ANALYSIS OF VARIANCE

Lecture-2

- ONE-WAY CLASSIFICATION:

Let us suppose that N number of observations x_{ij} (i = 1,2,…,k; j = 1,2,…,n_i) of a random variable X are grouped on some basis, in to k classes of sizes n_1, n_2 ,…,n_k respectively.

(N = $\sum_{i=1}^{k} n_i$) as exhibited in the following layout:

Class	Sample observations	Total	Mean
1	x_{11} x_{12} ……………… x_{1n1}	$T_{1.}$	$\bar{x}_{1.}$
2	x_{21} x_{22} ……………… x_{2n2}	$T_{2.}$	$\bar{x}_{2.}$
…..	……………………………	………	……..
i	x_{i1} x_{i2} ……………… x_{ini}	$T_{i.}$	$\bar{x}_{i.}$
…..	…………………………….	…….	……..
k	x_{k1} x_{k2} ……………… x_{knk}	$T_{k.}$	$\bar{x}_{k.}$
		G= Grand total	= $\bar{\bar{x}}_{..}$ Overall mean

The total variation in the observation 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 variation within the classes, i.e. that the inherent variation of the random variable within the observations of a class.

The first the first type of vision is due to assignable causes which can be detected and controlled by human efforts and the second type of radiation is due to chance causes which are beyond the control of human head

The main objective of analysis of variance technique is to examine if there is significant difference between the glass means in view of the inherent variability within the separate classes .

Remark: **Fixed effect model Vs. Random effect model**

Fixed effect 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, k classes are specifically chosen by the experimenter. In such a situation the model becomes fixed effect model.

Random effect model: Suppose we have a large number of classes under consideration and we want to test through an experiment if all these classifiers are equal or not do you do 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 classes in the experiment. And after studying and analyzing the sample data we want to draw conclusion, which would be valid for all the classes, whether they are included in the experiment or not.

*** Mathematical model for one-way ANOVA:**

If the factor levels under consideration are the only levels of interest, then the fixed effect or parametric linear mathematical model will be:

$$x_{ij} = \mu_i + e_{ij} \qquad (i= 1,2,\ldots k, j= 1,2,\ldots.,n_i)$$

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

$$= \mu + \alpha_i + e_{ij} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where μ = general mean effect

α_i = effect due to ith treatment , $\alpha_i = \mu_i - \mu$

e_{ij} = error effect due to chance

Also, note that,

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

$$= \sum_i n_i\mu_i - N\mu$$

$$= N\mu - N\mu \quad \text{since } \mu = \frac{\sum_{i=1}^{k} n_i\mu_i}{N} = \text{general mean effect}$$

$$= 0$$

* Assumptions:

i) All sample observations add are independent .

ii) Various effects are additive in nature.

iii) $e_{ij} \sim N(0, \sigma_e^2)$

*** Statistical Analysis of the Model**:

* Null Hypothesis:

H_0 : The data is homogeneous or all the treatments are equally effective.

i.e. they have same means.

Symbolically,

$H_0 : \mu_{1.} = \mu_{2.} = \ldots\ldots = \mu_{k.} = \mu$ Or equivalently we can write,

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

and the alternative is at least two treatment means are different.

Let us define mean of ith class and an overall mean as follows:

$$\bar{x}_{i.} = \frac{\sum_j x_{ij}}{n_i} = \text{mean of i}^{\text{th}} \text{ treatment}$$

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

- **Least square estimates of the parameters in the model:**

  The parameters μ and α_i in the model (1) can be estimated by using the principle of least squares on minimizing the error sum of squares.

  i.e. We minimize E = $\sum_i \sum_j (x_{ij} - \mu - \alpha_i)^2$

  The normal equations for estimating μ and α_i are

  $\frac{\partial E}{\partial \mu} = 0$ and $\frac{\partial E}{\partial \alpha_i} = 0$ respectively.

Now, $\frac{\partial E}{\partial \mu} = 0 \Rightarrow -2\sum_i \sum_j (x_{ij} - \mu - \alpha_i) = 0$

$$\Rightarrow \sum_i \sum_j x_{ij} - N\mu - 0 = 0 \text{ since } \sum_i n_i \alpha_i = 0$$

$$\Rightarrow \hat{\mu} = \bar{x}_{..}$$

Similarly,

$$\frac{\partial E}{\partial \alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{n_i} (x_{ij} - \mu - \alpha_i) = 0$$

$$\Rightarrow \sum_j x_{ij} - n_i \mu - n_i \alpha_i = 0$$

$$\Rightarrow \hat{\alpha}_i = \frac{\sum_j x_{ij}}{n_i} - \hat{\mu}$$

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

***Splitting of total S.S. in to various S.S.:**

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

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

Total sum of squares = SSE +SSt

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

Sum of squares due to error = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$

Sum of squares due to treatments = $\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$

*Degrees of freedom:

Total s.s. = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2$, is computed using all N observations which are subjected to one restriction given by $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..}) = 0$. Hence the degrees of freedom for total ss will be N-1.

S.S. t = $\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$, is based on k quantities subjected to one restriction, i.e. $\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..}) = 0$. Hence, the degrees of freedom for treatments s.s. will be k-1.

Finally, S.S.E. = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2$, is based on N quantities which are subjected to k restrictions, $\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.}) = 0 \; ; i = 1, 2, ..., k$. Thus, the degrees of freedom for S.S.E will be N-k.

Hence the degrees of freedom are also additive in nature.

***Mean Sum Of Squares:**

The mean sum of squares is nothing but the variance of the corresponding factor. i.e. it is obtained by dividing the sum of squares by its degrees of freedom. Thus,

Mean sum of squares(MSS) due to treatments = S.S.t / k-1

And mean sum of squares due to error is = S.S.E./ N-k i.e. MSSt $= \dfrac{\sum_{i=1}^{k} n_i (\bar{x}_{i.} - \bar{x}_{..})^2}{k-1}$

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

Hence, the ANOVA table for one-way classification is give as follows:

ANOVA table

Sources of variation	**Degrees of freedom**	**S.S.**	**M.S.S.**	**F-ratio**
Treatments	k-1	$S_t^2 = \sum_{i=1}^{k} n_i (\bar{x}_{i.} - \bar{x}_{..})^2$	$s_t^2 = S_t^2 / k-1$	$F_{cal} = s_t^2 / s_E^2$
Error	N-k	$S_E^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$	$s_E^2 = S_E^2 / N-k$	
Total	N-1	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2$		

Let, $F_\alpha = F_{tab}$ at α% level of significance with (k-1, N-k) d.f.

If $F_{cal} > F_{tab,}$ we reject H_0. Otherwise we accept it.

i.e. We conclude that at least two treatment means are different.

In order to find out which pair of treatments differ significantly, we follow the procedure given below.

***Least significance difference(Critical difference (C.D)):**

If the treatments shoe significant effect then we would be interested to find out which pairs of treatments differ significantly. For this we calculate critical difference or least significant difference. The C.D between any two treatments say, $\bar{x}_{i.}$ and $\bar{x}_{j.}$ at level of significance α is given by

C.D.($\bar{x}_{i.}$ - $\bar{x}_{j.}$) = t (α/2)% for error degrees of freedom * S.E.($\bar{x}_{i.}$ - $\bar{x}_{j.}$)

Where, S.E.($\bar{x}_{i.}$ - $\bar{x}_{j.}$)= $\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)MSSE}$, where, n_i and n_j are the sizes of the corresponding class.

If the difference $\left|\bar{x}_{i.}-\bar{x}_{j.}\right|$ > C.D. we conclude that the corresponding pair of treatments is not homogeneous. i.e. They differ significantly. Otherwise we say that the pair is homogeneous.

Remark:

For practical problems we use the following simplified formula:

1) Total S.S. = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2-\frac{G^2}{N}$, where G = grand total

2) Treatment S.S. = $\sum_{i}\frac{T_i^2}{n_i}-\frac{G^2}{N}$

3) Error S.S. = Total S.S. – Treatment S.S.

***REMARK:**

In a one-way ANOVA, the F statistic tests whether the treatment effects are all equal, i.e. that there are no differences among the means of the k groups.

A significant F value indicates that there are differences in the means, but it does not tell you where those differences are, e.g. group 1’s mean might be different than group 2’s mean but not different from group 3’s mean. To isolate where the differences are, you could do a series of pair wise t-tests. The problem with this is that the significance levels can be misleading. For example, if you have 7 groups, there will be 21 pair wise comparisons of means; if using the .05 level of significance, you would expect at least one statistically significant difference even if no differences exist. Therefore, various methods have been developed for doing multiple comparisons of group means.