UNIT 6 ONE-WAY ANALYSIS OF VARIANCE

Structure

6.1 INTRODUCTION

The analysis of variance is one of the powerful techniques of statistical analysis. Analysis of variance is used for testing of equality of means of several populations. It tests the variability of the means of the several populations. In the previous unit, we have discussed about the fundamental terms which are used in the analysis of variance. In that unit, we have also discussed the basic assumptions and models of analysis of variance.

As we have stated that the analysis of variance technique can be divided into two categories (i) parametric ANOVA and (ii) Non-parametric ANOVA. The parametric ANOVA can also be classified as one-way ANOVA if only one response variable is considered and MANOVA if two or more response variables are considered.

In this unit, we shall discuss the one-way analysis of variance. One-way analysis of variance is a technique where only one independent variable at different levels is considered which affects the response variable.

In this unit, the one-way analysis of variance model is discussed in Section 6.2. The basic assumptions under one-way analysis of variance are described in Section 6.3 whereas the estimates of each level mean of a factor are derived in Section 6.4. Test of hypothesis method is explained in Section 6.5 and the degrees of freedom for various sum of squares are determined in Section 6.6. The expectations of various sum of squares are derived in Section 6.7 whereas the analysis of variance table for one-way classification is described in Section 6.8.

Objectives

After studying this unit, you would be able to

- describe the one-way analysis of variance model;

- describe the basic assumptions under one-way analysis of variance;
- estimate of each level mean of a factor;
- determine the degrees of freedom for various sum of squares;
- derive the expectations of various sum of squares;
- construct the ANOVA table;
- test the hypothesis under one-way analysis of variance;
- identify differences in population means of k level of a factor; and
- determine which population means are different using multiple comparison methods.

6.2 ONE -WAY ANALYSIS OF VARIANCE MODEL

One-factor analysis of variance or one-way analysis of variance is a special case of ANOVA, for one factor of variable of interest and a generalization of the two sample t-test. The two sample t-test is used to decide whether two groups (two levels) of a factor have the same mean. One-way analysis of variance generalizes this to k levels (greater than two) of a factor.

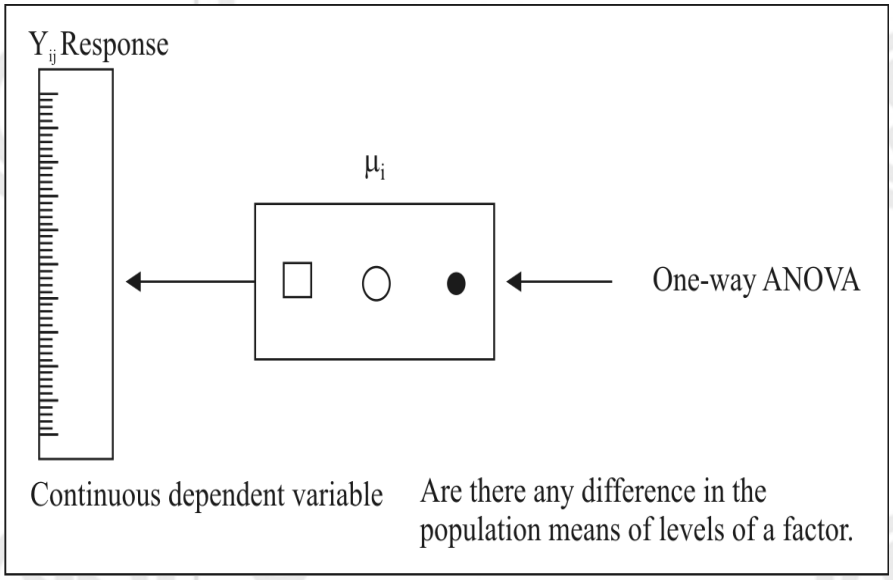
In the following, subscript i refers to the i^{th} level of the factor and subscript j refers to the j^{th} observation within a level of factor. For example y_{23} refers to third observation of the second level of a factor.

The observations on different levels of a factor can be exhibited below:

Level of a factor	**Observations**				**Totals**	**Means**
1	y_{11}	y_{12}	.....	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	.....	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.	…….	.	.	.
.	.	.	……...	.	.	.
.	.	.	…….	.	.	.
i	y_{i1}	y_{i2}	…….	y_{in}	$y_{i.}$	$\bar{y}_{i.}$
.	.	.	……...	.	.	.
.	.	.	……...	.	.	.
.	.	.	…….	.	.	.
k	y_{k1}	y_{k2}	.....	y_{kn}	$y_{k.}$	$\bar{y}_{k.}$

The linear mathematical model for one-way classified data can be written as

$$y_{ij} = \mu_i + e_{ij} \qquad \text{where, } i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n$$



Total observations are $n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = N$

Here, y_{ij} is continuous dependent or response variable, where μ_i is discrete independent variable, also called a explanatory variable.

This model decomposes the responses into a mean for each level of a factor and error term i.e.

Response = A mean for each level of a factor + Error term

The analysis of variance provides estimates for each level mean. These estimated level means are the predicted values of the model and the difference between the response variable and the estimated/predicted level means are the residuals.

That is

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

$$e_{ij} = y_{ij} - \mu_i$$

The above model can be written as $y_{ij} = \mu + (\mu_i - \mu) + e_{ij}$

or $y_{ij} = \mu + \alpha_i + e_{ij}, \quad \forall\ i = 1, 2, ..., k \text{ and } j = 1, 2, ..., n_i$

where, $\alpha_i = \mu_i - \mu$

A general mean effect given by

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

This model decomposes the response into an over all (grand) mean, the effect of the i^{th} factor level α_i and error term $e_{ij.}$ The analysis of variance provides estimates of the grand mean μ and the effect of the i^{th} factor level α_i. The predicted values and the responses of the model are

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

$$e_{ij} = y_{ij} - \mu - \alpha_i$$

α_i is the effect of the i^{th} level of the factor and given by

$$\alpha_i = \mu_i - \mu \qquad \forall \quad i = 1, 2, \ldots, k. \qquad \ldots (1a)$$

i.e. if the effect of i^{th} level of a factor increases or decreases in the yield/response variable by an amount α_i, then

$$\sum n_i \alpha_i = \sum n_i (\mu_i - \mu)$$

$$\sum n_i \alpha_i = \sum n_i \mu_i - \mu \sum n_i$$

$$= N\mu - N\mu$$

$$= 0$$

Under 6^{th} assumption, given in Section 6.3, the model becomes

$$E(y_{ij}) = \mu_i, \qquad \forall \quad i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n_i.$$

or $$E(y_{ij}) = \mu + \alpha_i, \qquad \forall \quad i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n_i.$$

6.3 BASIC ASSUMPTIONS OF ONE-WAY ANALYSIS OF VARIANCE

The following are the basic assumption of one-way ANOVA:

1. Dependent variable measured on interval scale;
2. k sample are independently and randomly drawn from the population;
3. Population can be reasonably to have a normal distribution;
4. k samples have approximately equal variance;
5. Various effects are additive in nature; and
6. e_{ij} are independently identically distributed normal with mean zero and variance σ_e^2.

Now, when we discuss the step by step computation procedure for one-way analysis of variance for k independent sample, the first step of the procedure is to make the null and alternative hypothesis.

We want to test the equality of the population means, i.e. homogeneity of effect of different levels of a factor. Hence, the null hypothesis is given by

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

Against the alternative hypothesis

$$H_1: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k \text{ (or some } \mu_i\text{'s are not equal)}$$

which, reduces to

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

Against the alternative hypothesis

$$H_1: \alpha_1 \neq \alpha_2 \ldots . \neq \alpha_k \neq 0 \text{ (or at least some } \alpha_i\text{'s are not zero)}$$

6.4 ESTIMATION OF PARAMETERS

The parameters μ and $\alpha_1, \alpha_2, ..., \alpha_k$ are estimated by the principle of least square on minimizing the error (residual) sum of squares. The residual sum of squares can be obtained as

$$e_{ij} = y_{ij} - \mu - \alpha_i \qquad \forall \quad i = 1, 2, \ldots, k \;\&\; j = 1, 2, \ldots, n_i.$$

$$e_{ij}^2 = (y_{ij} - \mu - \alpha_i)^2$$

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right)^2$$

By this residual sum of squares E, we partially differentiate it with respect to μ and partially differentiating with respect to $\alpha_1, \alpha_2, ..., \alpha_k$ and then equating these equations to 0, we get

$$\frac{\partial E}{\partial \mu} = -2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

or $$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} - N\mu - \sum_{i=1}^{k} n_i\alpha_i = 0, \quad \text{because} \quad kn_i = N$$

or $$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} - N\mu = 0, \quad \text{because} \sum_{i=1}^{k} n_i\alpha_i = 0$$

or $$N\mu = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

Therefore, $$\mu = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}$$

Similarly,

$$\frac{\partial E}{\partial \alpha_1} = -2\sum_{j=1}^{n_1}\left(y_{1j} - \mu - \alpha_1\right) = 0$$

$$\sum_{j=1}^{n_1} y_{1j} - n_1\mu - n_1\alpha_1 = 0$$

Or $$n_1\alpha_1 = \sum_{j=1}^{n_1} y_{1j} - n_1\mu$$

$$\alpha_1 = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} - \hat{\mu}$$

Therefore, $\alpha_1 = \bar{y}_{1.} - \bar{y}_{..}$

Similarly,

$$\alpha_2 = \bar{y}_{2.} - \bar{y}_{..}$$

Or in general $\alpha_i = \bar{y}_{i.} - \bar{y}_{..} \qquad \forall \; i = 1, 2, \ldots, k.$

6.5 TEST OF HYPOTHESIS

Small differences between sample means are usually present. The objective is to determine whether these differences are significant or in other words, are the difference more than what might be expected to occur by chance? If the differences are more than what might be expected to occur by chance, you have sufficient evidence to conclude that there are differences between the population means of different levels of a factor.
The hypothesis is :

H_0: Population means of k levels of a factor are equal.

H_1: At least one population mean of a level of a factor is different from the population means of other levels of the factor.

or $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

$H_1: \mu_1 \neq \mu_2 = \ldots \neq \mu_k$

Now, substituting these estimated values in the model given in equation (1), the model becomes

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

and then $$e_{ij} = (y_{ij} - \bar{y}_{..}) - (\bar{y}_{i.} - \bar{y}_{..})$$

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

Now substituting these values in equation (1) we get

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

Transporting by $\bar{y}_{..}$ to the left and squaring both sides and taking sum over i and j.

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

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

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

But $\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) = 0$, since the sum of the deviations of the observations from their mean is zero.

Therefore,

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

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

where,

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$ is known as total sum of squares (TSS), $\sum_{i=1}^{k}n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ is called between sum of squares or treatment sum of squares or sum of squares

due to different levels of a factor (SST) and $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$ is called within sum of squares or residual sum of squares or error sum of squares (SSE).

So, TSS = SST + SSE

6.6 DEGREE OF FREEDOM OF VARIOUS SUM OF SQUARES

The Total Sum of Squares (TSS) which is computed from the N quantities of the form (y_{ij} - $\bar{y}$ ..) will carry (N−1) degrees of freedom. One degree of freedom lost because of linear constraints

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)=0$$

Similarly, the Treatment Sum of Squares (SST) = $\sum_{i=1}^{k}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2$ will have (k−1) degrees of freedom.

Since $\sum_{i=1}^{k} n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)$ = 0 and the Error Sum of Squares i.e

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

will have (N − k) degrees of freedom, since it is based upon N observations which are subject to k linear constraints

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right) = 0 \text{ for } i = 1, 2, ..., k$$

Hence, degrees of freedom of various sum of squares also additive same as various sum of squares (TSS = SST + SSE), i.e. degrees of freedom

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

Mean Sum of Squares

The sum of squares divided by its degrees of freedom is called Mean Sum of Squares (MSS). Therefore,

MSS due to treatment (MSST) = SST/df = SST/ (k−1).

MSS due to error (MSSE) = SSE/df = SSE/ (N−k).

6.7 EXPECTATIONS OF VARIOUS SUM OF SQUARES

For obtaining appropriate test statistics for testing H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$, we have to find out expectations of various sum of squares. Our linear model is

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

We have

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mu + \alpha_i + e_{ij}\right)$$

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mu + \frac{1}{n_i}\sum_{j=1}^{n_i}\alpha_i + \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}$$

$$\bar{y}_{i.} = \mu + \alpha_i + \bar{e}_{i.} \qquad \ldots (2)$$

Now,

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mu + \alpha_i + e_{ij}\right)$$

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\alpha_i + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij} \quad \text{because} \sum_{i=1}^{k}\alpha_i = 0$$

$$\bar{y}_{..} = \mu + \bar{e}_{..} \qquad \ldots (3)$$

6.7.1 Expectation of Treatment Sum of Squares

$$E(SST) = E\left[\sum_{i=1}^{k} n_i\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2\right] \qquad \ldots (4)$$

Substituting the value of $\bar{y}_{i.}$ and $\bar{y}_{..}$ from equations (2) & (3), we get

$$E\,(SST) = E\left[\sum_{i=1}^{k} n_i\left\{\left(\mu + \alpha_i + \bar{e}_{i.}\right) - \left(\mu + \bar{e}_{..}\right)\right\}^2\right]$$

$$= E\left[\sum_{i=1}^{k} n_i\left(\alpha_i + \bar{e}_{i.} - \bar{e}_{..}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{k} n_i\left[\alpha_i^2 + \left(\bar{e}_{i.} - \bar{e}_{..}\right)^2 + 2\alpha_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)\right]\right]$$

$$= \sum_{i=1}^{k} E\left(n_i\alpha_i^2\right) + \sum_{i=1}^{k} E\left(n_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)^2\right) + 2\sum_{i=1}^{k} E\left(n_i\alpha_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)\right)$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\sum_{i=1}^{k} n_i\left(\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}\right) + 2\sum_{i=1}^{k} n_i\alpha_i E\left(\bar{e}_{i.} - \bar{e}_{..}\right)$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\sum_{i=1}^{k} n_i\left(\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}\right) + 0 \quad \left[\text{because } E\left(\bar{e}_{i.} - \bar{e}_{..}\right) = 0\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\left[\sum_{i=1}^{k} n_i\bar{e}_{i.}^2 + \sum_{i=1}^{k} n_i\bar{e}_{..}^2 - 2\sum_{i=1}^{k} n_i\bar{e}_{i.}\bar{e}_{..}\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + \sum_{i=1}^{k} n_i E\left(\bar{e}_{..}^2\right) - 2E\left[\sum_{i=1}^{k} n_i\bar{e}_{i.}\bar{e}_{..}\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + NE\left(\bar{e}_{..}^2\right) - 2E\left[\bar{e}_{..}\sum_{i=1}^{k} n_i\bar{e}_{i.}\right]$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + N E\left(\bar{e}_{..}^2\right) - 2E\left(\bar{e}_{..} N\bar{e}_{..}\right) \quad \left(\text{because } \bar{e}_{..} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{e}_{i.}\right)$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) - NE\left(\bar{e}_{..}^2\right) \quad \ldots (5)$$

Since, $e_{ij} \sim$ iid $N(0, \sigma_e^2)$ & $E(e_{ij}) = 0$

$V(e_{ij}) = E(e_{ij}^2) - (E(e_{ij}))^2$

$\sigma_e^2 = E(e_{ij}^2)$ … (6)

$E(\bar{e}_{i.}) = 0$ and $E(\bar{e}_{..}) = 0$

$V(\bar{e}_{i.}) = E(\bar{e}_{i.}^2) - (E(\bar{e}_{i.}))^2 = \sigma_e^2/n_i$

or $E(\bar{e}_{i.}^2) = \sigma_e^2/n_i$ … (7)

$V(\bar{e}_{..}) = E(\bar{e}_{..}^2) - (E(\bar{e}_{..}))^2 = \sigma_e^2/N$

or $E(\bar{e}_{..}^2) = \sigma_e^2/N$ … (8)

By Substituting values from equations (6), (7) and (8) in equation (5), we get

$$E(SST) = \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i \frac{\sigma_e^2}{n_i} - N\frac{\sigma_e^2}{N}$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + k\sigma_e^2 - \sigma_e^2$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + (k-1)\sigma_e^2$$

Dividing both side by (k - 1), we get

$$E\left(\frac{SST}{k-1}\right) = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i \alpha_i^2 + \sigma_e^2$$

$$E(MSST) = \sigma_e^2 + \frac{1}{(k-1)}\sum_{i=1}^{k} n_i \alpha_i^2$$

The second term in above equation like as variance since $\alpha_i = \mu_i - \mu$.
Therefore, under H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$, it is zero.

MSS due to treatment provides an unbiased estimate of σ_e^2 under H_0.

6.7.2 Expectation of Sum of Squares due to Error

$$E(SSE) = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\right]$$

Substituting the value of y_{ij} and $\bar{y}_{i.}$ from equations (1) & (2), we get

$$E(SSE) = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\left(\mu + \alpha_i + e_{ij}\right) - \left(\mu + \alpha_i + \bar{e}_{i.}\right)\right\}^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(e_{ij} - \bar{e}_{i.}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(e_{ij}^2 + \bar{e}_{i.}^2 - 2e_{ij}.\bar{e}_{i.}\right)\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 + \sum_{i=1}^{k} n_i \bar{e}_{i.}^2 - 2\sum_{i=1}^{k} \bar{e}_{i.}\left(\sum_{j=1}^{n_i} e_{ij}\right)\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 + \sum_{i=1}^{k} n_i \bar{e}_{i.}^2 - 2\sum_{i=1}^{k} n_i \bar{e}_{i.}^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 - \sum_{i=1}^{k} n_i \bar{e}_{i.}^2\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} E\left(e_{ij}^2\right) - \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \sigma_e^2 - \sum_{i=1}^{k} n_i \frac{\sigma_e^2}{n_i}$$

$$= N\,\sigma_e^2 - k\,\sigma_e^2$$

$$= (N - k)\,\sigma_e^2$$

$$E\,(SSE/\,N-k) = \sigma_e^2$$

$$E\,(MSSE) = \sigma_e^2$$

Therefore, the error mean squares always gives an unbiased estimate of σ_e^2.

Under H_0, E (MSST) = E (MSSE)

Otherwise, E (MSST) > E (MSSE)

Hence, the test statistics for testing H_0 is provided by the variance ratio or Snedecor's

F = MSST/MSSE with [(k−1), (N−k)] df.

Thus, if an observed value of F is greater than the tabulated value of F for {(k−1), (N−k)} df and specific level of significance (usually 5% or 1%), then H_0 is rejected otherwise, it may be accepted.

6.8 ANOVA TABLE FOR ONE-WAY CLASSIFICATION

The above analysis is presented in the following table:

ANOVA Table for One-way Classified Data

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio F
Treatments	k−1	$\sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = SST$	MSST = SST/(k−1)	F = MSST/MSSE With{(k−1), (N−k)}df
Error	N−k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = SSE$	MSSE = SSE/(N−k)	
Total	N−1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = TSS$		

If the treatment show significance effect at α level of significance then we are interested to find out which pair/pairs of treatments differ significantly, say the hypothesis H_{01}: $\mu_i = \mu_j$ or the null hypothesis is rejected.
With the help of

$$t = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{MSSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad \text{with df } N - k = k(n-1)$$

If $n_1 = n_2 = \ldots = n_k = n$ then this reduces to the

$$t = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{\frac{2MSSE}{n}}}$$

or $$|t| = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{\frac{2MSSE}{n}}} \quad \text{with } k(n-1) \text{ df}$$

If $t > t_{\alpha/2}[k(n-1)]$, then H_{01} is rejected at the level α. That is to say, H_{01} is rejected at the level of significance α if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq |t|_{\alpha/2[k(n-1)]} \times \sqrt{\frac{2MSSE}{n}}$$

Thus, to compare the factor level effect/group means two at a time, we have to calculate $|t|_{\alpha/2}$ (with df $k(n-1)$) $\times \sqrt{\frac{2MSSE}{n}}$

which is called the critical difference (CD) or the least significant difference (lsd) and if the difference between the observed class/groups /factor level effect means, i.e. $(\bar{y}_{i.} - \bar{y}_{j.})$, is greater than the CD, then H_{01}: $\mu_i = \mu_j$ is rejected at α level of significance, otherwise it is accepted. Here, $t_{\alpha/2}[k(n-1)]$ is the tabulated value of t-distribution with $k(n-1)$ df at upper $\alpha/2$ point.

Example 1: An investigator is interested to know the level of knowledge about the history of India of 4 different schools in a city. A test is given to 5, 6, 7, 6 students of 8th class of 4 schools. Their scores out of 10 is given below:

School I (S_1)	8	6	7	5	9		
School II (S_2)	6	4	6	5	6	7	
School III(S_3)	6	5	5	6	7	8	5
School IV(S_4)	5	6	6	7	6	7	

Solution: If $\mu_1, \mu_2, \mu_3, \mu_4$ denote the average score of students of 8th class of schools I, II, III, IV respectively. Then

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis H_1: Difference among $\mu_1, \mu_2, \mu_3, \mu_4$ are significant.

S.No.	S_1		S_2	S_3	S_4	S_1^2	S_2^2	S_3^2	S_4^2
1	8		6	6	5	64	36	36	25
2	6		4	5	6	36	16	25	36
3	7		6	5	6	49	36	25	36
4	5		5	6	7	25	25	36	49
5	9		6	7	6	81	36	49	36
6			7	8	7		49	64	49
7				5				25	
Total	35		34	42	37	255	198	260	231

Grand Total G = 35 + 34 + 42 + 37 = 148

Correction Factor (CF) $= \frac{G^2}{N} = \frac{148^2}{24}$, Since $N = n_1 + n_2 + n_3 + n_4$

$= 912.6667$

Raw Sum of Squares (RSS) $= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 = 255 + 198 + 260 + 231 = 944$

Total Sum of Squares (TSS) = RSS – CF = 944 – 912.6667 = 31.3333

Sum of Squares due to Treatments (SST)

$$= \frac{T_{1.}^2}{n_1} + \frac{T_{2.}^2}{n_2} + \frac{T_{3.}^2}{n_3} + \frac{T_{4.}^2}{n_4} - CF$$

$$= \frac{35^2}{5} + \frac{34^2}{6} + \frac{42^2}{7} + \frac{37^2}{6} - 912.6667$$

$$= 245 + 192.6667 + 252 + 228.1667 - 912.6667$$

$$= 5.1667$$

Sum of Squares due to Errors (SSE) = TSS − SST

$= 31.3333 - 5.1667 = 26.1666$

Now, $MSST = \frac{SST}{k-1} = \frac{5.1667}{3} = 1.7222$

$MSSE = \frac{SSE}{N-k} = \frac{26.1667}{20} = 1.3083$

ANOVA Table

Sources of Variation	DF	SS	MS	F
Between schools	3	5.1667	1.7222	$F = \frac{1.7222}{1.3083} = 1.3164$
Within schools	20	26.1666	1.3083	

Calculated F = 1.3164

Tabulated F at 5% level of significance with (3, 20) degree of freedom is 3.10.

Conclusion: Since Calculated F < Tabulated F, so we may accept H_0 and conclude that level of knowledge of schools I, II, III and IV do not differ significantly.

Example 2: If we have three fertilizers and we have to compare their efficacy, this could be done by a field experiment in which each fertilizer is applied to 10 plots, and then 30 plots are later harvested, with the crop field being calculated for each plot. The data were recorded in following table:

Fertilizer	**Yields (in tones) from the 10 plots allocated to that fertilizer**									
1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
2	3.07	3.29	4.04	4.19	.41	0.75	04.87	3.94	6.49	3.15
3	4.04	3.79	4.56	4.55	4.53	3.53	3.71	7.00	4.61	4.55

Solution:

H_0: Mean effect of Ist fertilizer = Mean effect of the IInd fertilizer = Mean effect IIIrd fertilizer

$H_0: \mu_1 = \mu_2 = \mu_3$

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

Steps for calculating different sum of squares

Grant Total = Total of all observation = $\sum\sum y_{ij} = G = 139.20$

Correction Factor (CF) = $G^2/N = 139.20 \times 139.20 / 30 = 645.89$

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^2 = 6385.3249$

Total Sum of Squares = RSS − CF =36.4449

Sum of Squares due to Treatments/Fertilizer (SST)

$$= \frac{y_{1.}^2}{10} + \frac{y_{2.}^2}{10} + \frac{y_{3.}^2}{10} - CF$$

$$= (54.5)^2/10 + (40)^2/10 + (44.9)^2/10 - CF$$

$$= 10.8227$$

Sum of Squares due to Error = TSS − SST = 36.4449 −10.8227

= 25.6222

Mean Sum of Squares due to Treatments/Fertilizers (MSST) = SST/df

= 10.8227/2 = 5.4114

Mean Sum of Squares due to Error (MSSE) = SSE/df

= 25.6221/27 = 0.9490

Variance ratio $F_{2,27}$ = MSST/MSSE = 5.414/0.9490 = 5.70

Tabulated $F_{2,27}$ = 3.35

Since calculated value of $F_{2,27}$ is greater than tabulated $F_{2,27}$ at 5% level of significance so we reject H_0. It means there is a significant difference among the effect of these three fertilizers.

Now, H_0 is rejected. So, we have to test which of the fertilizers most important among the pairwise comparison. So, pairwise comparison test will be applied.

To test $H_0: \mu_i = \mu_j$

against $H_1: \mu_i \neq \mu_j$

We have a test statistic

$$\left|\bar{y}_{i.} - \bar{y}_{j.}\right| \geq \left|t\right|_{\alpha/2}(N-k) \times \sqrt{\frac{2MSE}{10}}$$

or $$\left|\bar{y}_{1.} - \bar{y}_{2.}\right| = 5.45 - 4.0 = 1.45 \quad \ldots (9)$$

$$\left|\bar{y}_{2.} - \bar{y}_{3.}\right| = (4.00 - 4.49) = 0.49 \quad \ldots (10)$$

$$\left|\bar{y}_{1.} - \bar{y}_{3.}\right| = (5.45 - 4.49) = 0.96 \quad \ldots (11)$$

and

$$t_{\alpha/2} \text{ at } 27 \text{ df} \times \sqrt{\frac{2(0.9490)}{10}} = 2.05 \times \sqrt{\frac{1.8980}{10}}$$

$$= 2.05 \times 0.044 = 0.9020$$

Since, the calculated value of difference of mean effect 1 with 2 and 3 is greater than the value of the critical difference (0.9020). So, effect of fertilizer 1 has significant difference with effect of fertilizers 2 and 3. But there is no significant difference between the effects of fertilizer 2 and 3. Since mean value of fertilizer 1 $\left(\bar{y}_{1.} = 5.47\right)$ is greater than mean values of fertilizer 2 or 3 so we can say that fertilizer 1 can be preferred in comparison to fertilizer 2 and 3.

E 1) Three varieties A, B and C of wheat are shown in five plots each of the following fields per acre as obtained:

Plots	A	B	C
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Set up a table of analysis of variance and find out whether there is significant difference between the fields of these varieties.

E 2) The following figures relate to production in kg. of three varieties P, Q, R of wheat shown in 12 plots

P	14	16	18		
Q	14	13	15	22	
R	18	16	19	15	20

Is there any significant difference in the production of these varieties?

E 3) In 25 plots four varieties ν_1, ν_2, ν_3, ν_4 of wheat are randomly put and their yield in kg are shown below.

ν_1 2000	ν_3 2270	ν_2 2230	ν_4 2270	ν_4 2180
ν_2 2160	ν_1 2100	ν_2 2050	ν_3 2300	ν_2 2280
ν_1 2200	ν_1 2300	ν_4 2040	ν_3 2420	ν_1 2240
ν_4 2370	ν_1 2250	ν_2 2040	ν_2 2360	ν_1 2460
ν_3 2210	ν_1 2340	ν_2 2190	ν_1 2150	ν_3 2020

Perform the ANOVA to test whether there is any significant difference between varieties of wheat.

6.9 SUMMARY

In this unit, we have discussed:

1. The one-way analysis of variance model;
2. The basic assumptions in one-way analysis of variance;
3. Estimation of parameters of one-way analysis of variance model;
4. Test of hypothesis for one-way classified ANOVA;
5. How to obtain the expectation of various sum of squares in one-way ANOVA; and
6. The construction of one-way ANOVA table.

6.10 SOLUTIONS /ANSWERS

E1) Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ i.e. the mean fields of three varieties is the same,

Against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

The calculation is done on the basis of the given data and the results are as follows:

$G = \sum\sum y_{ij}$ = Sum of all observations

G = 8+10+7+14+11+7+5+10+9+9+12+9+13+12+14 = 150

N = Total number of observations = 15

Correction factor $(CF) = \frac{G^2}{N} = \frac{150 \times 150}{15} = 1500$

Raw Sum of Squares (RSS) $= \sum\sum y_{ij}^2$

$= 8^2 + 10^2 + 7^2 + 14^2 + 11^2 + 7^2 + 5^2 + 10^2$

$$+9^2+9^2+12^2+9^2+13^2+12^2+14^2$$

$$= 1600$$

Total Sum of Squares (TSS) = Raw Sum of Squares – CF

$$= 1600 - 1500 = 100$$

Sum of Squares due to Treatments (SST) = $\frac{T_A^2}{5}+\frac{T_B^2}{5}+\frac{T_C^2}{5}-CF$

$$= \frac{(50)^2}{5}+\frac{(40)^2}{5}+\frac{(60)^2}{5}-CF$$

$$= \frac{1}{5}[2500+1600+3600]-1500$$

$$= 1540 - 1500 = 40$$

Sum of Squares due to Error (SSE) = TSS - SST

$$= 100 - 40 = 60$$

Mean Sum of Squares due to Treatments (MSST) = $\frac{SST}{df} = 20$

Mean Sum of Squares due to Error (MSSE) = $\frac{MSSE}{df} = \frac{60}{12} = 5$

Therefore, $F = \frac{MSST}{MSSE} = \frac{20}{5} = 4$

ANOVA Table for One-way Classified Data

Sources of Variation (SV)	Degrees of Freedom	Sum of Squares	Mean Sum of Squares (MSS)	F-Statistic or Variation Ratio
Due to three varieties or due to treatments	2	40 (SST) or (SSB)	MSST = 20	$F_{2,12} = \frac{20}{5} = 4$
Due to error within groups	12	60 (SSE) or (SSW)	MSSE = 5	
Total	14	TSS = 100		

For $\nu_1 = 2, \quad \nu_2 = 12,$ the table value of F at 5% level of significance is 3.88 which can be seen from the statistical table. Since the calculated value is greater than the table value of F at 5% level of significance. So, we reject the null hypothesis and hence we conclude that the difference between the mean field of three varieties is significant.

Since the null hypothesis is rejected, then pairwise comparison test may be applied for testing the null hypothesis of equality of two population means. For this, critical difference (CD) will be calculated by using the formula

$$CD = \sqrt{\frac{2MSSE}{5}} \times t_{0.05} \text{ at error df}$$

$$CD = \sqrt{\frac{2\times 5}{5}} \times 2.571$$

$$= \sqrt{2} \times 2.571 = 1.41 \times 2.571$$

$$= 3.625$$

$$\left|\overline{T_1} - \overline{T_2}\right| = |10 - 8| = 2$$

$$\left|\overline{T_1} - \overline{T_3}\right| = |10 - 12| = 2$$

$$\left|\overline{T_2} - \overline{T_3}\right| = |8 - 12| = 4$$

Since $\left|\overline{T_1} - \overline{T_2}\right|$ and $\left|\overline{T_1} - \overline{T_3}\right|$ are less than CD. So we accept the null hypothesis which means that if we interested to take out of A and B varieties then we can take any of these two. Similarly, between A and C we can take any of varieties. But if we conclude to take out of B and C then we should prefer C because hypothesis of equality two mean is rejected and the mean value corresponding to C varieties is higher than B varieties.

E2) Null hypotheses H_0: $\mu_i = \mu_2 = \mu_3$, i.e. there is no difference in the production of these varieties P, Q and R against the alternative hypothesis $H_1 : \mu_i \neq \mu_2 \neq \mu_3$

$G = \sum y_{ij}$ = Grand total

= 14+16+18+14+13+15+22+18+16+19+19+20 = 204

N = Total number of observations = 12

Correction Factor (CF) = $\frac{G^2}{N} = \frac{204\times 204}{12} = 3468$

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^{\ 2}$

$$= 14^2+16^2+18^2+14^2+13^2+15^2+22^2$$
$$+18^2+16^2+19^2+19^2+20^2$$

= 3552

Total Sum of Squares (TSS) = RSS - CF

= 3552 - 3468 = 84

Sum of Squares due to Treatments (SST)

$$= \frac{T_P^2}{3} + \frac{T_Q^2}{4} + \frac{T_R^2}{5} - CF$$

$$= \frac{48\times 48}{3} + \frac{64\times 64}{4} + \frac{92\times 92}{5} - 3468$$

= 768+1024+1692.80 - 3468 = 16.8

Sum of Squares due to Error (SSE) = TSS - SST

= 84 - 16.8 = 67.2

Mean Sum of Squares due to Treatments (MSST) = $\frac{SST}{df} = \frac{16.8}{2} = 8.4$

Mean Sum of Squares due to Error (MSSE) = $\frac{SSE}{df} = \frac{67.2}{9} = 7.467$

Therefore,

$$F_{2,9} = \frac{MSST}{MSSE} = \frac{8.4}{7.467} = 1.125$$

ANOVA Table for One-way Classified Data

Sources of Variation (SV)	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-statistic or Variation Ratio
Between Varieties	2	16.8	8.4	$F_{2,9} = \frac{8.4}{7.46} = 1.12$
Due to Error	12	67.20	7.467	
Total	14	84		

For $\nu_1 = 2$ and $\nu_2 = 9$ the tabulated value of F at 5% level significance is 4.261. Since the calculated value F is less than the table value of F, we accept the null hypotheses and conclude that there is no any significance difference in their mean productivity of three varieties P, Q and R.

E3) We have

ν_1	ν_2	ν_3	ν_4
2000	2160	2210	2370
2200	2230	2270	2040
2100	2050	2300	2270
2300	2040	2420	2180
2250	2190	2020	
2340	2360		
2150	2280		
2240			
2460			

To simplify the calculations, we subtract some suitable number 2200 (say) from all observations and then dividing by 10 we have

S.No.	v_1	v_2	v_3	v_4	$(v_1)^2$	$(v_2)^2$	$(v_3)^2$	$(v_4)^2$
1	-20	-4	1	17	400	16	1	289
2	0	3	7	-16	00	9	49	256
3	-10	-15	10	7	100	225	100	49
4	10	-16	22	-2	100	256	484	4
5	5	-1	-18		25	1	324	
6	14	16			196	256		
7	-5	8			25	64		
8	4				16			
9	26				676			
Total	24	-9	22	6	1538	827	958	598

Null Hypothesis H_0: There is no significant difference in the effect of varieties.

Against the alternative hypothesis H: There is significant difference in the effect of varieties.

$$G = 24 + (-9) + 22 + 6 = 43$$

$$CF = \frac{G^2}{N} = \frac{(43)^2}{25} = 73.96$$

$$RSS = 1538 + 827 + 958 + 598 = 3921$$

$$TSS = RSS - CF = 3921 - 73.96 = 3847.04$$

$$SST = \frac{(24)^2}{9} + \frac{(-9)^2}{7} + \frac{(22)^2}{5} + \frac{(6)^2}{4} - 73.96$$

$$= 64 + 11.5714 + 96.8 + 9 - 73.96$$

$$= 107.4114$$

$$SSE = TSS - SST$$

$$= 3847.04 - 107.4114$$

$$= 3739.6286$$

$$MSST = \frac{SST}{df} = \frac{107.4114}{3} = 35.8038$$

$$MSSE = \frac{SSE}{df} = \frac{3739.6284}{21} = 178.0776$$

$$F = \frac{MSST}{MSSE} = \frac{35.8038}{178.0776} = 0.2011$$

ANOVA Table

Source of variation	SS	df	MSS	F
Between Varieties	107.4114	3	35.8038	F =0.2011
Due to errors	3739.6286	21	178.0776	
Total	3847.04	24		

Calculated F = 0.2011

Tabulated value of F at 5% level of significance with (3, 21) degrees of freedom is 3.07

Conclusion: Since calculated $|F|$ < Tabulated F, so we may accept H_0 and conclude that varieties ν_1, ν_2, ν_3, ν_4 of wheat are homogeneous.