CHAPTER 1

INTRODUCTION

The study of statistical aspects of growth and instability in agricultural crops has become a popular topic for research in India since Sen (1967) delivered the technical address on the occasion of twentieth anniversary of Indian Society of Agricultural Statistics. The subject has acquired added importance and has become a topic of more serious investigation after the introduction of new technologies, both in India and abroad. So much so that it is now difficult to find an informed observer on an Indian agriculture scene, who has not contributed something on the subject. Two specific purposes for which these researches are usually made are i) To examine whether new technology has helped to accelerate crop production and ii) To examine whether new technology has made the crop production more stable.

Considering the above views, the study pertaining to “Statistical Aspects of Growth and Instability of some important crops of Maharashtra” has been undertaken. The data on the important cash crops, like Sugarcane, Cotton and Groundnut were utilised for the purposes of model building and analysis. Considering the nation as a whole, Maharashtra contributes around 14, 34, and 7 percentage towards acreage’s and 16.5, 22, and 7 percentage towards production of Sugarcane, Cotton and Groundnut respectively. As today the science of Statistics is an indispensable tool in every field of investigation, these methods in their own rights have become much more essential and widely recognized as a research tool for solving research problems of every applied field. Moreover these tools are extensively used for formulating statistical models suitable in various fields.

As Statistics is a quantitative technology of scientific research, applications of Statistics have increased more diversely during the recent past. Due to diversification of scientific research into different branches of science the
Statistical technology also developed accordingly for analysis of different types of data and drawing inferences. Moreover this has become essential for proper and widely acceptable research.

In the interest of rapid economic development, we must necessarily try to accelerate the growth of agriculture. At the same time, we have to be very careful about Stability of crop production. Using the Statistical techniques, an attempt is made in this thesis to develop appropriate measures of stability, development of models which will help for proper analysis, and also the corrective measures are given under various circumstances.

For a selected crop within a season or period to find a competent crop in view of acreage is a problem of interest to the agricultural scientists, research workers, planners because after identification of competent crop, it is easy to find out the actual shift in acreage from one group of crops to another group within a season or period and it is easy to visualise the true picture of cropping pattern. Sale (1987) noted that variations in crop acreage under Kharif Jowar, Rabri Jowar, Sugarcane and Groundnut in Maharashtra are induced by change in relative prices and other non-price variables such as area under competing crops, yield and presowing period rainfall etc. In view of this, an attempt is made in chapter 2 to give a statistical method to find out the competing crops for a selected crop within a season or period. While considering the growth in crop acreage it is essential to consider the growth of net sown area for the season or period because the total growth in crop acreage depends on growth due to net sown area and growth due to actual crop acreage due to various reasons.

1.1.1 Karl Pearson’s Coefficient of Correlation

Within season or period for a selected crop negatively correlated crops are identified using Karl Pearson’s correlation coefficient. It is a measure of intensity or degree of linear relationship between two variables. Karl Pearson (1867-1936) a British Biometrician developed a formula called correlation coefficient between two random variables X and Y. Usually denoted by \( r(X, Y) \) or simply \( r_{XY} \) is a numerical measure of linear relationship between them and is defined as
\[ r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \]

\[(x_i, y_i) \quad i = 1, 2, 3 \quad \text{n is a bivariate distribution then} \]

\[
\text{Cov}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y} \\
\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\
\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2 - \bar{y}^2}
\]

1.1.2 Regression Technique

To test whether crops are truly competent with a selected crop, regression technique is used. Regression analysis is a body of statistical methods dealing with formulation of mathematical model that depict relationship among variables and the use of these modeled relationship for the purpose of prediction and statistical inference. It is defined as mathematical relationship between original units of data. In this analysis there are two types of variables one is regressor or predictor or explanator while another is called regressed or explained variable. A regression model is defined to identify the contribution of explained variable and a Student’s- \( t \) statistic is used to test the significance of regression coefficient. The line of regression on \( Y \) of \( X \) is given by equation \( Y = a + bX \). When there are more than one explained variables the model is called multiple regression model and the coefficients are estimated using principles of least square.

1.1.3 Semi log Curve:

For finding actual growth in crop area, net sown area and total area in acreage semi log curve function is fitted. It is called an exponential growth function and is defined by \( Y = ab^t \)

Mathematically the concept of growth rate can be expressed as under

\[
\frac{(Y_{t1} - Y_{t0})}{Y_{t0}} = \Delta Y \\
Y_{t0} \quad \text{is quantity of} \ Y \ \text{at time} \ t_0 \\
Y_{t1} \quad \text{is quantity of} \ Y \ \text{at time} \ t_1 \\
Y_{t1} - Y_{t0} = \Delta Y
\]
The compound growth rates were worked out by fitting semilog curve or say exponential growth function \( Y = ab^t \). Taking logarithms of both sides we get \( \log Y = \log a + t \log b \)

\[
t \log b = \log Y - \log a
\]

\[
\log b = (\log Y - \log a) / t
\]

Put \( b = 1 + r \)

\[
\log (1+r) = (\log Y - \log a) / t
\]

\[
r = \text{antilog}[\left(\log Y - \log a\right)/t] - 1
\]

The compound growth rates of individual crop acreages, net sown area are obtained using the above equation. The effect of net sown area has been eliminated from individual crop acreage that is by subtracting the growth rates of net sown area from growth rates of individual crop acreages, actual growth in crop acreage is computed. After computing the actual growth in crop acreages the groups are formed on the basis of sign of the growth rates of crop acreages after subtracting the growth rates of net sown area. The positive sign indicates that area is shifted to that crop or group of crops whereas negative sign indicates the crop area is shifted from that crop or group of crops. The per cent shift in area from one crop to other or from one group of crops to other group is computed.

An attempt is made in chapter 3 to identify a probability model for number of seeds per pod of groundnut which is an important character of groundnut. Groundnut \((Arachis hypogea)\) is a principal economic crop of the world. This is an important oilseed crop in Maharashtra. It occupies total area of 4.4 lakh hectares and its production is of 5.5 lakh tonnes. Considering oilseed scenario of Maharashtra, groundnut occupies a dominant position. It alone contributes 25.6 percent area and adds 43.26 percent to total production. In view of this an attempt is made in chapter 3 to identify a probability model for number of seeds per pod of groundnut.
1.2.1 Binomial Distribution

A random variable $X$ is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X=x) = px = n \cdot \frac{x}{\sum x} \cdot p^x \cdot q^{n-x}, \quad x = 0,1,2,3 \ldots n$$

$$= 0 \quad \text{otherwise} \quad \quad q = 1-p$$

1.2.2 Truncated Binomial Distribution

A random variable $X$ is said to follow truncated binomial distribution, truncated at point zero if it assumes only non-zero, non-negative values and its probability mass function is given by

$$P(X=x) = px = \frac{n \cdot \frac{x}{\sum x} \cdot p^x \cdot q^{n-x}}{(1- q^n)} \quad x = 1,2,3 \ldots n$$

$$= 0 \quad \text{otherwise} \quad \quad q = 1-p$$

Truncated Binomial distribution, truncated at zero, is found to hold good for a variety 88034 of groundnut. However, this model is not fit for rest of selected varieties. Refinement of the model to mixture over the parameter $n$ of truncated binomial is tried for the rest of the selected varieties and it is observed that this model holds good for the rest eight varieties. Chi-square test is used for testing the goodness of fit.

1.2.3 Chi-square Test

The suitability of a fitted distribution is verified by the chi-square test. The test enables us to find out whether a particular model is a good fit for the empirical data or not. The test was given by K. Pearson in 1900. In testing the goodness of fit, Chi-square value is calculated from the number of observed frequencies and expected frequencies in each distinguishable class. If $E$ is expected frequency and $O$ is observed frequency then

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where summation is taken over all classes.
In Chapter 4, an attempt is made to carry out the stability analysis using Eberhart and Russel model (1966). In case, 1) error variances are heterogeneous and Varieties are not responding to environment that is interaction of Variety x Environment is not differing significantly 2) interaction of Variety x Environment is non-linear that is Variety x Environment (linear) is not differing significantly.

In Stability analysis using Eberhart and Russel model, we classify varieties into stable varieties, varieties for favorable environment and varieties for unfavorable environment. The steps involved in analysis are as follows:

1. Collect Data of different varieties over different environments in respect of different characters.
2. Test whether varieties are differing significantly using ANOVA.
3. Test Homogeneity of error variances over the environments using 'Bartlett's test.
4. Carry out Pooled analysis when error variances are homogeneous.
5. In pooled analysis if (Variety X Environment) interaction is significant, partition the variance into VXE(linear) and Pooled deviations.
6. Test whether VXE (linear) is significant.
7. If VXE (linear) is nonsignificant, characterization of variety on the basis of linear regression Co-efficient may be misleading.
8. If V X E (linear) is significant test the genotype using two stability parameters defined by Eberhart and Russel, that is, regression Co-efficient and $S^2_d$, deviation from linearity using t and F test statistics.

In step 4 of the above procedure, if the error sums of squares are heterogeneous, we cannot proceed further for analysis. In view of this, an attempt is made to use adequate transformations of the data which helps us to extend analysis as explained in the steps above.

If (Variety X Environment) interaction is non-significant it implies that varieties are not responding to the environments and further analysis using Eberhart and Russel model will be meaningless. In view of this, an attempt is
made to define a simple measure, that is Coefficient of Variation which classifies the varieties in respect of their Stability.

1.3.1 Suppose there are ‘T’ varieties and ‘S’ environments and in each environment we have ‘R’ replications So we have ‘SxR’ observations on each variety Then the steps for computing the test statistics for classifying stable varieties are as under.

1. Obtain mean of each variety that is mean of SXR observations.
2. Obtain standard deviation of each variety using SXR observations
3. Define
   \[ \% \text{Coefficient of Variation} = \frac{SD}{\text{Mean}} \times 100 \]
   Obtain \% CV for each variety
4. Arrange Co-efficients of Variations in descending order and rank them accordingly

   The Variety which corresponds to the highest rank that is lowest Coefficient of Variation is the most stable variety compared to others, provided the Coefficients of Variation with highest rank and second highest rank are differing significantly

1.3.2 The test statistic for testing significance of Co-efficients of variation (CV) suggested by K Aruna Rao and A R S. Bhatt (1995) is used.

   To test
   \[ H_0: CV(\text{1st variety}) = CV(\text{2nd variety}) \]
   define \[ Z_2 = \frac{(S1/\bar{X}1 - S2/\bar{X}2)}{[\Sigma 1/m ( (su/\bar{x}1)^4 + 1/2 (su/\bar{x}1)^2 )]^{1/2}} \]
   where S1 and S2 are Standard deviations of corresponding samples
   \( \bar{X}1 \) and \( \bar{X}2 \) are means of corresponding samples
   m Sample size
   
   \[ Z_2 \] is a normal variate and if \( Z_2 \leq 1.96 \) then accept \( H_0 \) at 5 % level of significance \( Z_2 \leq 2.58 \) Accept \( H_0 \) otherwise reject \( H_0 \) at 1 % level of significance
1.3.3 If VXE (linear) is nonsignificant, characterization of variety on the basis of linear regression Co-efficient may be misleading, that is, as per the step 8, of the procedure of analysing data using Eberhart and Russel model is invalid In view of this an attempt is made to apply weights to the data and then analyse it for the stability over the environments

The model of Eberhart and Russel is

\[ Y_{i j} = m + B_i I_j + \delta_{i j} \]  

.. A

Where \( Y_{i j} \) is the mean of \( i \)th variety in \( j \)th environment,

\( m = \) Mean of all the varieties over all the environments

\( B_i = \) The regression Co-efficient of \( i \)th variety on the environment index which measures the response of this variety to varying environment

\( I_j = \) The environmental Index which is defined as the deviation of the mean of all the varieties at a given location from overall mean

and \( \delta_{i j} = \) The deviation from regression of \( i \)th variety at \( j \)th environment.

Let us apply weights to means of each environment that is

\[ Y_{i j} w_j = m' + B_i I_j' + \delta_{i j}' \]  

.. B

Now define \( Y_{i j} w_j = y_j' \)

where \( w_j = \eta_j / se^2 \)

\( \eta_j = \) replications in individual RBD analysis.

\( se^2 = \) Error mean sum of squares of individual RBD analysis.

That is we have tried to linearise, the response of varieties by applying weights.

Now using model B for stability analysis and following the same procedure given by Eberhart and Russel we observe the boost in sum of squares of Variety X Environment (linear) Now once the Variety X Environment (linear) part is significant, the characterization of varieties on the basis of linear regression will not be misleading as per the assumption of the model given by Eberhart and Russel. After giving weights to the data, analysis and interpretation are carried out as usual.
In Chapter five, an attempt is made by using Nonparametric Methods to test whether varieties are differing significantly over the environment i.e. variety x environment i.e. interactions are differing significantly. A Non Parametric method is presented to select a stable variety over the environments. As in parametric methods of analysis, we make assumptions like independence of observations, normality of data, additivity of effects etc. In practice it is observed that these assumptions are rarely satisfied and it is not sensible to apply parametric procedures without satisfaction of these assumptions. Non parametric test procedures make certain assumptions like continuity of probability density function, independence of sample observations, existence of lower order moment. These assumptions are weaker than those of parametric procedures. Hence Nonparametric test procedures are easily applicable. Lack of knowledge of underlying probability distribution, or when observations are in the form of ranks, scores, or in the form of binary digits that is 0 and 1, or Yes/No form, the Non parametric methods are used. So the techniques are useful when data do not follow the assumptions of parametric methods.

1.4.1 Friedman's Test

As in the pooled analysis, we use ANOVA technique to test whether varieties, environments, and variety x environments are differing significantly or not, but if the data fails to follow the underlying assumptions we propose Friedman's test to test whether the varieties are differing significantly and environments effect is differing significantly and to test whether (Variety x Environment) interactions are differing significantly or not.

Friedman (1937) suggested a test statistic to test K treatments applied to N subjects and is given as,

\[ S^2 = \frac{12 \sum (R_j)^2 - 3N(K+1)}{NK(K+1)} \]
Where

\[ N = \text{number of replications.} \]
\[ K = \text{number of treatments} \]
\[ R_j = \text{sum of } j^{th} \text{ treatment} \]

When the number of rows and columns are not too small, \( S^2 \) is distributed approximately as \( \chi^2 \) with (K-1) degrees of freedom. The Friedman test determines rank total \( R_j \) is differing significantly or not. In the data of ‘t’ treatments, ‘r’ replications, and ‘l’ locations after ranking the data in different stages that is in first stage data for each location is considered separately and txr ranks are assigned to all observations and the procedure is repeated for all locations. Rank totals of each treatment is worked out. In second stage data over all replications and all locations for each treatment is considered separately and txr ranks are assigned to the observations. Then rank totals of these ranks are worked out for each replication of each location. Putting \( N = \text{txr} \) and \( K = \text{pxr} \) the same test statistic \( S^2 \) is used for testing significance of Varieties, Environments, and Variety x Environment

1.4.2 Nonparametric Measure of Stability

For a given data of ‘K’ treatments, ‘N’ environments, ‘R’ replications, and ‘L’ locations, the actual observations of given character are the effects of different factors like effect of genotypes, environments, interactions and errors. This can be expressed in the form of linear model

\[ X_{ij} = \mu + g_i + e_j + (ge)_{ij} + e_{ij} \]

where \( \mu \) is the general mean
\[ g_i = \text{the effect of } i^{th} \text{ genotype } i = 1, 2 \ldots K \]
\[ e_j = \text{the effect of } j^{th} \text{ environment } j = 1, 2 \ldots N \]
\[ (ge)_{ij} = \text{the interaction effect of } i^{th} \text{ genotype and } j^{th} \text{ environment} \]
\[ e_{ij} = \text{the average error over replications} \]

Ranking of varieties is carried out within each environment for \( N \) environments so ranks of the observations will carry effect of genotypes and interactions. Without removing genotypic effect from the data stability measure cannot be
defined because stability shall be measured independently of genotypic effect. So removing the genotypic effect that is

\[ X_{ij}^* = X_{ij} - \bar{X}_i \]

ranking is based on \( X_{ij}^* \)

\[ r_{ij} = f(ge(\eta)+ e\eta) \]

The stability measure is defined as

\[
P_{11} = \frac{1}{N} \left[ \sum_{j=1}^{N} \left| r_{ij} - M_{d1} \right| / M_{d1} \right]
\]

\( N = \text{No of environments} \)
\( r_{ij} = \text{rank of } i^{th} \text{ variety in } j^{th} \text{ environment.} \)
\( M_{d1} = \text{Median of } i^{th} \text{ variety} \)

Median of ith variety is computed for each variety and used in the formula of the statistic \( P_{11} \), as while adopting Non parametric methods it is appropriate to consider the median as a measure of central tendency rather than the mean

**1.4.3 Test of Significance of Stability measure**

As the stability measure is defined, it is noteworthy to define a test of significance also. Howsoever good a measure may be, without a proper test procedure, it is useless for practical purposes. Since \( P_{11} \) is based on ranks which is distributed as a discrete uniform random variate assuming values 1 to k Under the null hypothesis of equal stability of genotypes, for a given genotype \( i \) the ranks \( r_{ij} (j = 1,2,3 \ N) \) represent a random sample from discrete Uniform distribution, it is possible to derive the mean and variance of each of the statistics.

If the measure is approximately normally distributed then the statistics,

\[
Z = \frac{P_{11} - E(P_{11})}{\sqrt{\text{Var}(P_{11})}} \sim N(0,1)
\]
Thus a test of significance is applied based on the normal distribution. The test of significance of stability measure based on $Z$ depends on the expectation and variance of $P_{li}$. The Mean and Variance of $P_{li}$ can be estimated through simulations and ready tables are available for various values of $N$ and $K$.

In chapter six an attempt is made to carry out Path analysis and $D^2$ analysis of the Sugarcane crop. Sugarcane is an important commercial crop of India as well as of Maharashtra. The yield of the Sugarcane all over India level is 66 metric tons per hectare whereas at Maharashtra state level the yield is 83 metric tones per hectare. If we compare the yield figures of outside countries they are 150 metric tones and 220 metric tonnes per hectare for Hawaii and Peru respectively. This implies that there is large variation in the yield figures and there is scope to improve the yield by developing new varieties.

The techniques of Path analysis and $D^2$ analysis help in the process of selection of genotypes. In Path analysis total variation is classified into Genotypic, Phenotypic and Environmental variation. Direct and indirect effect of yield contributing characters, say weight of cane in this case, are computed. Addition of direct and indirect effects on yields are decided by the correlation between yield and the character chosen. This implies that if the correlation coefficient between a causal factor and the effect is almost equal to its direct effect then correlation explains the true relationship and direct selection through this trait will be effective. If the correlation coefficient is positive, but direct effect is negative or negligible, the indirect effect seem to be cause of correlation. In such situations, the indirect causal factors are to be considered simultaneously for selection.

In $D^2$ analysis we classify the varieties into different clusters on the basis of distances between the varieties and characters which contribute towards divergence. Considering the intercluster distances and percent contribution of characters towards divergence, the varieties are selected for the crossing purpose. The varieties grouped together are less divergent than those which fall in different clusters. Clusters separated by the largest statistical distance ($D^2$) show maximum divergence.
1.5.1 Component of variances

Considering that the varieties tested here are genetically Uniform, the expected mean sum of square for error that is \( \sigma^2 e \) is purely a random environmental variance attributed to 1: Varietal difference 2: Environmental variation among individual of each genotype Thus \( E(MSV) = \sigma^2 e + r\sigma^2 g \)

\[ E(MSe) = \sigma^2 e \]

and therefore

\[ \sigma^2 g = (MSV-MSe)/r \]

Thus genotypic variance being \( \sigma^2 g \) and the environmental variance \( \sigma^2 e \) that is Phenotypic variance \( \sigma^2 p = \sigma^2 g + \sigma^2 e \).

The heritability is defined as ratio of genotypic and phenotypic variance

\[ h^2 = \frac{\sigma^2 g}{\sigma^2 p} \]

1.5.2 Direct and Indirect effect

\[ r(x1,Y) = a + r(x1,x2)b + r(x1,x3)c \]

correlation between \( x_1 \) and \( Y \) is that is \( r(x1,Y) \) is partitioned into three parts namely 1 due to direct effect of \( x_1 \) on \( Y \) which amounts to a 2 due to indirect effect of \( x_1 \) on \( Y \) via \( x_2 \) which amounts to \( r(x1,x2)b \) and 3. due to indirect effect of \( x_1 \) on \( Y \) via \( x_3 \) which equals \( r(x1,x3)c \)

\( a,b,c \) are defined as under

\[ \sigma x1/\sigma Y = 'a' \]

\[ \sigma x2/\sigma Y = 'b' \]

\[ \sigma x3/\sigma Y = 'c' \]

In a similar fashion we can define \( r(x2,Y) \) and \( r(x3,Y) \)
1.5.3 \( D^2 \) values and contribution of characters towards divergence

For each combination, the mean deviation that is \( Y_{1i} - Y_{1j} \) with \( i = 1,2,3 \) \( j \) is computed and the \( D^2 \) is calculated as sum of squares of these deviations that is

\[
D^2 = \sum(Y_{1i} - Y_{1j})^2
\]

On the basis of \( D^2 \) we classify varieties into different clusters and on the basis of intercluster distances and contribution of characters towards divergence, selection of the varieties for the purpose of crossing is made. Toucher’s method is used for grouping the varieties into different clusters [Singh and Choudhary 1996]

For the purpose of contribution of individual characters towards divergence in all the combinations of the variatal differences, each character is ranked on the basis of \( d_{ij} = Y_{1i} - Y_{1j} \) values. Rank 1 is given to the highest mean difference and rank \( p \) is to the lowest mean difference where \( p \) is the lowest mean difference. The percent contribution is calculated taking total combination of variatal differences equal to 100

In chapter seven an attempt has been made to describe a probability model for number of bolls per plant of Cotton as it is proved that this character is an important yield contributing character that is lint yield or seed cotton yield

Regression technique described in 1.1.2 (of this manuscript) is used to find the contribution of the character number of bolls per plant for explaining the variation in seed cotton yield and lint yield. The data for the character number of bolls per plant is collected for three environments. Regression analysis is carried out for each individual environment as well as the pooled data over three environments

Poisson model is considered for individual environments as well as for pooled data. It is observed that Poisson distribution fit to the data of individual environment but it does not fit for the pooled data over three environments.
It is generally considered the data based on large sample size follows Normal Distribution. In this case we have 48 observations for each environment and 144 observations for the pooled data so we examined the model of normal for each individual environment as well as for pooled data. While constructing the frequency distribution for observed data the number of classes are decided using the Strüge’s Rule (1926) Chi square test of goodness of fit referred to in 123, is used for testing the goodness of fit of Poisson, Truncated Poisson as well as the Normal Distributions to the data of individual environment as well as for the pooled data.

In addition to the fitting of the distributions, the tests of skewness, Kurtosis and the Normality are applied to the data of individual environment as well as for pooled data.

1.6.1 Poisson Distribution

A discrete random variable X is said to follow Poisson distribution if it assumes only non-negative values and its probability mass function is given by

\[ P(X,\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0,1,2,3,4, \quad \lambda > 0 \]

\[ = 0 \quad \text{otherwise.} \]

\( \lambda \) is known as parameter of the distribution.

This distribution was discovered by Poisson (1837) and the peculiar property of this distribution is mean = variance = \( \lambda \). As \( \lambda \to \infty \), Poisson Distribution tends to Normal Distribution.

1.6.2 Normal Distribution

A continuous random variable X is said to follow Normal distribution if it has probability density function \( f(x) \) given as

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right), \quad -\infty < x < \infty \]

The parameters are \( E(X) = \mu \) and \( V(X) = \sigma^2 \)

The sample mean \( \bar{x} \) and standard deviation \( s^2 \) are jointly sufficient for \( \mu \) and \( \sigma^2 \)
1.6.3 Struge's Rule (1926)

While we construct a frequency distribution for observed data, we always encounter with the problem of number of classes that is what should be the appropriate number of classes depending on number of observations. The number should not be generally too large or too less. Struge's (1926) stated that the number of classes C is given by the formula: $C = 1 + 3.32 \log(N)$ where $N$ is the number of observations.

1.6.4 Test of Skewness, Kurtosis and Normality

In addition to examining fitting of the Normal distribution to the data, tests like, test for skewness of the data that is $\beta_1$, test for kurtosis (peakedness) $\beta_2$, the test statistic $W_1$ and $W_2$ are computed for each environment and for the pooled data and the data have been judged for normality.

1. The measure of skewness in population is given by $\beta_1$ which is defined as $\beta_1 = \frac{\mu^3}{\mu^2}$ and the Co-efficient of skewness is defined on the basis of $\beta_1$ that is $\gamma_1 = \sqrt{\beta}$. For the Normal distribution $\beta_1 = 0$.

2. The measure of Kurtosis (peakness) is given by $\beta_2$ which is defined as $\beta_2 = \frac{\mu^4}{\mu^2}$ and $\gamma_2 = \beta_2 - 3$. A peaked distribution shows positive kurtosis and flat-topped distribution shows negative kurtosis [Snedecor and Cochran (1967)].

3. Fisher's test depending on $\beta_1$ and $\beta_2$ is used for testing the data for normality.

The statistic $W_1 = [\beta_1(n+1)(n+3)/6(n-2)]^{1/2}$ is a normal variate with mean 0 and variance 1 for large $n$. That is $W_1$ is a standard normal variate and the calculated values are compared with 1.96 and 2.58 at 5 percent and 1 percent level of significance respectively for testing normality.
Similarly the departure from normality in respect of kurtosis is tested by using the standard normal variate $W_2$

$$W_2 = \frac{(\beta_2 - 3 + 6/(n+1))[(n+1)^2 (n+3)(n+5)]^{1/2}}{24n(n-2)(n-3)}$$

Calculated values of $W_2$ are compared with 1.96 and 2.58 at 5 percent and 1 percent level of significance respectively.