

# Statistical Methods for Research and Product/Process Development

**Joshy C G**  
Fish Processing Division  
ICAR-Central Institute of Fisheries Technology, Cochin  
*Email: cgjoshy@gmail.com*

Statistics is a set of procedures for gathering, measuring, classifying, computing, describing, synthesizing, analyzing, and interpreting systematically acquired data. The data can be collected either in qualitative or quantitative in nature and can be presented in the form of descriptive statistics.

## Descriptive Statistics

Descriptive Statistics gives numerical and graphical procedures to summarize a collection of data in a clear and understandable way. Inferential statistics provides procedures to draw inferences about a population from a sample.

### Types of Descriptive Statistics

1. Graphs & Frequency Distribution

It summarize the distribution of individual observations or range of values in a given set of observations.

2. Measures of Central Tendency

It computes the indices enabling the researcher to determine the average score of a given set of data

3. Measures of Variability

It computes indices enabling the researcher to indicate how a given set of data spread out

## Measures of Central Tendency

The central tendency of a distribution is an estimate of the 'centre' of a distribution of values of a given set of distribution. The major measures of central tendencies are

1. Mean
2. Median
3. Mode
4. Harmonic mean
5. Geometric mean

**The mean** is the arithmetic average of data values. It computes by adding up the observations and divide by total number of observations. It is the most commonly used measure of central tendency and it is affected by extreme values (outliers).

**The median** is the "middle most observation" in a given set of observations. If  $n$  is odd, the median is the middle number and if  $n$  is even, the median is the average of the 2 middle numbers. Median is not affected by extreme values.

**The mode** is the most frequently observation in a given set of observations. Mode is not affected by extreme values.

**The harmonic mean** is the average of the reciprocal of the observations

**The geometric mean** is the  $n^{\text{th}}$  root of the products of the observations

Averages or measure of central tendency are representatives of a frequency distribution, but they fail to give a complete picture of the distribution. Measures of central tendency do not tell anything about the scatterness of observations within the distribution.

### Measures of Dispersion

Measures of Dispersion quantify the scatterness or variation of observations from their average or measures of central tendencies. It describes the spread, or dispersion, of scores in a distribution. The three most commonly used measures are

- a. Range
- b. Variance
- c. Standard Deviation

**Range** is the simplest measure of variability and it is the difference between the highest and the lowest observation in a given set of data. It is very unstable and unreliable indicator.

Range= H-L

**Variance** measures the variability of observations from its mean. It computes the sum of squared difference between observations and mean. Standard Deviation is the square root of variance.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

### Measures of Relative Dispersion

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal, then their variability can be compared directly by using their S.Ds. However, if their means are widely different or if they are expressed in different units of measurement, S.Ds cannot be used as such for comparing their variability. In such situations, the relative measures of dispersions can be used.

**The coefficient of variation (C.V)** is a commonly used measure of relative dispersion and it is ratio of SD to the Mean multiplied by 100.

C.V. = (S.D / Mean) x 100

**The C.V.** is a unit-free measure and it is always expressed as percentage. The C.V. will be small if the variation is small. Of the two groups, the one with less C.V. is said to be more consistent.

### Tests of Significance

Once sample data has been gathered, statistical inference allows assessing evidence in favor or some claim about the population from which the sample has been drawn. The method of inference used to support or reject claims based on sample data is known as testing of hypothesis. Statistical test is a procedure governed by certain rules, which

leads to take a decision about the hypothesis for its acceptance or rejection on the basis of the sample values. These tests have wide applications in agriculture, medicine, industry, social sciences, etc.

**Definitions:**

*Statistic: It is a function of units in the sample, like sample mean, sample variance*

*Parameter: It is a function of units in the population, like population mean, population variance*

*Statistical Hypothesis: A definite/tentative statement about the population parameters*

*Simple Hypothesis: If all the parameters are completely specified, the hypothesis is called a simple hypothesis*

*Composite hypothesis: If all the parameters are not completely specified by a hypothesis is called as composite hypothesis*

*Null Hypothesis (H<sub>0</sub>): The hypothesis under test for a sample study*

*Alternative Hypothesis (H<sub>1</sub>): The hypothesis tested against the null hypothesis*

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0 \quad (\text{Two-Tailed Test})$$

$$\mu < \mu_0 (\text{Left-Tailed Test})$$

$$\mu > \mu_0 \quad (\text{Right-Tailed Test})$$

*Level of Significance ( $\alpha$ ): The maximum size of the error (probability of rejecting H<sub>0</sub> when it is true) which we are prepared to risk. The higher the value of  $\alpha$ , less precise is the result*

*Test Statistic: It is a quantity calculated from sample of data. Its value is used to decide whether or not the null hypothesis should be rejected in the hypothesis test*

*Critical value(s): The critical value(s) for a hypothesis test is a value to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected. The critical value for any hypothesis test depends on the significance level at which the test is carried out, and whether the test is one-sided or two-sided.*

**Procedure of Testing Hypothesis**

- Step 1:        Setting up the hypothesis and level of significance  
                 Null hypothesis (H<sub>0</sub>) and Alternative hypothesis (H<sub>1</sub>)  
                 Level of significance formulation ( $\alpha$ )
- Step 2:        Data Collection and selection of appropriate test procedure  
                 Compute the Test Statistic
- Step 3:        Test Criteria
  - i) reject the null hypothesis, or
  - ii) not reject the null hypothesis
- Step 4:        Draw the Inference

The major statistic's used for tests of significance are

1. Normal Test
2. t - Test
3. Chi - Square Test
4. F - Test

**Normal test**

Test for the Mean of a Normal Population

When Population Variance is known

If  $x_i (i = 1, \dots, n)$  is a r.s of size  $n$  from  $N(\mu, \sigma^2)$ , then

$H_0 : \mu = \mu_0$  or

$H_0$  : the sample has been drawn from the population with mean  $\mu_0$

$H_1 : \mu \neq \mu_0$  (two-tailed) or  $\mu > \mu_0$  (right-tailed) or  $\mu < \mu_0$  (left-tailed)

$\text{Test Statistic } Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$	with n-1 degree of freedom
---	----------------------------

Depending on the alternative hypothesis selected, the test criteria is as follows:

H <sub>1</sub>	Test	Reject H <sub>0</sub> at level of significance $\alpha$ if
$\mu \neq \mu_0$	Two-tailed test	$ Z  > Z_{\alpha/2}$
$\mu < \mu_0$	Left-tailed test	$Z < -Z_{\alpha}$
$\mu > \mu_0$	Right-tailed test	$Z > Z_{\alpha}$

$Z_{\alpha}$  is the table value of Z at level of significance  $\alpha$ .

**Test for Difference of Means**

**Normal PopulationI: Sample size n<sub>1</sub>**

**Normal PopulationII: Sample size n<sub>2</sub>**

$H_0 : \mu_1 = \mu_2$

**Test Statistic: Normal test**

$$\text{Test statistic } Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\text{Under } H_0 \quad Z = \frac{x_1 - x_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{x_1 - x_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{if } \sigma_1^2 = \sigma_2^2 = \sigma^2$$

Population Variances are unknown but equal  $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

### t - tests

#### 1. Test for the Mean of a Normal Population when Small Sample ( $n < 30$ ) and Population Variance is Unknown

A r.s  $x_1, \dots, x_n \rightarrow N(\mu, \sigma^2)$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0 \quad \text{or} \quad \mu > \mu_0 \quad \text{or} \quad \mu < \mu_0$$

$$\boxed{t\text{-test } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}}$$

Test Statistic: with  $n-1$  degree of freedom,

$$\text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The null hypothesis is accepted or rejected accordingly.

#### 2. Test for the Difference of Two Population Means: when the population variances are unknown but assumed to be equal

Let  $\bar{x}_1$  be the sample mean of a sample of size  $n_1$  of first population with mean  $\mu_1$  and  $\bar{x}_2$  be the sample mean of a sample of size  $n_2$  from a population with mean  $\mu_2$ .

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad \text{or} \quad \bar{x}_1 - \bar{x}_2 = 0$$

$$\text{Test Statistic: Under } H_0 \text{ is } t = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

$$S^2 \text{ is estimated from the sample } s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$$

#### 3. Paired t-test for Difference of Means

Paired t-test is used when

- Sample size  $n_1 =$  sample size  $n_2 = n$
- Two samples are not independent (paired)
- Let  $(x_i, y_i), i=1, \dots, n$  be a random sample from a bivariate normal population
- Let  $d_i = x_i - y_i$

The null hypothesis is  $H_0 : \mu_1 - \mu_2 = 0$  i.e  $\bar{d} = 0$

The Test Statistic under  $H_0$  is  $t = \frac{\bar{d} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$

Where  $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$  and  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$

## Chi - Square Test

### 1. Test for the variance of a normal population

Let  $x_1, x_2, \dots, x_n$  ( $n \geq 2$ ) be a r.s from  $N(\mu, \sigma^2)$ .

$$H_0 : \sigma^2 = \sigma_0^2$$

Test Statistic under  $H_0$  is  $\chi^2 = \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma_0} \right)^2 \sim \chi_n^2$  when  $\mu$  is known and

$$\chi^2 = \sum_{i=1}^n \left( \frac{x_i - \bar{X}}{\sigma_0} \right)^2 \sim \chi_{n-1}^2 \text{ when } \mu \text{ is unknown}$$

### 2. Test of Goodness of Fit

To test the discrepancy between the observed and the expected frequency

$H_0$  : the fitted distribution is a good fit

$H_1$  : not a good fit

Test Statistic:

$O_i \rightarrow$  Observed frequency of  $i^{\text{th}}$  class

$E_i \rightarrow$  Expected frequency of  $i^{\text{th}}$  class,  $i = 1, \dots, n$ .

The test statistic is  $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim \chi_{n-r-1}^2$

### 3. Test of Independence

$H_0$ : The attributes are independent

$H_1$ : They are not independent

Test Statistic:  $\chi^2 = \sum_{j=1}^c \sum_{i=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(r-1)(c-1)}^2$

$O_{ij} \rightarrow$  Observed frequency,  $E_{ij} \rightarrow$  Expected frequency  $i = 1, \dots, r; j = 1, \dots, s$

$H_0$  is rejected at level  $\alpha$  if  $\chi^2 > \chi_{(r-1)(c-1)}^2$

## F- tests

### 1. To Test for the comparison of two population variances

Consider a sample of size  $n_1$  from a normal population  $N(\mu_1, \sigma_1^2)$  and another sample of size  $n_2$  from second normal population  $N(\mu_2, \sigma_2^2)$ . **The null hypothesis to test the significance of two population variances is  $H_0: \sigma_1^2 = \sigma_2^2$  ie  $s_1^2 = s_2^2$**

The test statistic is  $F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$

The computed value of F is compared with the tabulated value and the inference is drawn accordingly.

### 2. Equality of Several Population Means

The null hypothesis to test the equality of several means is  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

The total variability in the data is being partitioned into different known variability components using a statistical technique called analysis of variance (ANOVA). But the statistic used to test the significance of equality several means is F-test/F-statistic

The statistic is  $F = \frac{\text{Variation among the sample means}}{\text{Variation within the samples}}$

## Fundamentals of Design of Experiments

### Introduction

Any scientific investigation involves formulation of certain assertions (or hypotheses) whose validity is examined through the data generated from an experiment conducted for the purpose. Thus experimentation becomes an indispensable part of every scientific endeavour and designing an experiment is an integrated component of every research programme. Three basic techniques fundamental to designing an experiment are *replication*, *local control (blocking)*, and *randomization*. Whereas the first two help to increase precision in the experiment, the last one is used to decrease bias. These techniques are discussed briefly below.

*Replication* is the repetition of the treatments under investigation to different experimental units. Replication is essential for obtaining a valid estimate of the experimental error and to some extent increasing the precision of estimating the pairwise differences among the treatment effects. It is different from *repeated measurements*. Suppose that the four animals are each assigned to a feed and a measurement is taken on each animal. The result is four independent observations on the feed. This is *replication*. On the other hand, if one animal is assigned to a feed and then measurements are taken four times on that animal, the measurements are not independent. We call them *repeated measurements*. The variation recorded in repeated measurements taken at the same time reflects the variation in the measurement process, while variation recorded in repeated measurements taken over a time interval reflects the variation in the single animal's responses to the feed over time. Neither reflects the variation in independent animal's responses to feed. We need to know about the latter variation in order to generalize any conclusion about the feed so that it is relevant to all similar animals.

For inferences to be broad in scope, it is essential that the experimental conditions should be rather varied and should be representative of those to which the conclusions of the experiment are to be applied. However, an unfortunate consequence of increasing the scope of the experiment is an increase in the variability of response. Local control is a technique that can often be used to help deal with this problem.

Blocking is the simplest technique to take care of the variability in response because of the variability in the experimental material. To block an experiment is to divide, or partition, the observations into groups called blocks in such a way that the observations in each block are collected under relatively similar experimental conditions. If blocking is done well, the comparisons of two or more treatments are made more precisely than similar comparisons from an unblocked design.

The purpose of randomization is to prevent systematic and personal biases from being introduced into the experiment by the experimenter. A random assignment of subjects or experimental material to treatments prior to the start of the experiment ensures that observations that are favoured or adversely affected by unknown sources of variation are observations "selected in the luck of the draw" and not systematically selected. Lack of a random assignment of experimental material or subjects leaves the experimental procedure open to experimenter bias.

### Contrasts and Analysis of Variance

The main technique adopted for the analysis and interpretation of the data collected from an experiment is the analysis of variance technique that essentially consists of partitioning the total variation in an experiment into components explainable to different sources of variation due to the controlled factors and error. Analysis of variance clearly indicates a difference among the treatment means. The objective of an experiment is often much more specific than merely determining whether or not all of the treatments give rise to similar responses.

#### Contrasts

Let  $y_1, y_2, \dots, y_n$  denote  $n$  observations or any other quantities. The linear function

$C = \sum_{i=1}^n l_i y_i$ , where  $l_i$ 's are given number such that  $\sum_{i=1}^n l_i = 0$ , is called a *contrast* of  $y_i$ 's.

Let  $y_1, y_2, \dots, y_n$  be independent random variables with a common mean  $\mu$  and variance  $\sigma^2$ . The expected value of the random variable  $C$  is zero and its variance is  $\sigma^2 \sum_{i=1}^n l_i^2$ . In

what follows we shall not distinguish between a contrast and its corresponding random variable.

**Sum of squares (s.s.) of contrasts.** The sum of squares due to the contrast  $C$  is defined

as  $C^2 / \sigma^{-2} \text{Var}(C) = C^2 / \left( \sum_{i=1}^n l_i^2 \right)$ . Here  $\sigma^2$  is unknown and is replaced by its unbiased

estimate, *i.e. mean square error*. It is known that this square has a  $\sigma^2 \chi^2$  distribution with one degree of freedom when the  $y_i$ 's are normally distributed. Thus the sum of squares due to two or more contrasts has also a  $\sigma^2 \chi^2$  distribution if the contrasts are

independent. Multiplication of any contrast by a constant does not change the contrast. The sum of squares due to a contrast as defined above is not evidently changed by such multiplication.

**Orthogonal contrasts.** Two contrasts,  $C_1 = \sum_{i=1}^n l_i y_i$  and  $C_2 = \sum_{i=1}^n m_i y_i$  are said to be orthogonal if and only if  $\sum_{i=1}^n l_i m_i = 0$ . This condition ensures that the covariance between  $C_1$  and  $C_2$  is zero.

When there are more than two contrasts, they are said to be mutually orthogonal if they are orthogonal pair wise. For example, with four observations  $y_1, y_2, y_3, y_4$ , we may write the following three mutually orthogonal contrasts:

$$(i) \quad y_1 + y_2 - y_3 - y_4$$

$$(ii) \quad y_1 - y_2 - y_3 + y_4$$

$$(iii) \quad y_1 - y_2 + y_3 - y_4$$

The sum of squares due to a set of mutually orthogonal contrasts has a  $\sigma^2 \chi^2$  distribution with as many degrees of freedom as the number of contrasts in the set.

### Response Surface Methodology for Product/Process Optimization

Response surface methodology (RSM) is such a collection of statistical and mathematical techniques useful for developing, improving and optimizing processes. It also has important applications in the design, development and formulation of new products, as well as in the improvement of existing product designs. The most extensive applications of RSM are in the industrial world, particularly in situations where several input variables potentially influence some performance measure or quality characteristic of the product or process.

In general, suppose that the scientist or engineer is concerned with a product, process or system involving a response  $Y$  that depends on controllable input factors  $x_1, x_2, \dots, x_p$ . The relationship between  $Y$  and the  $x$ 's is defined as

$$Y = f(x_1, x_2, \dots, x_p) + \varepsilon$$

Where the form of the true response function  $f$  is unknown and perhaps very complicated, and  $\varepsilon$  is a term that represents other sources of variability not explained or accounted by  $f$ . Thus  $\varepsilon$  includes effects such as measurement error on the response, other sources of variation that are inherent in the process or system, the effect of other variables, and so on. We treat  $\varepsilon$  as a statistical error term with mean zero and constant variance i.e.  $\varepsilon \sim N(0, \sigma^2)$ , then

$$\begin{aligned} E(Y) &= E[f(x_1, x_2, \dots, x_p)] + E(\varepsilon) \\ &= f(x_1, x_2, \dots, x_p) \end{aligned}$$

Because the form of the true response function  $f$  is unknown, we must approximate it. In fact, successful use of RSM is critically dependent upon the experimenter's ability to develop a suitable approximation for  $f$ . Usually, a low order polynomial in some relatively small region of the independent variable space is appropriate. In many cases a first order or a second order model is used. The major objectives and applications of RSM are

1. To determine and quantify the relationship between response variables and settings of a group of experimental factors (independent variables) i.e. Mapping a response surface over a particular region of interest
2. To find the settings of experimental factors that produces the best value or the best set of values of the response variables i.e. Optimization of the responses

The major steps involved in RSM to improve an existing process/product or formulation of new product are

1. Formulation of experimental design in terms of independent variables
2. Formulation of hypothesis
3. Execution of experiments and generation of experimental data
4. Development of empirical model to predict the response variables in terms of independent variables
5. Model adequacy checking and testing of hypothesis
6. Optimization of response variables in terms of independent variables

### Formulation of Experimental Design

Factorial designs are widely used in experiments involving several factors (independent variables) to investigate the main and interaction effects of the factors on response variables. The factorial designs can be classified into two groups viz: symmetrical and asymmetrical factorial experiments. A good response surface design should possess the properties viz., detectability of lack of fit, the ability to sequentially build up designs of increasing order and the use of a relatively modest, if not minimum, number of design points. Examples on some experimental situations, where response surface methodology can be usefully employed are

**Example 1:** To Optimize the high pressure process parameters viz: pressure, ramp rate and holding time and to see its effect on high pressure treated Indian white prawn. The levels of various factors are

	Factors	Levels
1.	Pressure (MPa)	150, 250, 350
2.	Ramp Rate	300, 400, 500
3.	Holding Time(Min)	5, 10, 15

**Example 2:** For value addition to the agriculture produce, food-processing experiments are being conducted. In these experiments, the major objective of the experimenter is to obtain the optimum combination of levels of several factors that are required for the product. To be specific, suppose that an experiment related to osmotic dehydration of

the banana slices is to be conducted to obtain the optimum combination of levels of concentration of sugar solution, solution to sample ratio and temperature of osmosis. The levels of the various factors are the following

	Factors	Levels
1.	Concentration of sugar solution	40%, 50%, 60%, 70% and 80%
2.	Solution to sample ratio	1:1, 3:1, 5:1, 7:1 and 9:1
3.	Temperature of osmosis	25°C, 35°C, 45°C, 55°C and 65°C

In this situation, response surface designs for 3 factors each at five equispaced levels can be used.

In general response surface methodology is useful for all the factorial experiments in agricultural experimental programme that are under taken so as to determine the level at which each of these factors must be set in order to optimize the response in some sense and factors are quantitative in nature.

Examples of experimental design setup for RSM

1. All the factorial experiments where the factors are quantitatively measured
2. Central Composite Design
3. Box-Behnken Design
4. Simplex lattice mixture design
5. Simplex centroid mixture design
6. D-optimal design

### Development of Empirical Models

In practice the mathematical form of  $f$  discussed in the introduction is not known; we, therefore, often approximate it, within the experimental region, by a polynomial of suitable degree in variables  $x_{iu}$  (independent variables). The adequacy of the fitted polynomial is tested through the usual analysis of variance. Polynomials which adequately represent the true input-response relationship are called **Response Surfaces** and the designs that allow the fitting of response surfaces and provide a measure for testing their adequacy are called **response surface designs**. If the function  $f$  is of degree one in  $x_{iu}$  *s.i.e.* the response can be represented as

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + \dots + \beta_v x_{vu} + e_u$$

And we call it a first-order response surface in  $x_1, x_2, \dots, x_v$ .

The second-order (quadratic) response surface can be represented as

$$y_u = \beta_0 + \sum_{i=1}^v \beta_i x_{iu} + \sum_{i=1}^v \beta_{ii} x_{iu}^2 + \sum_{i=1}^{v-1} \sum_{i'=i+1}^v \beta_{ii'} x_{iu} x_{i'u} + e_u$$

This functional form has many applications in most of the agricultural experiments

The analysis of variance table for a second order response surface design is given below.

### Analysis of variance for second order response surface

Source	d.f.	S.S.
Due to regression coefficients	$2v + \binom{v}{2}$	$\hat{b}_0 \sum_{u=1}^N y_u + \sum_i \hat{b}_i \left( \sum_{u=1}^N x_{iu} y_u \right) + \sum_i \hat{b}_{ii} \left( \sum_{u=1}^N x_{iu}^2 y_u \right) + \sum_{i \neq i'} \sum \hat{b}_{i i'} \left( \sum_{u=1}^N x_{iu} x_{i'u} y_u \right) - CF$
Error	$N - 2v - \binom{v}{2} - 1$	By subtraction = SSE
Total	$N - 1$	$\sum_{u=1}^N y_u^2 - CF$

In the above table CF = correction factor =  $\frac{(\text{Grand Total})^2}{N}$ . For testing the lack of fit the sum of squares is obtained using (2.16) and then sum of squares is obtained by subtracting the sum of squares due to pure error from sum of squares due to error. The sum of squares due to lack of fit and sum of squares due to pure error are based on  $N' - 2v - \binom{v}{2} - 1$  and  $N - N'$  degrees of freedom respectively.

The lack of fit is tested using the statistic  $F = \frac{SS_{LOF} / (N' - p)}{SS_{PE} / (N - N')}$

where  $N$  is the total number of observations,  $N'$  is the number of distinct treatments and  $p$  is the number of terms included in the model.  $SS_{PE}$  (sum of squares due to pure error) has been calculated in the following manner: denote the  $l^{th}$  observation at the  $u^{th}$  design point by  $y_{lu}$ , where  $l = 1, \dots, r_u (\geq 1)$ ,  $u = 1, \dots, N'$ . Define  $\bar{y}_u$  to be average of  $r_u$  observations at the  $u^{th}$  design point. Then, the sum of squares for pure error is

$$SS_{PE} = \sum_{u=1}^{N'} \sum_{l=1}^{r_u} (y_{lu} - \bar{y}_u)^2 \quad (2.16)$$

Then sum of squares due to lack of fit ( $SS_{LOF}$ ) = sum of squares due to error -  $SS_{PE}$

It is suggested that in the experiments conducted to find an optimum combination of levels of several quantitative input factors, at least one level of each of the factors should be higher than the expected optimum. It is also suggested that the optimum combination should be determined from response surface fitting rather than response curve fitting, if the experiment involves two or more than two factors.

### Optimization of Response

The result of model-building procedure is an equation. Once the model is developed, the next stage is to optimize the process. Different type of optimization methods are

1. Method of steepest ascent/descent
2. Method of graphical evaluation of response surface plot

3. Method of desirability function analysis
4. Method of genetic algorithm

### References

- Snecdecor.G.W. and Cochran.W.G. 1989.Statistical Methods.John Wiley & Sons. New York. P 503
- Das.M.N and Giri. N.C. 1986. Design and analysis of experiments. New age international P. Ltd. New Delhi. P 488.
- Khuri, A.I. and Cornell, J.A. (1987).*Response Surfaces: Designs and Analysis*. New York: Marcel Dekker.
- Myers, R.H. and Montgomery, D.C. (1995).*Response Surfaces Methodology: Process and product optimization using designated experiments*. John Wiley and Sons.

\*\*\*\*\*