C H A P T E R - I

DESCRIPTIVE STATISTICS FOR STATISTICAL MODELS

1.1 Introduction :

This chapter consists a literature on descriptive statistics for statistical models. In section (1.2) a model is described; steps involved in proposing model are discussed and some examples of the parametric models are also given. In section (1.3), the need , definition , and some desirable properties of descriptive statistics have been discussed. In the last section the chapter wise summary of this dissertation is given. In the following we discuss in detail about models.

1.2 Models :

When we remember our childhood, and the toyes that we used, then we realized that these toyes playes an important role in understanding the world around us. For example, models of a train, aeroplane, doll, building sets etc., are sufficient to describe the real ones, up to a certain extent. An exact model might be very much complicated, and would take very long time to understand it, and might be very expensive to build; however a knowledgeable person can propose various models to describe a

phenomenon or a thing under study.

A proposed models may not be able to give the descriptions of every part. When a particular model gives a good description of one aspect of situation, at the same time it may give poor description of another aspect. For example a aeroplane-toy looks like a aeroplane but it can not fly or it may fly but does not look like a plane. Thus the models are only an approximations to the aspects of reality.

In a similar manner some real life phenomenon where certain kind of randomness appears can be modeled through some mathematical or statistical or probabilistic or stochastic models. The purpose of presenting model mathematically is that it is clearly defined and easily communicated. The benefit from this we get is; its strength and weakness may be analyzed. It may be possible that there are some errors in model due to precise situations. Hence we lack a clear knowledge of the underlying mechanisms and relationships in the situations. Thus we may be faced with possibility of uncertainty and finally model will be wrong. Wrong in the sense that may be wrong in the form or in numerical values. To understand an appropriate model we have to use statistical methods.

The problem is how to select a statistical model to the situation under study ; Warren Gilchirst (1984) in the book

"Statistical Modeling" suggests the five major stages in statistical modeling which are as follows;

I. Identification :

At this stage we find or choose an appropriate model for a given situation.

II. Estimating and fitting :

The general form of the model may be of interest. But to use it in practice it must give some numerical form. In this stage we give some numerical values to our model assumed.

III. Validation :

The process of comparison of the model with the observed world is called validation. This stage is proposed to examine whether the model is a good description of prototype in terms of its behaviour and of the application proposed.

IV. Application :

At this stage we are concerned with how the nature of the application may be taken into account when carrying out the process of identification, estimation and validation.

V. Iteration :

This is the last stage in statistical modeling. At this stage we consider the ways in which we might iterate the process identification, estimation validation and application and can make an improvement in the process of modeling. There are two types of models:

i) Parametric models and

ii) Non-parametric models.

In parametric models we have to make some assumptions about the form of distribution of the parent population. In case of parametric models the distribution in question is characterized by a parameter and specify some conditions on parameter. In the chapter III of this dissertation we describe in detail the non-parametric models.

The purpose of model fitting is to use the data to estimate the form of the relationship between the variable under study(Y say) and one or more measurements X_j ($j = 1,2, \ldots, k$). Now we give few examples of parametric models.

Example (1.2.1) :

i) Family of all normal distribution is characterized by expection and variance.

ii) Family of Pearson curves characterized by first four moments i.e. by the classical measures; measures of location, scale, skewness and kurtosis. For detail we refer [Ref. Kendall and Stuart, 1948].

Some of the densities of family of Pearson curves are as follows;

1. Type I (Beta Distribution) : The density is

$$f(x) = \begin{cases} \frac{1}{B(p,q)} x^{p-1} (1-x)^{q-1} & 0 \le x \le 1;\ p, q > 0. \\ 0 & \text{otherwise.} \end{cases} \qquad (1.2.1)$$

2. Type II : The density is

$$f(x) = \begin{cases} \frac{1}{a\ B(1/2, m-1/2)} (1-x^2/a^2)^m & -a \le x \le a. \\ 0 & \text{otherwise.} \end{cases} \qquad (1.2.2)$$

3. Type III (Gamma Distribution) : The density is

$$f(x) = \begin{cases} \frac{1}{\Gamma(\lambda)} x^{\lambda-1} e^{-x} & 0 \le x \le \infty\ ;\ \lambda > 0 \\ 0 & \text{otherwise.} \end{cases} \qquad (1.2.3)$$

4. Type IV : The density is

$$f(x) = \begin{cases} k\,(1 + x^2/a^2)^{-m} \exp\{-\nu \arctan(x/a)\} & m > 1/2. \\ 0 & \text{otherwise.} \end{cases} \qquad (1.2.4)$$

5. Type VI : The density is

$$f(x) = \begin{cases} \dfrac{1}{B(p,q)} \dfrac{x^{p-1}}{(1+x)^{p+q}} & 0 \le x \le \infty. \\ 0 & \text{otherwise.} \end{cases} \qquad (1.2.5)$$

Moreover in designs of experiments we model through CRD, RBD, LSD, BIBD etc. For example in RBD we use the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}. \qquad (1.2.6)$$

where,

Y_{ij}- The observation received i^{th} treatment in j^{th}block.

μ - Overall mean effect.

α_i - Effect due to i^{th} treatment.

β_j -Effect due to j^{th}block.

ε_{ij}- i.i.d. N(0,1).

Gauss Markov set up is also a model, which is a linear model.

To propose suitable model, relevant information (data) is to be obtained. The data so collected might be huge and needs manipulation from the point of utility. An objective of statistics is to get a statistic which describes the characteristics of the entire mass of data. This is possible by analyzing huge data. There are differnet methods of analyzing data depending upon what type of analysis is required. For example every nation collects and compiles statistical data to provide descriptive information relative to all sorts of things such as taxes and agricultural crops etc. We are able to use statistical methods that not only describe important features, but also allow us to use the collected data for decision making through generalizations and predictions. We shall categorize the statistical methods into one of the two major areas called

(i) descriptive statistics and

(ii) statistical inference.

In the following we describe descriptive statistics.

1.3 Need of descriptive statistics :

Descriptive statistics comprises methods concerned with collecting and describing a set of data so as to yield meaningful information. It deals with measures of different aspects of population (or a distribution of population values). The

population may be finite or infinite. Typical example of descriptive measures are : the mean or median as measures of location, the standard deviation or inter quartile range as measures of scale and some others are the classical measures of skewness, kurtosis and correlation.

Usually when defining such a measures one has in mind not a single population for which the measure is to be computed, but for a family of such populations. In particular, one is typically concerned with a family to which the given population is assumed to belong. The choice of suitable measure then depend upon the nature of the family. In the following we give the definition of descriptive statistics and its properties. One may refer Harry and Althoen (1994).

Definition (1.3.1) : Descriptive statistics :

An index which describes some characteristics of a frequency or relative frequency distribution is called as descriptive statistic.

Properties :

1. A descriptive statistic should be single valued.

By stating this property author wants to say that the descriptive property of collection of data should be represented

by a single number.

2. A descriptive statistics should be algebraically tractable.

One should be able to calculate, transform or otherwise manipulate a descriptive statistics using ordinary arithmetic operations, such as addition, subtraction, multiplication, roots and powers.

3. A descriptive statistics should consider every observed value.

By definition a summary (numerical or otherwise) involves some loss of information. This loss is reduced if every numerical value that appears in the collection of measurements is represented in the calculation of a descriptive statistics.

4. A descriptive statistics should consider the frequency of every observed value.

A collection of data is likely to include some numbers that appears more often than others. A descriptive statistics should somehow characterize the entire collection of data, each numerical value should therefore contribute to the statistics in proportion to the frequency with which it observed.

Some well known measures of location are mean, median, percentile etc. and measures of dispersion are range, standard deviation, quartile range etc. Other descriptive measures that describe the nature of the distribution are measure of skewness,

which indicates the degree of departure from symmetry and measures of kurtosis, which indicates the concentration (peakedness) of a distribution. Standard measures of skewness and kurtosis are respectively, β_1 and β_2 which are defined,

$$\beta_1 = \mu_3^2/\mu_2^3 \qquad (1.3.1)$$

and

$$\beta_2 = \mu_4/\mu_2^2 \qquad (1.3.2)$$

It should be noted that the above measures have simple interpretation and computed easily in univariate case. These measures have been investigated in many works Bickel and Lehmann(1975, 1976); Oja(1981a); Vanzwet(1964). But it is difficult to generalise these concepts for multivariate cases. For example how should one interprete the terms skewness or kurtosis? Can we compare two distributions with respect to these properties? It is also necessary to introduce such measures to specified classes of distributions. The underlying class may be the class of all non-parametric distributions. The above described measures can be extended to both mutivariate models and non-parametric models. Through this dissertation we will try to discuss the litrature on discriptive statistics for multivariate as well as for non-parametric models.

When the data consists several number of variables, one needs a suitable method to analyse the data. In such situation we use some standard techniques for analysing the data. All these techniques put together form multivariate statistics. Multivariate statistics techniques are widely used in social sciences and other braches. As compared to the univariate statistics it is noticed that the study of multivariate statistics needs depth of knowledge and analytical ability.

Genreally multivariate statistics represents the expansion of more familiar univariate and bivariate statistics. Univariate and bivariate statistics are special cases or just simplifications of more general multivariate models. With the help of multivariate statistics we can analyse and study simultaneously more than one variable. The decriptive statistics for multivariate models have been introduced by Mardia(1970); Oja (1983). We see in detail about the desriptive multivariate statistics in chapter II of this dissertaion.

In statistical estimation problem generally we assumes that the observed quantities are independent random variables with common probability distribution F_{θ_o} which belongs to the set of distributions $\{F_{\theta_o} \ \theta_o \in \Theta\}$, Θ is called the parameter space. The problem is to estimate θ_o, the true value of the parameter, based

on observations,that is one may try to find a mapping from the space of all possible sets of observations to the parameter space Θ which takes values 'close' to θ_o with high probability (computed when θ_o, $\theta_o \in \Theta$,is the true value of the parameter, if F_{θ_o} is the underlying distribution). Such a setup is called as parametric model. For example the normal models which assumes the real valued observations are normally distributed with unknown mean and known/unknown variance; the parameter is to be estimated is mean/the pair (mean, variance).

In non-parametric models no assumptions are made about the underlying distributions except that the distribution function being sampled is absolutely continuous or purely discrete. These models do not specify any condition about the parameter(s) of parent population. In this case there is a parameter which indexes the family of absolutely continuous distribution functions, but it is not numerical; hence the parameter set can not be represented as a subset of $\mathbb{R}^k$ for any $k \geq 1$. In non-parametric models we consider an empirical distribution function. In the third chapter we have discussed the descriptive statistics for non-parametric models introduced by Bickel and Lehmann (1975, 1976).

In the last chapter we propose some new measures. The first section is introductory one. In second section we give some new measures of location which are obtained by combing two measures, and related some results are also given. [Ref. Rattihalli(1996)]. In section 4.3 the measure for peakedness introduced by Horn (1983) has been discussed. Using this measure we can find the kurtosis of distributions even when moments do not exists for example Cauchy. Examples and properties of this measure are also discussed.

Generally we say that for normal distribution kurtosis is 3. If kurtosis is not equal to 3, then the distribution is not normal. In the last section of chapter IV we made an interesting attempt which contains wide class of symmetric distributions whose kurtosis is equal to 3. Such class was introduced by Kale and Sebastian (1996).

* * * * *