One way ANOVA

In this session we discuss the basic concepts in Analysis of Variance techniques and one way Analysis of variance for fixed effects. The entire topic is divided into the following subdivisions.

1. Gauss-Markoff Linear Model
2. Analysis of variance Technique
3. One way ANOVA
4. One way ANOVA Model
5. Decomposition of the total sum of squares
6. Statistical Analysis
7. ANOVA Table
8. Comparing pairs of treatment means
9. Conclusion

1. Gauss-Markoff Linear Model

Consider a set of an independent random variables y_i whose expectations are given as linear functions of p unknown parameters say $\beta_1, \beta_2, \ldots \beta_p$, ($p \leq n$) with known coefficients a_{ij}'s and whose variances are a constant σ^2.

Then $E(y_i) = a_{i1}\beta_1+a_{i2}\beta_2+\ldots a_{ip}\beta_p$, i=1…n

$V(y_i) = \sigma^2$, $Cov(y_i, y_j) = 0$, $i \neq j$.

(1)

Writing $y= (y_1,y_2,\ldots y_n)^1$; $\beta = (\beta_1, \beta_2, \ldots \beta_p)^1$; $\alpha_j=(a_{1j}, a_{2j},\ldots a_{nj})^1$, j=1...p($\leq$n)

$A = (\alpha_1, \alpha_2, \ldots \alpha_p)$

Then model (1) can be written as

$E(y) = A\beta$

$D(y) = \sigma^2 I$ where D(y) denote the dispersion matrix of y and I is identity matrix of order n.

(2)

Or Model (2) can also be written as

$Y = A \beta + e$, where $E(e) = 0$ and $D(e) = \sigma^2 I$

(3)

Model (1) or (2) or (3) is known as Gauss Markov Model.

Estimable function: A linear function $a^1\beta$ of parameters with known "a" is said to be an estimable parametric function (or estimable) if there exists a linear function $L^1y$of Y such that $E(L^1y) = a^1\beta$ for all β. [where $L = (l_1,l_2,\ldots l_p)^1$; $y=(y_1,y_2,\ldots y_n)^1$, $a=(a_1,a_2,\ldots a_p)^1$, $\beta = (\beta_1, \beta_2, \ldots \beta_p)^1$]

Theorem 1: A linear parametric function $L^1\beta$ admits a unique least square estimate if and only if $L^1\beta$ is estimable.

Theorem 2(Gauss Markoff Theorem): If the linear parametric function $L^1\beta$ is estimable then the linear estimator $L^1\hat{\beta}$ where $\hat{\beta}$ is a solution of $X^1X\hat{\beta} = X^1Y$ is the best linear unbiased estimator of $L^1\beta$ in the sense of having minimum variance in the class of all linear and unbiased estimators of $L^1\beta$

Gauss Markoff theorem states that in a linear regression model in which the errors have expectations zero and are uncorrelated and have equal variances, the best linear unbiased estimator(BLUE), i.e. giving the lowest variance of the estimate as compared to other unbiased estimators of the coefficients is given by the ordinary least squares

The same linear model(Gauss Markoff model) is used in the analysis of variance Technique.

The unknown parameters βj are called effects (due to j^{th} cause), j=1,2,…p. When all the parameters β_j are fixed constants, the model (1) is called fixed effect model. In such case, if one of the coefficient a_{ij}, i=1,2…n say $a_{i1} = 1$ for every i=1,2..n so that the model can be written as $E(y_i) = \beta_1+a_{i2}\beta_2+\ldots a_{ip}\beta_p$, then β_1 is an additive constant and is called a general effect.

2. Analysis of Variance (ANOVA) **Technique**: Provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes *t*-test to more than two groups. ANOVAs are helpful because they possess an advantage over a two-sample t-test. Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. For this reason, ANOVA are useful in comparing two, three or more means. The technique in the analysis of variance involves breaking down of total variation into orthogonal components. Each orthogonal components represents the variation due to a particular factor contributing in the total variation.

Assumptions:

1. The observations are independent
2. Parent population from which observation are taken is normal
3. Various treatments(factors) and environmental effects are additive in nature
4. The distributions of the residuals are normal. i.e. random errors are distributed as iid normal with mean 0 and variance σ^2.

3. One way ANOVA:

The objective in the one way classification (one way ANOVA) is to test the hypothesis about the equality of several means on the basis of several samples which have been drawn from univariate normal populations with different means but the same variances.

Example.The yield of milk may be affected by different treatments(diet) fed to the cows. Our interest is to study the effect of these different type of diet(food) given at random to these cows on the average milk production. Diet is here referred as treatment. Response variable is the yield of milk.

Suppose we have p treatments and we wish to compare these p treatments on the response usually referred as yield. . The observed response from each of the p treatments is a random variable. The data can be tabulated as follows.

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	..	y_{1n_1}	$y_{1.}$	$\overline{y}_{1.}$
2	y	y	..	y	$y_{2.}$	$\overline{y}_{2.}$

	2 1	2 2 .		2 n 2		
.	.	.	..	.	.	.
..	.	.	..	.	.	.
.	.	.	..	.	.	.
p	y p 1	y p 2	..	y p n p	$y_{p.}$	$\overline{y}_{p.}$

Y_{ij} represent the j^{th} observation taken from treatment i. There will be in general n_i observations under the i^{th} treatment

Let $y_{i.}$ represent the total of the observations and $\overline{y}_{i.}$ average of observations under the i^{th} treatment, Similarly $y_{..}$ represent the grand total of all the observations and $\overline{y}_{..}$ represent grand average of all the observations. And are expressed as follows.

$$y_{i.} = \sum_{j=1}^{n} y_{ij}, \qquad \overline{y_{i.}} = \frac{y_{i\cdot}}{n_i}, \quad i=1,2\ldots p$$

$$y_{..} = \sum_{i=1}^{p} \sum_{j=1}^{ni} y_{ij} \qquad \overline{y}_{..} = \frac{y_{\cdot\cdot}}{N},$$ where N is total number of observations N= $\sum_{i=1}^{p} n_i$

4. **One way ANOVA Model.**: We define the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i=1,2\ldots p, \; j=1,2,\ldots n_i \tag{4}$$

Where μ general effect

α_i – i^{th} treatment effect;

ϵ_{ij}- error term which are independent and identically distributed normal random variables with mean 0 and variance σ2

Observe that the response variable y_{ij} is a liner function of the model parameters.

Equation (4) is called the one way analysis of variance model because only one factor is investigated.

Here we are interested in testing the equality of p treatment means.

Hence the appropriate hypotheses is

H0: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$(i.e. all the treatment has the same effect on the yield)

Against the alternative

H1: atleast one $\alpha_i \neq \alpha_j$, for all I,j (i.e. the treatment effects are different)

5. **Decomposition of the total sum of squares:**

The analysis of variance technique is partitioning the total variability into its components parts.

The corrected total sum of squares is

$SST == \sum_{i=1}^{p} \sum_{j=1}^{ni} (y_{ij} - \overline{y_{..}})^2$ and is used as overall variability in the data.

Consider

$$\sum_{i=1}^{p} \sum_{j=1}^{ni} (y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p} \sum_{j=1}^{ni} (y_{ij} - \overline{y_{i..}} + \overline{y_{i..}} - \overline{y_{..}})^2$$

$$= \sum_{i=1}^{p} n_i (\overline{yi.} - \overline{y_{..}})^2 + \sum_{i=1}^{p} \sum_{j=1}^{ni} (yij - \overline{yi_{..}})^2 + 2$$

$$\sum_{i=1}^{p} \sum_{j=1}^{ni} (\overline{y_i.} - \overline{y..})(y_{ij} - \overline{y_{i..}})$$

The third term becomes zero hence we have

$$\sum_{i=1}^{p} \sum_{j=1}^{ni} (y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p} n_i (\overline{yi.} - \overline{y_{..}})^2 + \sum_{i=1}^{p} \sum_{j=1}^{ni} (yij - \overline{yi_{..}})^2 \quad 6$$

The total corrected sum of squares can be partitioned into a sum of squares of the differences between the treatment averages and the grand average, plus a sum of squares of the differences of the observations within treatments averages. The differences between the observed treatment averages and the grand average is a measure of the differences between treatment means where as the differences of observations with in a treatment from the treatment averages can be due to only random error.

Thus we can write the equation 6 as

SST = SSTR + SSE where SSTR is called the sum of squares due to treatments .and SSE is called the sum of squares due to error.

Where SST = $\sum_{i=1}^{p}\sum_{j=1}^{ni}(y_{ij}-\overline{y_{..}})^2 = \sum_{i=1}^{p}\sum_{j=1}^{ni}y_{ij}^2 - \frac{y_{..}^2}{N}$ (simplified formula)

SSTR = $\sum_{i=1}^{p} n_i(\overline{yi.}-\overline{y_{..}})^2$ = $\sum_{i=1}^{p}\frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$

There are N total observations, thus SST has N-1 degrees of freedom, There are p treatment means so SSTR has p-1 degree of freedom, and (N-1)-(p-1) = N-p degrees of freedom for error.

6. **Statistical Analysis**:

We now investigate a formal test of the hypothesis

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

Against the alternative

H_1: atleast one $\alpha_i \neq \alpha_j$, for all I,j

We have assumed that the errors ϵij are normally and independently distributes with mean zero and variance σ^2. The observations y_{ij} are normally and independently distributed with mean μ + αi and variance σ^2. Thus SST is a sum of squares in normally distributed random variables hence can be shown that SST/σ^2 is distributes as chi-square with N-1 degree freedom .Further we can show that SSE/σ^2 is chi-square N-p degrees of freedom and that SSTR /σ^2 is chi - square variate with p-1 degrees of freedom if the null hypothesis i.e. H0 : $\alpha_i = 0$ is true. It also implies that SSTR/ σ^2 and SSE/σ^2 independently distributed chi-square random variables. Therefore if the null hypothesis is true, the ratio

$$F_\alpha = \frac{SSTR/(p-1)}{SSE/(N-p)} = \frac{MSSTR}{MSSE} \qquad (8)$$

is distributed as F with p-1 and N-p degrees of freedom.

Equation (8) is the test statistics for testing H_0: there is no differences in treatment means.

7. **Analysis of variance Table(ANOVA Table)**

Sources of variation	Degree of freedom	Sum of squares	Mean sum of squares	F-Value
Treatments	p-1	SSTR	MSSTR = SSTR/(p-1)	MSSTR/ MSSE
Error	N-p	SSE	MSSE = SSE/(p-1) (k-1)	
total	N-1	TSS		

We reject H_0 and conclude that there are differences in the treatment means if $F_0 > F_{\alpha,,p-1, N-p}$ Where F_0 is computed from equation 8 and Fα, p-1, N-p, is the table value referring to F table at α level significance corresponding to p-1and N-p degrees freedom.

8. **Comparing pairs of treatment means**

Case of rejection of H_0.:

if $F_0 > F_{\alpha,,p-1, N-p,}$ then

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

is rejected. This means that at least one αi

is different from other effects which is responsible for the rejection of the H0. Now our interest is to find out such αi and divide the population into groups such that the means of the population with in the groups are same. This can be done by pairwise testing of αi's.

i.e. we test H_0: αi = αi , (i≠j) against H1: αi≠ αi

This can be tested using the following t-statistics:

$$T = \frac{\overline{y_{i.}} - \overline{y_{j.}}}{\sqrt{MESS\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$ which follows t distribution with N-p degrees of freedom

under H0.

Thus the decision rule to reject H0 at α if the observed difference

$\overline{y_{i.}} - \overline{y_{j.}} > t_{\alpha,n-p}\sqrt{MESS\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ The quantity $t_{\alpha,n-p}\sqrt{MESS\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ is called critical difference. Where $t_{\alpha, N-p}$ is the table value referring to t table at α level significance corresponding to N-p degrees freedom.

9. Conclusion: WE have introduced the Gauss- moakoff linear model, defined estimable function and defined Gauss markoff theorem. Discussed the meaning of Analysis of Variance Technique We have presented the one way analysis of variance model, stated the hypothesis to be tested and discussed the analysis . The entire analysis procedure we summarized in ANOVA Table.. We have discussed the procedure to find out the factors which are responsible for rejecting H0 when the stated hypothesis H0 is rejected.