

Chapter 1

WHY STATISTICS?

1.1 Introduction

Rational decision making at any stage needs information. If the government of a country wants to find out how many schools have to be started in a particular region, how many hospitals have to be built in a particular district or whether how many highways have to be extended or improved, it needs to collect the information on the present state of affairs and make an assessment on the needs and finances available prior to taking action. If an industry wants to expand its base in manufacturing, it needs to have information on its capacity to produce the product and on the demand for the product. If the government of a region wants to find out the average income of a household in a particular area or percentage employed in a particular state, a scientific decision can be made only after it obtains observations or data on such issues. If a manufacturer of drugs wants to introduce a new drug into the market for some disease, it has to study the performance of the new drug it discovered through clinical trials and possibly compare its performance to other drugs presently used possibly in the market for the same disease so that it can arrive at a rational conclusion. Such a procedure needs data on the recovery aspects of the new drug besides its side effects as compared to the old drug. If a meteorologist wants to predict or forecast the weather during a particular period in a year, say, in the month of November in a particular year, he or she can make a rational decision only if the information on the weather patterns during the month of November and possibly earlier months in the previous years is available. If a stock broker wants to know how the share prices of a company are likely to fluctuate in a specified time of the year, information on the volatility in the share prices of the company during the previous trading seasons

is needed besides the general patterns of volatility of share prices in the stock market. If a commodity salesman wants to find how the prices of a particular commodity is likely to fluctuate in a given time of the year, he or she has to have information on the prices during the corresponding previous trading season. There is a common link that is apparent in all these situations. Any rational decision making process involves data and information. The subject of Statistics deals with developing the methods and the techniques for such a decision making process. Modelling of an observed data is an important statistical problem as it possibly brings out or indicates the causes leading to or underlying the phenomenon giving rise to this data. If a particular model does not explain the data or does not help in forecasting or predicting the data reasonably accurately or the actual observations do not follow the pattern indicated by the model, it is necessary to take a fresh look at the data and try to refine or improve the model. Of course it is important not to clutter the model to be constructed with too much information for the simple reason that too many restrictions lead to a model which is problem specific and will not be applicable to similar but slightly different problems. A model should be such that it encompasses or unifies several similar problems taking into account the common features of these problems for unified study or discussion. A basic requirement for such a study of the model building is to know the distribution or shape of the data and other relevant features of the data.

1.2 Representation of Data

We will now describe a method of representing a data or information through an example.

Suppose we are interested in estimating the average income of all the people employed in a particular city. A first stage in obtaining the estimate is to prepare a list of all people, choose a representative sample of the people employed in the city and then collect the information from the sample chosen. It is clear that if our sample consists of people only from the low income group or if all the people sampled come from the high income strata, the sample chosen is not representative. How to choose a representative sample from the city? We will not deal with this question here. The subject of study of methods of choosing samples is known as *Survey sampling*. Suppose we have obtained such a set of observations. Let us denote the sample of observations by $X_i, 1 \leq k \leq n$. We choose an ori-

gin x_0 and a positive real number h , called binwidth and divide the whole real line into intervals $[x_0 + mh, x_0 + (m + 1)h)$ of length h , called bins. On every such bin of length h , erect a rectangle with its height equal to $\frac{1}{nh} \times$ (number of observations falling in that bin). The graph so obtained from the data is called a *histogram* for the data. The shape of a histogram for a data depends on the choice of the origin x_0 as well as the bin width h of the rectangles erected. There is no unique representation for any data using a histogram because the choice of the origin and the choice of the bin-width h determine the groups into which the data is divided. For practical reasons, it is convenient to choose h so that the number of groups is neither too large nor too small. This can be achieved by finding the *range* r of the observations, that is, the difference between largest and the smallest of the observations. If we want to divide the data into k classes, we can choose h to be $\frac{r}{k}$. An important use of a histogram is that it will indicate the shape of the data, for instance, whether it is unimodal or multimodal, whether it is bell-shaped or whether it is skewed from the left or from the right and so on. This is especially useful in modelling the data as we shall see later in this book. In order to develop statistical methods useful for analysis any data, we will first introduce the notion of probability of an event and study related concepts in the next chapter.