Introduction to Statistics

Origin and Growth of Statistics

The origin of statistics can be traced back to the primitive man, who put notches on trees to keep an account of his belongings. During 5000 BC, kings used to carry out census of populations and resources of the state. Kings of olden days made their crucial decisions on wars, based on statistics of infantry, and elephantry units of their own and that of their enemies. Later it enhanced its scope in their kingdoms' tax management and administrative domains. Thus, the word 'Statistics' has its root either to Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' each of which means a 'political state'. The word 'Statistics' was primarily associated with the presentation of facts and figures pertaining to demographic, social and political situations prevailing in a state/government. Its evolution over time formed the basis for most of the science and art disciplines. Statistics is used in the developmental phases of both theoretical and applied areas, encompassing the field of Industry, Agriculture, Medicine, Sports, and Business analytics.

Statistics is concerned with scientific method for collecting, organizing, summarizing, presenting, analyzing, and interpreting of data. The word statistics is normally referred either as numerical facts or methods.

Statistics is used in two different forms-singular and plural. In plural form it refers to the numerical figures obtained by measurement or counting in a systematic manner with a definite purpose such as number of accidents in a busy road of a city in a day, number of people died due to a chronic disease during a month in a state and so on. In its singular form, it refers to statistical theories and methods of collecting, presenting, analyzing, and interpreting numerical figures.

Though the importance of statistics was strongly felt, its tremendous growth was in the twentieth century. During this period, lot of new theories, applications in various disciplines were introduced. With the contribution of renowned statisticians several theories and methods were introduced, naming a few are Probability Theory, Sampling Theory, Statistical Inference, Design of Experiments, Correlation and Regression Methods, Time Series and Forecasting Techniques.

In early 1900s, statistics and statisticians were not given much importance but over the years due to advancement of technology it had its wider scope and gained attention in all fields of science and management. We also tend to think statistician as a small profession but a steady growth in the last century is impressive. It is pertinent to note that the continued growth of statistics is closely associated with information technology. As a result, several new inter- disciplines have emerged. They are Data Mining, Data Warehousing, Geographic Information System, Artificial Intelligence etc. Now-a-days, statistics can be applied in hardcore technological spheres such as Bioinformatics, Signal processing, Telecommunications, Engineering, Medicine, Crimes, Ecology, etc.

Today's business managers need to learn how analytics can help them make better decisions that can generate better business outcomes. They need to understand the statistical concepts that can help analyze and simplify the flood of data around them. They should be able to leverage analytical techniques like decision trees, regression analysis, clustering and association to improve business processes.

Definitions of Statistics

Statistics has been defined by various statisticians.

'Statistics is the science of counting' ***-A. L .Bowley***

'Statistics is the science which deals with the collection, presentation, analysis and interpretation of numerical data' **- *Croxton and Cowden***

Wallist and **Roberts** defines statistics as *"Statistics is a body of methods for making decisions in the face of uncertainty"*

Ya-Lun-Chou slightly modifies Wallist and Roberts definition and come with the following definition: *"Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risk."*

It may be seen that most of the above definitions of statistics are restricted to numerical measurements of facts and figures of a state. But modern thinkers like Sacrist defines statistics as

'By statistics we mean the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated to reasonable standards of accuracy collected in a systematic manner for a predetermined purpose and placed in relation to each other'.

Among them, the definition by Croxton and Cowden is considered as the most preferable one due to its comprehensiveness. It is clear from this definition that statistics brings out the following characteristics.

Characteristics of Statistics:

- **Aggregate of facts collected in systematic manner for a specific purpose**.

  Statistics deals with the aggregate of facts and figures. A single number cannot be called as statistics. For example, the weight of a person with 65kg is not statistics but the weights of a class of 60 persons is statistics, since they can be studied together, and meaningful comparisons are made in relation to the other. This reminds us of Joseph Stalin's well known quote, "One death is a tragedy; a million is a statistic." Further the purpose for which the data is collected is to be made clear, otherwise the whole exercise will be futile. The data so collected must be in a systematic way and should not be haphazard.

- **Affected by large number of causes to marked extent.**

  Statistical data so collected should be affected by various factors at the same time. This will help the statistician to identify the factors that influence the statistics. For example, the sales of commodities in the market are affected by causes such as supply, demand, and import quality etc. Similarly, as mentioned earlier if a million deaths occur the policy makers will be immediately in action to find out the causes for these deaths to see that such events will not occur.

- **Numerically expressed.**

The statistical facts and figures are collected numerically for meaningful inference. For instance, the service provided by a telephone company may be classified as poor, average, good, very good and excellent. They are qualitative in nature and cannot be called statistics. They should be expressed numerically such as 0 to denote poor,1 for average, 2 for good, 4 to denote very good and 5 for excellent. Then this can be regarded as statistics and is suitable for analysis. The other types of quality characteristics such as honesty, beauty, intelligence, defective etc. which cannot be measured numerically cannot be called statistics. They should be suitably expressed in the form of numbers so that they are called statistics.

- **Enumerated or estimated with a reasonable degree of accuracy.**

The numerical data are collected by counting, measuring or by estimating. For example, to find out the number of patients admitted in a hospital, data is collected by actual counting or to find out the obesity of patients, data are collected by actual measurements on height and weight. In a large scale study like crop estimation, data are collected by estimation and using the powerful sampling techniques, because the actual counting may or may not be possible. Even if it is possible, the measurements involve more time and cost. The estimated figures may not be accurate and precise. However certain degree of accuracy must be maintained for a meaningful analysis.

- **To be placed in relation to the other.**

One of the main reasons for the collection of statistical data is for comparisons in order to make meaningful and valid comparisons, the data should be on the same characteristic as far as possible. For instance, we can compare the monthly savings of male employees to that of the female employees in a company. It is meaningless if we compare the heights of 20 year-old boys to the heights 20 year- old trees in a forest.

Types of Statistics

There are two kinds of Statistics, which are descriptive Statistics and inferential Statistics. In descriptive Statistics, the Data or Collection Data are described in a summarized way, whereas in inferential Statistics, we make use of it in order to explain the descriptive kind. Both are used on a large scale. Statistics is mainly divided into the following two categories.

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

- Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars, or scatter plots.
- Descriptive Statistics are just descriptive and thus do not require normalization beyond the Data collected.

Inferential Statistics

In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarized we use Inferential Statistics to describe the Meaning of the collected Data.

- Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.
- Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
- Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

Functions of Statistics

The functions of statistics can be elegantly expressed as 7 - C's as:

S.NO	Functions	What it does
1	Collection	The basic ingredient of statistics is data. It should be carefully and scientifically collected
2	Classification	The collected data is grouped based on similarities so that large and complex data are in understandable form.
3	Condensation	The data is summarized, precisely without losing information to do further statistical analysis.
4	Comparison	It helps to identify the best one and checking for the homogeneity of groups,
5	Correlation	It enables to find the relationship among the variables
6	Causation.	To evaluate the impact of independent variables on the dependent variables.
7	Chance	Statistics helps make correct decisions under uncertainty.

Scope and Applications

In ancient times the scope of statistics was limited. When people hear the word 'Statistics' they think immediately of either sports related numbers or a subject they have studied at college and passed with minimum marks. While statistics can be thought in these terms there is a wide scope for statistics. Today, there is no human activity which does not use statistics. There are two major divisions of statistical methods called descriptive statistics and inferential statistics and each of the divisions are important and satisfies different objectives. The descriptive statistics is used to consolidate a large amount of information. For example, measures of central tendency, like mean are descriptive statistics. Descriptive statistics just describes the data in a condensed form for solving some limited problems. They do not involve beyond the data at hand.

Inferential statistics, on the other hand, are used when we want to draw meaningful conclusions based on sample data drawn from a large population. For example, one might want to test whether

a recently developed drug is more efficient than the conventional drug. Hence, it is impossible to test the efficiency of the drug by administering to each patient affected by a particular disease, but we will test it only through a sample. A quality control engineer may be interested in the quality of the products manufactured by a company. He uses a powerful technique called acceptance sampling to protect the producer and consumer interests. An agricultural scientist wanted to test the efficacy of fertilizers should test by designed experiments. He may be interested in farm size, use of land and crop harvested etc. One advantage of working in statistics is that one can combine his interest with almost any field of science, technology, or social sciences such as Economics Commerce, Engineering Medicine, and Criminology and so on.

The profession of statistician is exciting, challenging and rewarding. Statistician is the most prevalent title but professionals like Risk analyst, Data analyst, Business analyst have been engaged in work related to statistics. We have mentioned earlier that statistic has applications to almost all fields. Here in this section, we highlight its applications to select branches.

- **Statistics and actuarial science**

  Actuarial science is the discipline that extensively applies statistical methods among other subjects involved in insurance and financial institutions. The professionals who qualify in actuarial science course are called actuaries. Actuaries, in the earlier days used deterministic models to assess the premiums in insurance sector. Nowadays, with modern computers and sophisticated statistical methods, science has developed vastly.

- **Statistics and Commerce**

  Statistical methods are widely used in business and trade solutions such as financial analysis, market research and manpower planning. Every business establishment irrespective of the type must adopt statistical techniques for its growth. They estimate the trend of prices, buying and selling, importing, and exporting of goods using statistical methods and past data. Ya-Lun-Chou says, “It is not an exaggeration to say that today nearly every decision in business is made with the aid of statistical data and statistical methods.”

- **Statistics and Economics**

Statistical methods are very much useful to understand economic concepts, such as mandatory policy and public finance. In the modern world, economics is taught as an exact service which makes extensive use of statistics. Some of the important statistical techniques used in economic analysis are Times series, Index Numbers, Estimation theory and Tests of significance, stochastic models. According to Engberg "No Economist would attempt to arrive at a conclusion concerning the production or distribution of wealth without an exhaustive study of statistical data." In our country many state governments have a division called Department of Economics and Statistics for the analysis of Economic data of the state.

- **Statistics and Medicine**

In medical field, statistical methods are extensively used. If we look at the medical journals one can understand to what extent the statistical techniques play a key role. Medical statistics deals with the applications of statistical methods like tests of significance and confidence intervals to medicine and health science including epidemiology, public health. Modern statistical methods help the medical practitioners to understand how long a patient affected by a dreaded disease will survive and what are the factors that influence a patient to be alive or dead.

- **Statistics and Agriculture**

Experimentation and inference based on these experiments are the key features of general scientific methodology. Agricultural scientists conduct experiments and make inferences to decide whether the particular variety of crop gives a better yield than others or a particular type of fertilizer etc, there are several institutes where research is being done by making use of statistical methods like analysis of variance (ANOVA), factorial experiments etc., falls under the hut of Design of experiments.

- **Statistics and Industry**

Statistical methods play a vital role in any modern use of science and technology. Many statistical methods have been developed and applied in industries for various problems. For example, to maintain the quality of manufactured products the concept of statistical quality control is used. The quality in time domain study of mechanical, electrical, or electronic items the concept of 'Reliability' has emerged. Total quality management and six-sigma theories make use of statistical concepts.

- **Statistics and Information Technology**

Information Technology is the applications of computers and telecommunication equipment to store, retrieve, transmit and manipulate data. Now-a-days, several industries are involved in information technology and massive amounts of data are stored every day. These data are to be analyzed meaningfully so that the information contained in the data is used by the respective users. To address this issue, fields such as data mining, Machine learning have emerged. Data mining an interdisciplinary sub field of computer science is the computational process of discovering patterns in large data sets involving methods such as artificial intelligence and statistics. Persons trained in statistics with computing knowledge have been working as data analytics to analyze such huge data.

- **Statistics and Government**

Statistics provides statistical information to government to evolve policies, to maintain law and order, to promote welfare schemes and to other schemes of the government. In other words, statistical information is vital in overall governance of the state. For instance, statistics provide information to the government on population, agricultural production, industrial production, wealth, imports, exports, crimes, birth rates, unemployment, education, minerals and so on.

Limitation of Statistics

Although Statistics has wide field of application, it has some limitations. Some of these limitations are as follows.

- **Qualitative Aspect Ignored:**

The statistical methods don't study the nature of phenomenon which cannot be expressed in quantitative terms. Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. It needs conversion of qualitative data into quantitative data. So, experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

- **It does not deal with individual items:**

It is clear from the definition given by Prof. Horace Sacrist, "By statistics we mean aggregates of facts and placed in relation to each other", that statistics deals with only aggregates of facts or items, and it does not recognize any individual item. Thus, individual terms as death of 6 persons in an accident, 85% results of a class of a school in a particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

- **It does not depict entire story of phenomenon:**

When even phenomena happen, that is due to many causes, but all these causes cannot be expressed in terms of data. So, we cannot reach at the correct conclusions. Development of a group depends upon many social factors like, parents' economic condition, education, culture, region, administration by government etc. But all these factors cannot be placed in data. So, we analyze only that data we find quantitatively and not qualitatively. So, results or conclusion are not 100% correct because many aspects are ignored.

- **Statistics can be misused**

Only the experts or statistician can handle statistical data properly. It is likely to be misused the Statistics by non-statistical persons in handling data and interpreting the result.

- **Statistical laws are not exact:**

Generally statistical laws are probabilistic in nature. Based on probability or interpolation, we can only estimate the production of paddy in 2008 but cannot make a claim that it would be exactly 100 %. Here only approximations are made.

- **Results are true only on average:**

As discussed above, here the results are interpolated for which time series or regression or probability can be used. These are not true. If average of two sections of students in statistics is same, it does not mean that all the 50 students is section A has got same marks as in B. There may be much variation between the two. So, we get average results.

- **To Many methods to study problems:**

In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case.