and the model is a fixed effect model.

In the fixed effect model, the conclusions about the test of hypothesis regarding the parameters α_i's will apply only to k-treatments (factor levels) considered in the experiment.

These conclusions cannot be extended to other remaining treatments (factors) which are not considered in the experiment.

Random effects model:

Suppose we have a large number of classes (treatments) and we want to test, through an experiment if all these class effects are equal or not. Due to consideration of time, money or administrative convenience it may not be possible to include all the factor levels in the experiment. In such a situation, we take only a random sample of factor levels in the experiment and after studying and analysing the sample data, we draw conclusions which would be valid for all the factor levels whether included in the experiment or not. In such a situation the parameters α_i's in the model will not be fixed constants but will be random samples and the model is random effect model.

In the random effect model if the null hypothesis of the homogeneity of class (treatment) effects is rejected, then to test to test the difference between two class(treatments) effects we cannot apply the t-test because all treatments are not included in the experiment.

5·2. ONE-WAY CLASSIFICATION

Let us suppose that N observations y_{ij}, $(i = 1, 2, ..., k;\ j = 1, 2, ..., n_i)$ of a random variable Y are grouped, on some basis, into k classes of sizes $n_1, n_2, ..., n_k$ respectively, $\left(N = \sum_{i=1}^{k} n_i\right)$ as exhibited in *Table 5·1*.

TABLE 5·1 : ONE-WAY CLASSIFIED DATA

Class	*Sample Observations*				*Total*	*Mean*
1	y_{11}	y_{12}	...	y_{1n_1}	$T_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	...	y_{2n_2}	$T_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	...	y_{in_i}	$T_{i\cdot}$	$\bar{y}_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	...	y_{kn_k}	$T_{k\cdot}$	$\bar{y}_{k\cdot}$

The total variation in the observation y_{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 variation *within the classes, i.e.*, the 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 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 inherent variability within the separate classes.

In particular, let us consider the effect of k different rations on the yield in milk of N cows (of the same breed and stock) divided into k classes of sizes $n_1, n_2, ..., n_k$ respectively, $N = \sum_{i=1}^{k} n_i$. Here the sources of variation are :

(i) Effect of the ration (treatment) : t_i ; $i = 1, 2, \ldots, k$.

(ii) Error (ε) produced by numerous causes of such magnitude that they are not detected and identified with the knowledge that we have and they together produce a variation of random nature obeying Gaussian (Normal) law of errors.

Mathematical Model. In this case the *linear* mathematical model will be :

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij} \;;\; \text{where } (i = 1, 2, \ldots k \;;\; j = 1, 2, \ldots, n_i) \qquad \ldots (5\cdot1) \end{aligned}$$

(i) y_{ij} is the yield from the jth *cow*, ($j = 1, 2, \ldots, n_i$) fed on the ith ration ($i = 1, 2, \ldots, k$). ... (5·2)

(ii) μ is the general mean effect given by :

$$\mu = \sum_{i=1}^{k} n_i \mu_i / N \qquad \ldots (5\cdot2a)$$

where μ_i is the fixed effect due to the ith ration, *i.e.*, if there were no treatment differences and no chance causes then the yield of each cow will by μ,

(iii) α_i is the effect of the ith ration given by : $\quad \alpha_i = \mu_i - \mu$, ($i = 1, 2, \ldots, k$) ... (5·2b)

i.e., the ith ration increases (or decreases) the yield by an amount α_i. On using (5·2a) and (5·2b), we get

$$\sum_{i=1}^{k} n_i \alpha_i = \sum_i n_i (\mu_i - \mu) = \sum_i n_i \mu_i - \mu \sum n_i = N.\,\mu - \mu.\,N = 0 \qquad \ldots (5\cdot2c)$$

(iv) ε_{ij} is the error effect due to chance. ... (5·2d)

ANOVA FOR FIXED EFFECT MODEL

The fixed effect or parametric model used is

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

$$= \mu + \alpha_i + \varepsilon_{ij} \quad (i=1,2\ldots k;\ j=1,2\ldots n_i) \text{ where } \alpha_i = \mu_i - \mu$$

where y_{ij} is the yield from the ith row and jth column

μ is the general mean effect given by $\mu = \sum_{i=1}^{k} n i \mu i / N$

μ_i is the fixed effect due to ith treatment

α_i is the effect of the ith treatment given by

$$\alpha_i = \mu_i - \mu$$

ε_{ij} is the error effect due to chance

ASSUMPTIONS IN THE MODEL

i) all the observations (y_{ij}'s) are independent and $y_{ij} \sim N(\mu_i, \sigma_e^2)$
ii) different effects are additive in nature

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

under the third assumption, the model becomes

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

STATISITICAL ANALYSIS OF THE MODEL

Null hypothesis: we have to test the equality of the population means.
Hence the null hypothesis is given by

$$H_0 = \mu_1 = \mu_2 = \ldots = \mu_k = \mu$$

which reduces to $H_0 = \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ since $\alpha_i = \mu_i - \mu$

Alternate hypothesis:

H_1: atleast two of the means $\mu_1, \mu_2, \ldots \mu_k$ are different.

let us write

(ii) ALTERNATIVE HYPOTHESIS:

$$H_1 = \mu_1 \neq \mu_2 \neq \cdots\cdots \neq \mu_K$$

we know that,

$\overline{y_{i.}}$ = mean of the i^{th} class

$$= \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}, \quad i = 1, 2, 3 \cdots, k.$$

$$\overline{y}_{..} = \text{overall mean} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij})$$

$$= \frac{1}{N} \sum_{i=1}^{k} n_i \overline{y_{i.}}$$

(iii) Least square Estimates of parameters.

The parameters μ & α_i ~~given & above model are given by above~~ of the model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ are estimated by the principle of least squares, minimising the error sum of squares.

The mathematical model is

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ ——— } (1)$$

$$\Rightarrow \epsilon_{ij} = y_{ij} - \mu - \alpha_i \text{ ——— } (2)$$

squaring & taking $\sum_i \sum_j$ on both sides we get,

$$E = \sum_i \sum_j \epsilon_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2$$

Estimation of μ.

$$E = \sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2$$

$$\frac{\partial E}{\partial \mu} = -2 \sum_i \sum_j (y_{ij} - \mu - \alpha_i) = 0$$

$$= \sum_i \sum_j y_{ij} - \sum_i \sum_j \mu - \sum_i \sum_j \alpha_i = 0$$

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

$$\Rightarrow \mu = \frac{\sum_i \sum_j y_{ij}}{N} = \bar{y}_{..}$$

$$\boxed{\because \sum n_i \alpha_i = 0 \text{ from note 1}}$$

$$\hat{\mu} = \bar{y}_{..}$$

Estimate of α_i

$$E = \sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2$$

$$\frac{\partial E}{\partial \alpha_i} = -2 \sum_j (y_{ij} - \mu - \alpha_i) = 0$$

$$= \sum_j y_{ij} - \sum_j \mu - \sum_j \alpha_i = 0$$

$$= \sum_j y_{ij} - n_i \hat{\mu} - n_i \alpha_i = 0$$

$$\Rightarrow n_i \alpha_i = \sum_j y_{ij} - n_i \hat{\mu}$$

$$\alpha_i = \frac{1}{n_i} \sum_j y_{ij} - \frac{n_i \hat{\mu}}{n_i}$$

$$= \frac{1}{n_i} \sum_j y_{ij} - \hat{\mu}$$

$$= \bar{y}_i - \bar{y}_{..}$$

hence the model becomes.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_i)$$

TOTAL SUM OF SQUARE:

$$T.S.S = S.S.E + S.S.T$$

$$T.S.S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

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

$$\doteq \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum_i n_i (\bar{y}_i - \bar{y}_{..})^2 +$$

$$2\left[\sum_i \{(\bar{y}_{i.} - \bar{y}_{..}) \sum (y_{ij} - \bar{y}_{i.})\}\right]$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\left[\because \sum_{j=1}^{n_i} (y_{ij} - y_{i.}) = 0\right] \quad (3)$$

From (3), we get.

$$S_E^2 = S.S.E = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$S_t^2 = S.S.T = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2.$$

DEGREES OF FREEDOM:

(i) The degrees of freedom for T.S.S (S_T^2) is $N-1$.

(ii) The degrees of freedom for treatment S.S (S_t^2) is $k-1$

(iii) The df of error sum of square (S_E^2) is $N-k$.

MEAN SUM OF SQUARES (M.S.S)

(i) The M.S.S due to treatments

$$= \frac{S_t^2}{k-1} = \phi_t^2$$

(ii) M.S.S due to Error $= \dfrac{S_E^2}{N-k} = \phi_E^2$

ANOVA TABLE:

The ANOVA table for one way classified data.

Sources of variation	sum of squares	d.f	mean sum of squares	variance ratio
Treatment	S_t^2	$k-1$	$\phi_t^2 = \dfrac{S_t^2}{k-1}$	$F = \dfrac{\phi_t^2}{\phi_E^2} \sim F_{k-1, N-k}$
Error	S_E^2	$N-k$	$\phi_E^2 = \dfrac{S_E^2}{N-k}$	
Total	S_T^2	$N-1$		

Conclusion:

If $F = \dfrac{\phi_t^2}{\phi_E^2}$ is less than the table value of $F_{k-1, N-k}$, we accept H_0 otherwise reject H_0.

VARIANCE OF THE ESTIMATES:

The linear model for one way classification is given by

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} ; \quad i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, n_i$$

The least square estimates of μ & α_i are given by

$$\hat{\mu} = \bar{y}_{..} \text{ \& } \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \quad \text{—— (1)}$$

Then,

$$\text{Var}(\hat{\mu}) = E[\hat{\mu} - E(\hat{\mu})]^2$$

$$= E[\bar{y}_{..} - E(\bar{y}_{..})]^2 \quad \text{from (1)}$$

$$= E(\bar{y}_{..} - \mu)^2$$

$$= E(\bar{\epsilon}_{..}^2) = \text{var}(\bar{\epsilon}_{..}) = \frac{\sigma_e^2}{N}$$

But $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$

we know that $\bar{y}_i = \mu + \alpha_i + \bar{\epsilon}_i$

and $\bar{y}_{..} = \mu + \bar{\epsilon}_{..}$

Then we have

$$\hat{\alpha}_i = \mu + \alpha_i + \bar{\epsilon}_{i\cdot} - (\mu + \bar{\epsilon}_{\cdot\cdot})$$

$$= \alpha_i + \bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}$$

$$\therefore E(\hat{\alpha}_i) = \alpha_i \quad \because \epsilon_{ij} \sim N(0, \sigma_e^2)$$

$$\therefore \hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}$$

$$V(\hat{\alpha}_i) = E[\hat{\alpha}_i - E(\hat{\alpha}_i)]^2$$

$$= E[\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2$$

$$= E(\bar{\epsilon}_{i\cdot}^2) + E(\bar{\epsilon}_{\cdot\cdot}^2) - 2E(\bar{\epsilon}_{i\cdot}\,\bar{\epsilon}_{\cdot\cdot}) \quad \text{--- } (2)$$

But

$$E[(\bar{\epsilon}_{i\cdot})(\bar{\epsilon}_{\cdot\cdot})] = E\left[\left(\frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_{ij}\right)\left(\frac{1}{N}\sum_{i}^{k}\sum_{j}^{n_i} \epsilon_{ij}\right)\right]$$

$$= \frac{1}{n_i N} E\left[(\epsilon_{i1} + \epsilon_{i2} + \cdots + \epsilon_{in_i})\left\{\begin{array}{l} \epsilon_{11} + \epsilon_{12} + \cdots + \epsilon_{1n_i} \\ \vdots \\ + \epsilon_{i1} + \epsilon_{i2} + \cdots + \epsilon_{in_i} \\ \vdots \\ + \epsilon_{k1} + \epsilon_{k2} + \cdots + \epsilon_{kn_i} \end{array}\right.\right.$$

$$= \frac{1}{n_i N}\left[E(\epsilon_{i1} + \epsilon_{i2} + \cdots + \epsilon_{in_i})^2\right]$$

all co-variance terms vanish, since ϵ_{ij}'s are un correlated.

$$\therefore E(\bar{\epsilon}_{i\cdot}\,\bar{\epsilon}_{\cdot\cdot}) = \frac{1}{n_i N}\left[var(\epsilon_{i1}) + var(\epsilon_{i2}) + \cdots + var(\epsilon_{in_i})\right]$$

$$= \frac{1}{\cancel{n_i} N} \cancel{n_i}\, \sigma_e^2$$

Substituting 3 ~~from~~ in 2, we get.

$$\therefore \operatorname{var}(\hat{\alpha}_i) = E(\bar{\epsilon}_{i.}^2) + E(\bar{\epsilon}_{..}^2) - 2E(\bar{\epsilon}_{i.}\,\bar{\epsilon}_{..})$$

$$= \frac{\sigma_e^2}{n_i} + \frac{\sigma_e^2}{N} - 2\frac{\sigma_e^2}{N}$$

$$= \frac{\sigma_e^2}{n_i} - \frac{\sigma_e^2}{N}$$

$$= \sigma_e^2\left(\frac{1}{n_i} - \frac{1}{N}\right)$$

TWO WAY CLASSIFICATION :

Suppose n observations are classified into K categories (or classes), say $A_1, A_2, \ldots\ldots, A_K$ according to some criterion A; and into h categories say $B_1, B_2 \ldots B_h$ according to some criterion B, having kh combinations A_i, B_j; $i = 1, 2 \ldots k$, $j = 1, 2 \ldots h$; often called cells.

This Scheme of classification according to two factors or criteria is called two-way classification and its analysis is called two-way analysis of variance..

The number of observations in each cell may be equal or different, but we shall consider one observation per cell so that

So that $n = hk$, i.e., The total number of cells is $n = hk$

In the two way classification, the values of the response Variables are affected by two factors.

For example, the yield of milk may be affected by differences in treatment, (i.e) rations as well as the differences in variety, (ie) breed and stock of the cows. Let us now suppose that the n cows are divided into n different groups or classes according to their breed and stock, each groups containing k cows and then let us consider the effect of k treatment (ie) rations given at random to cows in each group on the yield of milk.

Let $Y_{ij} = \left[\text{Yield of milk from the cow } j^{th} \text{ breed or stock, fed on the ration } i \right]$

$i = 1, 2, \ldots k; \; j = 1, 2, \ldots h$

Treatments (Rations)	Varieties of Cows $1 \quad 2 \cdots j \cdots h$	Row Totals $(\sum_j y_{ij})$	Row Means $(\sum_j y_{ij})/h$	COLUMN Means $(\sum_i y_{ij})/k$
1	$y_{11} \quad y_{12} \cdots y_{ij} \cdots y_{1h}$	$T_{1.}$	$\bar{y}_{1.}$	$\bar{y}_{.1}$
2	$y_{21} \quad y_{22} \cdots y_{2j} \cdots y_{2h}$	$T_{2.}$	$\bar{y}_{2.}$	$\bar{y}_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i1} \quad y_{i2} \cdots y_{ij} \cdots y_{in}$	$T_{i.}$	$\bar{y}_{i.}$	$\bar{y}_{j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$y_{k1} \quad y_{k2} \cdots y_{kj} \cdots y_{kn}$	$T_{k.}$	$\bar{y}_{k.}$	$\bar{y}_{.n}$
Column Total	$T_{.1} \quad T_{.2} \cdots T_{.j} \cdots T_{.n}$	$G = \sum\sum y_{ij}$		

TWO WAY ANOVA for fixed Effect Model

The mathematical model is given by

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

$$= \mu + (\mu_{i} - \mu) + (\mu_{.j} - \mu) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu) + \varepsilon_{ij}$$

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

Where, $\sum_{i=1}^{k} \alpha_i = 0 = \sum_{j=1}^{h} \beta_j$

and $\sum_{i=1}^{k} \gamma_{ij} = 0 \;\forall j \;;\; \sum_{j=1}^{h} \gamma_{ij} = 0 \;\forall i$

In this case, the interaction effect $\gamma_{ij} = 0$ and the mode reduces to,

H_{11} : Atleast two $\mu_{i.}$'s are different

H_{12} : Atleast two $\mu_{.j}$'s are different

Least square Estimates of parameters,

The linear model is given by,

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

$$\varepsilon_{ij} = y_{ij} - \mu - \alpha_i - \beta_j$$

Why applying the principal of least square we get,

$$E = \sum_i \sum_j \varepsilon_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

The normal equation for estimating

$$\frac{dE}{d\mu} = 0 \Rightarrow -2 \sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)$$

$$\frac{dE}{d\alpha_i} = 0 \Rightarrow -2 \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)$$

$$\frac{dE}{d\beta_j} = 0 \Rightarrow -2 \sum_i (y_{ij} - \mu - \alpha_i - \beta_j)$$

Since, $\sum_i \alpha_i = 0 = \sum_j \beta_j$, we get from the above

equations,

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where, $\mu = \sum_i \sum_j \mu_{ij}$

$$\mu_i = \frac{1}{h} \sum_{j=1}^{h} \mu_{ij}$$

$$\mu_j = \frac{1}{k} \sum_{i=1}^{k} \mu_{ij}$$

$$\alpha_i = \mu_i - \mu \; ; \; \sum_{i=1}^{k} \alpha_i = 0$$

$$\beta_j = \mu_j - \mu \; ; \; \sum_{j=1}^{h} \beta_j = 0$$

Statistical Analysis of a fixed effect model.

Null Hypothesis (H_0):

The treatments and the varieties of homogenius.

H_{01} : The treatment are homogeneous

H_{02} : The variety are homogeneous.

i.e), $H_{01} : \mu_{1.} = \mu_{2.} = \cdots = \mu_{i.} = \mu$

$H_{02} : \mu_{.1} = \mu_{.2} = \cdots = \mu_{.j} = \mu$

Alternative Hypothesis :

H_{1t} : At least two of the $\mu_{i.}$'s are different.

H_{1v} : At least two of the $\mu_{.j}$'s are different

Or their equalent.

$$\hat{\mu} = \frac{1}{nk} \sum_i \sum_j y_{ij} = \bar{y}_{..}$$

$$\hat{\alpha}_i = \frac{1}{n} \sum_j y_{ij} - \hat{\mu}$$

$$= \bar{y}_{i.} - \bar{y}_{..}$$

$$\hat{\beta}_j = \frac{1}{k} \sum_i y_{ij} - \hat{\mu}$$

$$= \bar{y}_{.j} - \bar{y}_{..}$$

Thus the linear model becomes

$$y_{ij} = \bar{y}_{..} + (\bar{y}_i - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

Partitioning of the sum of squares,

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j \left[(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) \right]^2$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 +$$

$$2 \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + 2 \sum_i \sum_j (\bar{y}_{.j} - \bar{y}_{..}) (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + 2 \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})$$

Now,

$$\sum_i\sum_j(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})=\sum_i\left[(\bar{y}_{i.}-\bar{y}_{..})\sum_j(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})\right]$$

$$=\sum_i\left[(\bar{y}_{i.}-\bar{y}_{..})\left\{\sum_j(y_{ij}-\bar{y}_{i.})-\sum_j(\bar{y}_{.j}-\bar{y}_{..})\right\}\right]=0,$$

$$\therefore \sum_i\sum_j(y_{ij}-\bar{y}_{..})^2=h\sum_i(\bar{y}_{i.}-\bar{y}_{..})^2+k\sum_j(\bar{y}_{.j}-\bar{y}_{..})^2+\sum_i\sum_j(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2$$

Or, $$S_T^2=S_t^2+S_v^2+S_E^2$$

Where,

$S_T^2=\sum_i\sum_j(y_{ij}-\bar{y}_{..})^2$ is the total s.s

$S_t^2=h\sum_i(\bar{y}_{i.}-\bar{y}_{..})^2$ is s.s due to treatment

$S_v^2=k\sum_j(y_{.j}-\bar{y}_{..})^2$ is the s.s due to varieties,

and, $S_E^2=\sum_i\sum_j(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2$ is the error or residuals

Degrees of freedom for various s.s

S_T^2 being computed in $N=hk$ quantities $(y_{ij}-\bar{y}_{..})$

Which are subject to one linear constraint $\sum_i \sum_j (y_{ij}-\bar{y}_{..})=0$

Will carry $(N-1)$

$$S_t^2 \longrightarrow (n-1) \longrightarrow \sum_i (\bar{y}_{i.} - \bar{y}_{..}) = 0$$

$$S_v^2 \longrightarrow (k-1) \longrightarrow \sum_j (y_{.j} - \bar{y}_{..}) = 0$$

$$S_E^2 \longrightarrow (N-1)-(k-1)-(h-1) = (h-1)(k-1)$$

Thus the partitioning of d.f is as follows

$$[h(k-1)] = (k-1)+(h-1)+(h-1)(k-1)$$

Which implies that the d.f are additive.

Test statistic

Mean s.s due to

$$\text{treatments} = \frac{S_t^2}{k-1} = s_t^2$$

$$\text{Mean s.s due to varieties,} = \frac{S_v^2}{h-1} = s_v^2$$

$$\text{Error Mean S.S} = \frac{S_E^2}{(h-1)(k-1)} = s_E^2$$

Over j from 1 to h and dividing by h, we get

$$\frac{1}{h}\sum_j y_{ij} = \frac{1}{h}\left[h\mu + h\alpha_i + \sum_j \beta_j + \sum_j \varepsilon_{ij}\right] \Rightarrow$$

$$\bar{y}_{i.} = \mu + \alpha_j + \bar{\varepsilon}_{i.}$$

Over i from 1 to k and dividing by k and using $\sum_i \alpha_i = 0$

We get,

$$\bar{y}_{.j} = \mu + \beta_j + \bar{\varepsilon}_{ij}$$

Over i and j both and dividing by hk and using,

We shall get,

$$\bar{y}_{..} = \mu + \bar{\varepsilon}_{..}$$

ANOVA TABLE FOR TWO-WAY CLASSIFIED

Sources of variation	S.S	d.f	M.S.S	Variance Ratio (F)
Factor A (ROWS)	S.S.A	$(k-1)$	$\frac{S.S.A}{k-1} = M.S.S.A$	$F_A = \frac{M.S.S.A}{M.S.S.E} = \frac{s_t^2}{s_E^2} - F(k-1),(k-1)(h-1)$
Factor B (COLUMNS)	S.S.B	$(h-1)$	$\frac{S.S.B}{h-1} = M.S.S.B$	$F_B = \frac{M.S.S.B}{M.S.S.E} = \frac{s_v^2}{s_E^2} - F(h-1),(k-1)(h-1)$
Error	S.S.E	$(k-1)(h-1)$	$\frac{S.S.E}{(k-1)(h-1)} = M.S.S.E$	
Total	T.S.S	$hk-1$		

Conclusion :

If $F_t \le F_{(k-1),(h-1)(k-1)}$, We accept H_{01}

otherwise,

If $F_v \le F_{(h-1),(h-1)(k-1)}$, We accept H_{02} otherwise reject H_{02}.