UNIT 5 INTRODUCTION TO ANALYSIS OF VARIANCE

Structure

5.1 INTRODUCTION

As its name suggests, the analysis of variance focuses on variability. It involves the calculation of several measures of variability, all of which comes down to one or another version of the measure of variability such as the sum of squared deviations or mean sum of squared deviations. The statistical technique known as “Analysis of Variance”, commonly referred to by the acronym ANOVA was developed by Professor R. A. Fisher in 1920’s. Variation is inherent in nature, so analysis of variance means examining the variation present in data or parts of data. In other words, analysis of variance means to find out the cause of variation in the data. The total variation in any set of numerical data of an experiment is due to number of causes which may be (i) Assignable causes; and (ii) Unassignable / Chance causes.

The variation in the data due to assignable causes can be detected, measured and controlled whereas the variation due to chance causes is not in the control of human being and cannot be traced or find out separately.

The reason, this analysis is called analysis of variance rather than multi-group mean analysis (or something like that), is because it compares group means by analysing comparisons of variance estimates. Analysis of variance facilitates the analysis and interpretation of data from field trials and laboratory experiments in agriculture and biological research. Today, it constitutes one of the principal research tools of the biological scientists, and its use is spreading rapidly in the social sciences, the physical sciences, in the engineering, in management, etc.

In Unit 11 of MST-004, we compared means from two independent groups by using t-test. But if we are interested to test more than two independent groups then t-test cannot be applied and firstly we have to apply analysis of variance technique. An F-test is used to test the means of several groups. This F-test was named 'F' in honor of Professor R. A. Fisher by G. W. Snedecor. ANOVA is helpful because it possesses an advantage over a two sample t-test. The multiple two sample t-test would result in an increase of chance of committing a type I error.

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two population means. In an situation when we have more than two population to consider at a time and want to test the means of these population are same. For example, five doses of a drug are applied to four patients each and responses / values of dependent variable (observations) of these twenty patients are obtained. Now, we may be interested in finding out whether the effect of these five doses of drug on the patients is significantly differs. The answer to this problem is provided by the technique of analysis of variance. Thus, the analysis of variance technique is used to test the homogeneity of several population means.

Thus, the analysis of variance technique is a powerful statistical tool for tests of significance in comparing more than two means.

The notations and basic definitions of the important terms are provided in Section 5.2, so that you would become familiar with the basic terminologies, notations and the various types of analysis of variance technique. Section 5.3 describes the various assumptions involved in analysis of variance whereas Section 5.4 explains the various types of linear models used in analysis of variance. Applications of analysis of variance are explored in Section 5.5.

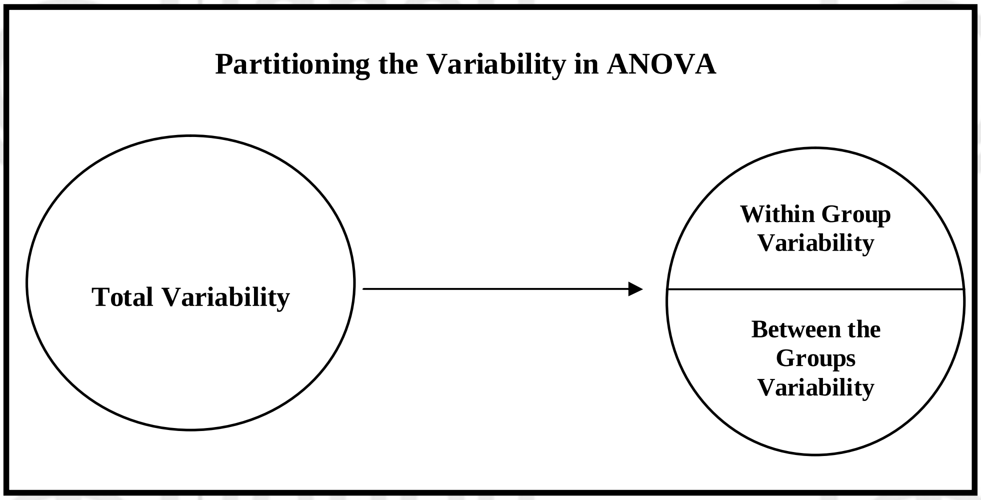
Objectives

After reading this unit, you would be able to

- become familiar with the analysis of variance technique;
- describe the various types of analysis of variance;
- describe the various assumptions involved in analysis of variance;
- define the various types of linear models used in analysis of variance; and
- describe the applications of analysis of variance.

5.2 ANALYSIS OF VARIANCE

According to Professor R. A. Fisher, Analysis of Variance (ANOVA) is "Separation of variance ascribable to one group of causes from the variance ascribable to other group". So, by this technique, the total variation present in the data are divided into two components of variation one is due to assignable causes (between the groups variability) or other is variation due to chance causes (within group variability).


Partitioning the Variability in ANOVA
Total Variability
Within Group Variability
Between the Groups Variability

The analysis of variance technique solves the problems of estimating and testing to determine, whether to infer the existence of true difference among "treatment" means, among variety means and under certain conditions among other means with respect to the problem of estimation. The analysis of variance is simply the form of the method of least squares discussed in Unit 5 of MST-002. Analysis of variance technique can be classified as follows:

1. Parametric ANOVA.
2. Non Parametric ANOVA.

Parametric ANOVA can be classified as simply ANOVA if only one response variable is considered. If more than one response variables are under consideration than it is called multivariate analysis of variance (MANOVA).

If we consider, only one independent variable which affects the response / dependent variable then it is called One-way ANOVA. If the independent variables / explanatory variables are more than one i.e. n (say) then it is called n-way ANOVA. If n is equal to two than the ANOVA is called Two-way classified ANOVA.

Factorial ANOVA is used when the experimenter wants to study the interaction effects among the explanatory variables. Repeated measure ANOVA is used when the same subjects (experimental units) are used for each treatment (levels of explanatory variable). Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

The F test in ANOVA has been used with normality assumption. But there are some cases where the distribution of response variable and its transformation is not normal. In such cases, where the variable Y_{ij} is non-normal and the appropriate information which will make normal is unknown, we need non-parametric ANOVA methods. You can set Kruskal-Wallis One-way ANOVA by ranks. One-way repeated measures ANOVA for non-parametric is Friedman test among the explanatory variables. This technique can be used in different statistical linear models, like in Fixed effect model, Random effect model and Mixed effect model.

5.3 Basic Definitions of the Terms Used in Analysis of Variance

Once you have familiarized yourself with the terminology of analysis of variance, you will find it easier to grasp many of the parametric techniques that you read about in this unit. Some of the terms described below may be referred to by one of many names, as indicated below:

1. **Variable**
   A characteristic or attribute varies in a measurable way between subjects in a sample.

2. **Response/Dependent Variable Y**
   It describes the measurements, usually on a continuous scale, of the variable of interest (e.g. weight: what causes variation in weight?).

3. **Explanatory /Independent /Predictor Variable/Treatment/ Factor Effect X**
   Explanatory variable is controlled by experimenter, which may affect response. Response does not affect explanatory variable. The non-random measurements or observations (e.g. treatments fixed by experimental design), which are hypothesized in a statistical model to have predictive power over the response variable. This hypothesis is tested by calculating sums of squares and looking for a variation in Y between levels of X that exceeds the variation within levels. An explanatory variable can be categorical (e.g. sex, with 2 levels of male and female), or continuous (e.g. height with a continuum of possibilities). The explanatory variable is assumed to be 'independent' in the sense of being independent of the response variable: i.e. weight can vary with height, but height is independent of weight. The values of X are assumed to be measured precisely, without error, permitting an accurate estimate of their influence on Y. Explanatory variable may be fixed, random or mixed as per statistical model used.

4. **Variates/ Replicates/Observations/Scores/Data Points**
   The replicate observations of the response variable (Y_1, Y_2, …, Y_i, …, Y_N) are measured at each level of the explanatory variable. These are the data points, each usually obtained from a different subject to ensure that the sample size reflects N independent replicates (i.e. it is not inflated by non-independent data 'pseudo replication').

5. **Sample**
   The collection of observations measured at a level of X (e.g. body weights were measured from one sample of males and another of females to test the effect of Sex on Weight). If X is continuous, the sample comprises all the measures of Y on X (e.g. Weight on Height).

6. **Sum of Squares**
   The squared distance between each data point (Y_i) and the sample mean ($\bar{y}$), summed for all N data points. The squared deviations measure variation in a form which can be partitioned into different components that sum to give the total variation (e.g. the component of variation between samples and the component of variation within samples).

7. **Variance and Standard Deviation**
   The variance in a normally distributed population is described by the average of N squared deviations from the mean. Variance usually refers to a sample, however, in which case it is calculated as the sum of squares divided by (N−1) rather than N. Its positive root is then the standard deviation (SD) which describes the dispersion of normally distributed variables (e.g. 95% lying within 1.96 standard deviations of the mean when N is large).

8. **Statistical Model Y = f (X) + e, where f (X) is Some Function of X**
   It is a statement of the hypothesized relationship between the response variable and the predictor variable. A simple model would be: Weight = Sex + e. The '=' does not signify a literal equality, but a statistical dependency. So the statistical analysis is going to test the hypothesis that variation in the response variable on the left of the equals sign (Weight) is explained or predicted by the factor on the right (Sex), in addition to a component of random variation (the error term e). An analysis of variance will test whether significantly more of the variation in ‘Weight’ falls between the categories of 'male' and 'female', and so is explained by the independent variable ‘Sex’ that lies within each category (the random variation ‘e’). The error term is often dropped from the model description though it is always present in the model structure, as the random variation against which to calibrate the variation between levels of X in testing for a significant explanation (the F-ratio).

9. **Null Hypothesis H_0**
   While a statistical model can propose a hypothesis, that response variable Y depends on independent variable X, the statistical analysis can only seek to reject a null hypothesis that Y does not vary with X. This is because it is always easier to find out how different things are than to know how much they are the same, so the statistician's easiest objective is to establish the probability of a deviation away from random expectation rather than towards any particular alternative. If the analysis reveals a sufficiently small probability that the null hypothesis is true, then we can reject it and state that Y evidently depends on X in some way or hypothesis of null hypothesis.

10. **One-Way ANOVA Y = f (X) + e**
    An analysis of variance (ANOVA) is used to test the model hypothesis that variation in the response variable Y can be partitioned into the different levels of a single explanatory variable X (e.g. Weight = Sex). If X is a continuous variable, then the analysis is equivalent to a linear regression, which tests for a significant slope in the best fit line describing change of Y with X (e.g. Weight with Height).

11. **Two-Way ANOVA $Y = X_1 + X_2 + X_1 X_2 + e$**
    Test of the hypothesis that variation in Y can be explained by one or both variables X_1 and X_2. If X_1 and X_2 are categorical and Y has been measured only once in each combination of levels of X_1 and X_2, then the interaction effect $X_1.X_2$ cannot be estimated. Otherwise a significant interaction term means that the effect of X_1 is modulated by X_2 (e.g. the effect of Sex (X_1), on Weight (Y) depends on Nationality (X_2). If one of the explanatory variables is continuous, then the analysis is equivalent to a linear regression with one line for each level of the categorical

variable (e.g. graph of -Weight by Height, with one line for males and one for females): different intercepts signify a significant effect of the categorical variable, different slopes signify a significant interaction effect with the continuous variable.

12. **Error/Residual**
    The amount by which an observed value of variable differs from the value predicted by the model is known as error. Errors or residuals are the segments of scores not accounted for by the analysis. In analysis of variance, the errors are assumed to be independent of each other, and normally distributed about the sample means. They are also assumed to be identically distributed for each sample (since the analysis is seeking only a significant difference between sample means), which is known as the assumption of homogeneity of variances.

13. **Normal Distribution**
    It is a bell-shaped frequency distribution of a continuous variable. The formula for the normal distribution contains two parameters: the mean, giving its location, and the standard deviation, giving the shape of the symmetrical 'bell'. This distribution arises commonly in nature when myriad independent forces, themselves subject to variation, combine additively to produce a central tendency. The technique of analysis of variance is constructed on the assumption that the component of random variation takes a normal distribution. This is because the various sum of squares that are used to describe variance in an ANOVA accurately reflect the true variation between and within samples only if the residuals are normally distributed about sample.

14. **Degrees of Freedom (df)**
    The F-ratio in an analysis of variance is always presented with two sets of degrees of freedom, the first corresponding to one less than a level (p) of the explanatory variable i.e. $(p-1)$ and the second to the remaining error degrees of freedom i.e. $(n-p)$ in one-way classified data.

15. **Variance Ratio / F-statistic**
    The statistic calculated by analysis of variance, which reveals the significance of the hypothesis that response variable depend on explanatory variable. It comprises the ratio of two mean squares i.e. Mean sum of squares due to explanatory variable divided by Mean sum of squares due to error. A large proportion indicates a signified effect of independent variable.

16. **Significance**
    This is the probability of mistakenly rejecting a null hypothesis that is actually true.

17. **P -Value**
    In biological sciences a critical value $p = 0.05$ is generally taken as marking an acceptable boundary of significance. A large F-Ratio signifies a small probability that the null hypothesis is true.

18. **Mean Sum of Squares**
    Mean sum of squares is the average sum of squares. In other words, “the sum of squares of deviations from the ‘mean X’ or ‘error e’ divided by its appropriate degrees of freedom”.

19. **Population**
All subjects / experimental units possess a common characteristic that is being studied whereas a group of subjects selected from the target population is known as sample.

20. **Parameter and Statistic**
Characteristics /measures obtained from a population are known as parameters whether the characteristics/measures obtained a sample are known as statistic.

E1) Describe the analysis of variance and differentiate between One-way and Two-way ANOVA.

5.4 BASIC ASSUMPTIONS IN ANALYSIS OF VARIANCE

The analysis of variance has been studied from several approaches, the most common of which is to use a linear model that relates the response variable to the treatments (explanatory variables). Even when the statistical model is non linear, it can be approximated by a linear model for which an analysis of variance may be appropriate.

When analysis of variance is used as a method of statistical inference for inferring properties of the "Population" from which the data are drawn (taken) then certain assumptions about the "Population" and the sampling procedure by means of which the data are obtained, must be fulfilled, if the inference are to be valid. So, ANOVA makes certain assumption about the nature of the experimental data that have to be at least approximately true before the method can be validly applied.

Before explaining the assumptions of ANOVA the response variable Y demonstrated in the following two-way table in which there are r-rows (varieties) and c-columns (treatments).

	1	2	...	j	...	c	Total	Row Means
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1c}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2c}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.			.	.	.	.
.	.	.			.	.	.	.
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ic}	$y_{i.}$	$\bar{y}_{i.}$
.	.	.			.	.	.	.
.	.	.			.	.	.	.
r	y_{r1}	y_{r2}	...	y_{rj}	...	y_{rc}	$y_{r.}$	$\bar{y}_{r.}$
Total	$y_{\cdot 1}$	$y_{\cdot 2}$	...	$y_{\cdot j}$	...	$y_{\cdot c}$	$y_{..}$	
Column Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.c}$		$\bar{y}_{..}$

y_{ij} = is the value of the response variable Y occurring in the ith row and jth column.

$y_{i.} = \sum_{j=1}^{c} y_{ij}$ for all j =1, 2, …, c; = Total corresponding to the ith row

$\bar{y}_{i.} = \frac{1}{c}\sum_{j=1}^{c} y_{ij}$ for all j = 1, 2,…, c; = Mean corresponding to the ith row

$y_{.j} = \sum_{i=1}^{r} y_{ij}$ for all i =1, 2, …, r; = Total corresponding to the jth column

$\bar{y}_{.j} = \frac{1}{r}\sum_{i=1}^{r} y_{ij}$ for all i = 1, 2,…, r; = Mean corresponding to the jth column

$y_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}$ for all i=1, 2, …, r; j=1, 2, …, c; = Total of all the observations

$\bar{y}_{..} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}$ for all i=1, 2, …, r; j=1,2, …, c; =Mean of all the observations

Following are the assumptions satisfied by the ANOVA technique:

5.4.1 Assumption of Randomness

The values y_{ij} are (observed values/response variable/dependent variable) random variables that distributed about so called true mean values (expected value) μ_{ij} (i=1, 2, ..., r; j=1,2, ..., c) are fixed constants.

In statistical language this assumption states that, of some particular type of experiment leading to value y_{ij} were repeated indefinitely, then the value y_{ij} would vary at random about an average value equal to μ_{ij}, which is therefore a parameter that characterizes the expected value of the y_{ij}.

From this assumption an unbiased estimator of any linear function of the μ_{ij} with known coefficient is obtained by the same linear function of the y_{ij}.

Further, if the variances of the y_{ij} about their respective means and their inter correlations are known, then the variances of any linear function of the y_{ij} can be evaluated and provides a measure of the precision of this linear function of the y_{ij} as an unbiased estimator of corresponding linear function of the μ_{ij}.

5.4.2 Assumption of Additivity

True means (μ_{ij}) are simple additive functions of the corresponding marginal means and the general mean, that is,

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

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

or $y_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) + e_{ij}$ for (i=1, 2, ..., r; j=1, 2, ..., c)

then the statistical inferences that may be based upon the y_{ij} are of a much more satisfactory. Or we can say that the observed value/response variable y_{ij} can be the sum of three parts:

1. The overall mean of the observations.

2. Treatments/factors/classification effects.
3. A random element or error effect drawn from normally distributed populations. The random element reflects the combined effect of natural variation between observations and errors of measurement.

When the above additivity assumption is satisfied, the difference between arbitrary pair of row-wise marginal means, e.g. $\mu_{1.}$ and $\mu_{2.}$ (say) is a comprehensive measure of the average difference in effectiveness of the factors identified with these rows. When the above assumption does not satisfied then $\mu_{1.} - \mu_{2.}$ is merely a measure of the average difference between the effects of the corresponding row factors when the column factors are as in the experiment concerned. Similarly, the actual mean difference in effectiveness of a pair of column factors will depend upon row factors concerned. Hence, when additivity does not prevail, we say that there are interactions between row factor and column factor. Thus, additivity implies that the true mean yield of level of one factor is greater (or less) than the true mean of another level of a factor by an amount additive constant not a multiplier that is the same for each of the another factor concerned, and conversely the true mean yield with levels of another factor is greater (or less) than the true mean yield with levels of another factor by an amount that does not depend upon the previous factor concerned, which is exactly that there are no "Interaction" between first factor & second factor.

When assumptions 1 and 2 are satisfied that the difference between any pair of row wise means of the observations y_{ij} e.g., $\overline{y}_{1.} - \overline{y}_{3.}$ is an unbiased estimator of the general average difference in effectiveness of the row factors concerned i.e. of $\mu_{1.} - \mu_{3.}$. Similarly, for column wise means of the observations.

5.4.3 Equality of Variances or Homoscedasticity and Zero Correlation

The random variables y_{ij} are homoscedastic and mutually uncorrelated, that is, they have a common variance σ^2 although means vary from group to group. The variances should be constant in all groups and all covariance among them are zero so, it is desirable before using ANOVA technique the assumption of homogeneity of variance and normality should jointly be tested.

In general, it is not possible to derive from the observations y_{ij}, the unbiased estimates of variances of the y_{ij} and any particular linear functions of them, unless the assumptions 1, 2 and 3 are not satisfied.

If assumptions 1, 2 and 3 are all satisfied, an unbiased estimate of the variances of the differences of two observed row means and an unbiased estimate of the variances of the differences of two observed column means can be evaluated.

Suppose, assumption 1 and assumption 2 are satisfied and assumption 3 is not. In this case, four values of σ^2 would be the expected value of sum of square due to rows, due to column, due to residual and due to total, which may be denoted by σ_r^2, σ_c^2, σ_e^2 and σ_t^2. These will be the complex weighted means of the variances and covariances of the $y_{ij.}$

So, it is desirable before using ANOVA technique the assumption of homogeneity of variance should be tested. Although the means may vary from group to group the variances should be constant in all the groups. The following procedures are widely used for this purpose:

Bartlett's Test for Homogeneity of Variances

The null hypothesis for equal variances is

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

against $H_1: \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_k^2$

From each group of size n_i taken from i^{th} population, i = 1, 2, ... , k we can calculate $S_1^2, S_2^2, \ldots, S_k^2$ where,

$$S_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{i.}\right)^2 \qquad \text{for all } i = 1, 2, \ldots, k$$

Now, let $\nu_i = n_i - 1$; where, ν_i are the degrees of freedom of S_i^2

and $\nu = \sum_{i=1}^{k} \nu_i$ for all i = 1, 2, …, k

and $S^2 = \sum_{i=1}^{k} \nu_i S_i^2$

The Bartlett's test statistic M is defined by

$$M = \nu \log S^2 - \sum_{i=1}^{k} \nu_i \log S_i^2$$

When none of the ν_i's degrees of freedom is small, M is distributed as $\chi^2_{k-1.}$ The χ^2 approximation is generally acceptable if all n_i are at least 5. This is slightly biased test. So it can be improved by dividing M by the factor

$$C = 1 + \frac{1}{3(k-1)}\sum_{i=1}^{k}\left(\frac{1}{\nu_i} - \frac{1}{\nu}\right)$$

Instead of M, it is suggested to use M/C for the test statistic.

Levene’s Test for Homogeneity of Variances

Levene's test offers a more robust alternative to Bartlett's procedure. That means it will be less likely to reject a true hypothesis of equality of variances just because the distribution of the sampled populations are not normal. When non normality is suspected, Levene's procedure is a better choice then Bartlett's goodness of fit test because the value of the response variance occurring in the row and column.

Levene's test is used to test if k samples have equal variances. Equal variances across samples are called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene’s test can be used to verify that assumption.

Levene's test is an alternative to the Bartlett test. The Levene’s test is less sensitive than the Bartlett test to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal distribution, then Bartlett's test has better performance.

The Levene’s test is defined as:

$$H_0: \sigma_1 = \sigma_2 = \ldots = \sigma_k$$

$$H_1: \sigma_i \neq \sigma_j \quad \text{for at least one pair (i, j).}$$

Given a variable Y (response variable) with sample of size N divided into k subgroups, where n_i is the sample size of the i^{th} subgroup, the Levene's test statistic is defined as:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i \left(\bar{Z}_{i.} - \bar{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \bar{Z}_{i.}\right)^2}$$

where, Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = \left|Y_{ij} - \bar{Y}_{i.}\right|$

   where, $\bar{Y}_{i.}$ is the mean of the i^{th} subgroup.

2. $Z_{ij} = \left|Y_{ij} - \tilde{Y}_{i.}\right|$

   where, $\tilde{Y}_{i.}$ is the median of the i^{th} subgroup.

3. $Z_{ij} = \left|Y_{ij} - \bar{Y}'_{i.}\right|$

where, $\bar{Y}'_{i.}$ is the 10% trimmed mean of the i^{th} subgroup. $\bar{Z}_{i.}$ are the group means of the Z_{ij} and $\bar{Z}_{..}$ is the overall mean of the Z_{ij}.

The three choices for defining Z_{ij} determine the robustness and power of Levene's test. By robustness, we mean the ability of the test to not falsely detect unequal variances when the underlying data are not normally distributed and the variables are in fact equal. By power, we mean the ability of the test to detect unequal variances when the variances are in fact unequal.

Using the trimmed mean which is performed best when the underlying data followed a Cauchy distribution (i.e., heavy-tailed) and the median which is performed best when the underlying data followed a χ_4^2 (i.e., skewed) distribution. Using the mean provided the best power for symmetric or moderate-tailed distributions.

Although the optimal choice depends on the underlying distribution, the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. If you have knowledge of the underlying distribution of the data, this may indicate using one of the other choices.

The Levene's test rejects the hypothesis at level of significance α that the variances are equal if

$$W > F_{(\alpha,\, k-1,\, N-k)}$$

where, $F_{(\alpha, k-1, N-k)}$ is the upper critical value of the F distribution with (k – 1) and (N – k) degrees of freedom at a significance level of α.

5.4.4 Assumptions of Normality

y_{ij} are jointly distributed in multivariate normal distribution. It is not possible to conduct exact test of significance based on the y_{ij} alone e.g. test of significance on Snedecor's F-distribution. Fortunately, normality, in addition to assumption 1 to 3 is sufficient for exact test of significance.

So, when assumptions 1 to 4 are all satisfied, then all of the usual analysis of variance procedure for estimating and testing to determine whether to infer the existence of, fixed linear relations e.g. non-zero difference among population means are strictly valid. In particular, an unbiased estimator of any given linear function of the parameters μ_{ij} is provided by the identical linear function of the observations y_{ij}, and an unbiased estimate of its variance can be derived from the residual or error mean square and exact confidence limits for the value of the given linear function of the parameters can be deduced with the aid to Student's t-distribution.

An easy method of checking the assumption of a single normal distribution is to construct a histogram of the data. The data should be examined for departures from normality before the tests are applied. However, the tests are robust to small departures from normality i.e. they work fairly well as long as the curve of the data is bell shaped and the tails are not heavy. Another method for testing the normality assumption is the normal probability plot. Goodness of fit for χ^2-test may be applied for testing normality. This assumption can be relaxed when the sample size is very large. If the assumptions do not hold, then a transformation of the y_{ij} into another scale will often allow ANOVA to be carried out. For more detailed use of transformations you can use Snedecor and Cochran (1980).

In many applications and applied microbiology in which bacterial number are being estimated, the assumption may not hold. The case when the sample sizes are small and in whole numbers with many zeros are unlikely to be normally distributed or wide range of observations may be present leading to heterogeneous variances.

Let us answer some activity questions.

E 2) Describe the assumptions of randomness in ANOVA.

E 3) Describe the assumptions of additivity in ANOVA.

E 4) Describe the assumptions of normality in ANOVA.

5.5 LINEAR MODELS USED IN ANALYSIS OF VARIANCE

Let $y_1, y_2, \ldots, y_n$ be the n observations of dependent variable/response variable. By using the additively assumptions of ANOVA, we shall assume that observed value y_{ij} be composed of two factors

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

where, μ_i is the true value and e_{ij} is the error. The true value μ_i is that part which is due to assignable causes and the portion that remains is the error, which is due to various chance causes. The true value μ_i is again assumed to be a linear function of k unknowns $T_1, T_2, \ldots, T_k$ called effects, i.e.

$$\mu_i = a_{i1}T_1 + a_{i2}T_2 + \ldots + a_{ik}T_k$$

where, a_{ij} are known and each being usually taken to be 0 or 1. This set up, which is fundamental to analysis of variance, is called linear model. So,

$$y_{ij} = a_{i1} T_1 + a_{i2} T_2 + \ldots + a_{ik} T_k + e_{ij}$$

5.5.1 Fixed Effect Model (Model 1)

A model in which all the effects T_i's are unknown constants, which we call parameters, is known as fixed effect model or model 1 or linear hypothesis model. It is often the case that one of the T_j's is constant and $a_{ij} = 1$ for that j and i. Such a T_j is called general mean or additive constant.

The fixed effect model of ANOVA applies to situations in which the experimenter applies one or more treatments (levels of a factor) to the subject/experimental units of the experiment to see if response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as whole.

5.5.2 Random Effects Models (Model 2) or Variance Components Model

A model in which all the effects T_i's are random variable except possibly the additive constant is called random effect model or model 2 or variance component model.

In Statistics, random effect model also called variance components model, is a kind of hierarchical linear model. It assumes that the data set /observations being analysed consist a hierarchy of different population whose differences relate to that hierarchy.

Random effects models are used when the treatments (levels of a factor) are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting, the treatments differ from ANOVA (model 1).

Suppose we are interested in knowing whether all the factors level effect (class effects) are equal or not. Now due to consideration of time, cost, or space, it is not possible to include in our experiment all the available factor levels (classes) effects. We can include only a sample of these factor levels and we want to infer about all the factor levels. Whether included in the experiment or not, form the result of the classes included in the experiment. In the random model, we shall consider balanced cases. Balanced cases are those cases, in which observations under different factor levels are the same. In higher order classification, if number of observations in cell is equal then it is called balanced.

In random effect models, main interest lies in estimation of variance components while in fixed effect models interest lies in estimation and testing the treatment differences. The difference of fixed effect model and random effect model can be seen from the following example:

Suppose a certain drug is thought to have an effect on the ability to perform mental arithmetic. The quantity of drug used may vary from 0 to 100 milligrams.

One possible experiment would be to test all possible levels of this drug on groups subjects and then use of analysis of variance to detect differences.

Because limited funds, only six levels of the drug may be tested. Then there will be two situations of implementation and analysis this experiment:

1. One possibility would be to systematical choosen a set of levels covering the range of doses. For example, one might choose 0, 20, 40, 60, 80 and 100. The differences in mental arithmetic scores could then be analysed using fixed effect model because doses have been fixed.
2. Another possibility is to choose the six dose levels randomly from the set of numbers 1 to 100. If this experiment were repeated then different levels might be chosen. The ith dose levels changes and because of this the effect of ith treatment is a random variable.

For instance, when an experimenter selects two or more treatments or two or more varieties, for testing, he rarely, if ever, draws them at random from a population of possible treatment or varieties he select those that he believes are most promising. In that situation, fixed effect model (model 1) is generally appropriate. On the other hand, when an experimenter selects a sample of varieties or treatments from a group of population for a study of the effects of various treatments or varieties, he can ensure that they are a random sample from the population of varieties or treatments by introducing randomisation into the sampling procedure for example, by using random number table. In this situation a random effect model (model 2) would clearly be appropriate.

Example of Random effect model

Suppose m large elementary schools are chosen randomly from among thousands in large country. Suppose also that n pupils of the same age are chosen randomly at each selected school. Their scores in a standard aptitude test are as curtained. Let y_{ij} be the score of the jth pupil in the ith school, then the following model is suggested:

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

where, μ is the average test score for the entire population. In this model α_i is the school specific random effect. It measures the difference between the average score at school i and the average score in the entire country and it is random because the school has been randomly selected from a larger populations of schools. The term, e_{ij}, is the individual specific error. That is, it is the deviation of the jth pupil's score from the average for the ith school. Again, this is regarded as random because of the random selection of pupils within the school, even though it is fixed quantity for any given pupil.

5.5.3 Mixed Effect Model

A model in which at least one T_i's is a random variable and at least one T_i' is constant (non-nagative constant) is called a mixed model.

Remark:

In general random effect is efficient, and should be used (over fixed effect) if the assumptions underlying it are believed to be satisfied. For random effect to work in the school example, it is necessary that the school specific effects be orthogonal to the other covariates of the model. This can be tested by running random effects, then fixed effects and doing a Housman specification to be seen test. If the test rejects, then random effects is blazed and fixed effects is the correct estimation procedure.

E 5) Describe the fixed effect model in ANOVA.

E 6) Describe the random effect model in ANOVA.

5.6 USES OF ANOVA

The following are some of the uses of ANOVA:

1. **To Test the Homogeneity of Several Means (k groups) or**

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

   If H_0 is rejected then we can say that there is a significant difference between these k groups or there is a significant effect of these k independent variables.

2. **To Test the Relationship between Two Variables**

   This test provides evidence that dependent variable Y_{ij} and independent variable X_{ij} are related in their movements. If Y_{ij} do not relate with X_{ij} then we expect H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, which is the null hypothesis for testing the absence of relationship.

3. **Test for Linearity of Regression**

   After the relationship is established, the next step will be to find the appropriate regression function. At the first stage we try to find out whether the linear regression fit the observed data. So, the null hypothesis is now

   H_0: $\mu_i = \alpha + \beta X_i$

   with the sample model $Y_{ij} = \mu_i + e_{ij}$, when α and β are the parameters.

4. **Test for Polynomial Regression**

   The test procedure for testing the null hypothesis

   H_0: $\mu_i = \alpha + \beta_1 X_i + \beta_2 X_i^{\,2} + \ldots + \beta_k X_i^{\,k}$

   That the relationship between X and Y can be explained by a polynomial of degree k.

5. **Some Other Uses of ANOVA**

   - Test of homogeneity of a group of regression coefficients.
   - Test for equality of regression equations from p groups.
   - Test for multiple linear regression model.

E7) Explain briefly the uses of analysis of variance.

5.7 SUMMARY

In this unit, we have discussed:

1. The basic ideas and various types of analysis of variance;
2. Basic assumptions in analysis of variance;
3. Basic terminologies and notations in analysis of variance;
4. Various types of linear models in analysis of variance; and
5. Uses of ANOVA.

5.8 SOLUTIONS/ANSWERS

E 1) As same as Section 5.2 without Sub-sections 5.2.1 and 5.2.2.

E 2) As same as Sub-section 5.4.1.

E 3) As same as Sub-section 5.4.2.

E 4) As same as Sub-section 5.4.4.

E 5) As same as Sub-section 5.5.1.

E 6) As same as Sub-section 5.5.2.

E 7) As same as Section 5.6