CHAPTER

# 1 Introduction

Statistics has its own vocabulary. We are frequently reminded about the fact that we are living in the information age. The subject statistics as it seems, is not a new discipline, but it is as old as the human society itself. It has been used right from the existence of life on this earth although the sphere of its utility was very much restricted. In the olden days statistics was regarded as the "science of Statecraft" and was the byproduct of the administrative activity of the state.

The word statistics seems to have been derived from the Latin word, status or the Italian word ‘statista’ or the German word, statistik or the French word ‘statique’ each of which means a political state.

In the ancient time the scope of statistics was primarily limited to the collection about

1. Age and sex-wise population of the county
2. Property and wealth of the Country.

Presently the statistics has been extended to Mechanical science, Pharmaceutical science, Medical science, life science and Paramedical science for collection and to analysis of data.

Historical evidences about the prevalence of a very good system of collecting vital statistics and registration of births and deaths even before 300 BC are available in Kautilya.

Sixteenth century saw the applications of the statistics for the collection of the data relating to the movements of heavenly bodies-stars and planets to know about their positioning and for the prediction of eclipses.

Seventeenth century witnessed the origin of vital statistics. Statistics was systematically applied to study of the birth and death statistics. The computation of mortality tables and the calculation of expectation of life at different ages led to the idea of "life insurance".

Modern stalwart in the development of statistics who contributed how to apply statistics in different fields to analyze the data. Francis Galton (1822-1921) who designed the regression analysis technique, Karl Pearson (1857-1936) developed the correlation analysis technique, also he developed Chi-square test ($\lambda^2$- test) of goodness of fit to analyze non-Parametric test.

RA Fisher is the real giant in the field of statistics, because he is pioneer in estimation of theory, sampling distribution, analysis of variance and design experiments. These are the statistical tools which are most frequently used in pharmaceutical calculation to analyze the experimental data and to verify the level of significance of the result.

Before RA Fisher, W.S.S Gosset one who developed the t-test to analyze the data of small size.

## SOME BASIC CONCEPTS

**Data:** The raw material of statistics is data. Statistics are a set of numerical data, in fact only numerical data constitute statistics. Thus the raw material of statistics always originates from the operation of counting (enumeration) or measurements.

The person conducts the statistical measures are the characteristics under study to carry out further statistical analysis is known as investigator.

For our purpose we can define data as numbers - these data will be collected in two ways:

1. Measurement
2. Counts

When an Investigator measures the weights and temperatures of patients is measurement and counts is the total number of patients in different age Group is counting.

Another example is considered when hospital administrator counts the number of patients who have admitted and discharged in a day can be taken as count.

The entire structure of the statistical analysis for any enquiry is based upon systematic collection of data. When an investigator while doing clinical trials experiment, he has to collect data for every half an hour till the completion of clinical trial.

Distribution of concentration of drug in the blood and complete elimination of drug from the blood. That will become more accurate data to apply statistical tools to analyze and do draw inference about the results obtained.

## STATISTICS

One can define the statistics as a field of study concerned with:

1. The collection, organization, summarization and analysis and
2. Drawing inferences about a body of data when only a part of the data is observed.

The field of utility of statistics has been increasing steadily in different field like, medical pharmaceutical, paramedical, life science and other different areas, people defined it in different ways.

Generally the Statistical methods embodying the theory and techniques used for collecting, analyzing and drawing inferences for the numerical data. Hence, statistics can be defined as the science of collection, presentation, analysis and interpretation of numerical data.

## BIO STATISTICS

When the tools of statistics are applied to analyze the medical, biological and pharmaceutical sciences data, then it can be named as biostatistics.

Biostatistics is contraction of biology and statistics. Sometimes referred to as biometry or biometrics, is the application of statistics to a wide range of topics in biology.

# PHARMACEUTICAL STATISTICS

Pharmaceutical statistics is the application of statistics, to the matters of concerning the pharmaceutical industry. The example is the design of experiments to analyze the product, to analyze the clinical trial experiments etc.

***Example***

1. To evaluate the activity of drug
   Ex: heat of caffeine on attention compare the analgesic effect of plant extract and NSAID
2. To explore whether the changes produced by the drug are due to the action of drug by chance
3. To compare the action of two or more different drugs or different dosages of the same drug are studied using statistical methods
4. To find an association between disease and risk factors such as coronary artery disease and smoking
5. Design and Analysis clinical trials in medicine

The science of biostatistics encompasses the design of biological experiments especially in medicine and pharmaceutical science.

# VARIABLES

When we observe a characteristic we find that it takes on different values in different persons, places or things, we label those values are as a variable. The reason for doing this is that characteristic is not same when observed in different possessors of it. In general variables are termed as the measurements of the values which are the characteristics of data collected after performing the experiment.

*Example*: Blood pressure, sugar level, heartbeat, heights of patients, age of patients which area observed in clinic.

## Types of Variables

## Quantitative Variable

A quantitative variable is one that can be measures in the usual sense.

When we measure height of patients come to clinics, that variable can be called as quantitative variables. Measurements made on quantitative variable convey information regarding amount.

## Qualitative Variables

Characteristic can be categorized only.

For example when a person with diseased diagnose, is designated as belongings to ethnic group of a person or object is said to possess or not to possess some characteristic of interest. In this case measurement consists of categorizing. These variables are termed as qualitative variables

### Discrete Variable

Those variable which cannot take all the possible values within a given specified range are termed as discrete (discontinuous) variable. Discrete variable can take on a countable number of values. These variables are commonly observed in biological and Pharmaceutical experiments or clinical trial experiments. For example when patients are treated using a particular drug, then the different kinds of side effects of drugs will be measured. This type of measurement can be considered or can be included in discrete variable group.

Even the number of daily admissions to a general hospital in a day may be considered as discrete variable, because number of admissions each day must be represented by a whole number, such as 0, 1, 2 or 3.

The number of decayed, missing of filled teeth in a dental checkup camp in a village can be considered as a discrete variable.

### Continuous Variable

Those variable which can take all the possible values (integral as well as fractional) within same range or interval can be termed as continuous variable (i.e., within a specified lower and upper limit). The limiting factor for the total number of possible observations or results is the sensitivity of the measuring instrument.

For Example, the age of patients in a hospital is a continuous variable because age can take all possible values (it can be measured to a nearest fraction)

Time: years, month, day, minutes, Seconds etc., in a certain range, say to 10 years or 10-20 years etc.,

More precisely a variable is said to be continuous if it is possible of passing from any given value to the next value by infinitely small gradation

*Ex*: Height, weight, temperature are continuous variable

## FREQUENCY DISTRIBUTION

The organization of the data pertaining to a quantitative phenomenon. The frequency distribution can be defined as a table in which data's are grouped into classes and number of items which are falling in each class will be recorded.

It is an important function of statistics to facilitate the comprehension and meaning of large quantities of data by constructing simple data summaries.

The Frequency distribution can also be defined as a table or categorization of the frequency of occurrence of variables in various class intervals.

Sometimes it can also be defined set of data is simply called a distribution

For a sampling of continuous data, in general, a frequency distribution is constructed by classifying the observations.

*A frequency distribution is constructed for three main reasons*

1. To facilitate the analysis of data;

2. To estimate frequencies of the unknown population distribution from the distribution of sample data;
3. To facilitate the computation of various statistical measures.

## Kind of Frequency Distribution

Frequency distribution can be classified into three types which area based on the methods of arranging data in the table. The three types of frequency distributions are named as

(a) Series of individual observations;

(b) Discrete series and;

(c) Continuous series.

**a) Series of individual observations**

The individual observations are a series, where items are listed singly after observation, as distinguished from listing them in groups. If height of 10 patients are given individually it will form a series of individual observation

**Example: 1**

| Patient Number | Height (Inches) |
|---|---|
| 1 | 62" |
| 2 | 65" |
| 3 | 68" |
| 4 | 64" |
| 5 | 60" |
| 6 | 63" |
| 7 | 67" |
| 8 | 69" |
| 9 | 70" |
| 10 | 58" |

This individual series data has to be arranged in either ascending or descending order for some statistical calculation purpose

**Example: 2**

B P of 10 Patients recorded in a hospital before including them in a clinical trial

| Sl. No.: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BP | 125 | 135 | 140 | 120 | 110 | 115 | 160 | 190 | 170 | 160 |

**b) Discrete series:** In case of discrete series data's are presented in a way that exact measurements of items or subjects are clearly mentioned and there will be a definite difference between the variables of different group of items.

Here each group is distinct and separate from other classes. There will be no continuity from one class to another

**Example**

| Disease Type: | Asthma | Heart Put | Diabetic | Cancer, | Arthritis |
|---|---|---|---|---|---|
| No. of Patients | 25 | 70 | 120 | 30 | 100 |

(in Village with Particular diseases)

**c) Continuous Series**

Continuous series is one where measurements are taken in approximations and the data's are expressed in the form of class interval. In this case the variable can take any intermediate value between the lowest and highest value in the distribution. In case of continuous series the class intervals theoretically continues from the beginning of the frequency distribution to the end without having any break. Always it can be distinguished from the discrete frequency distribution of series, because here it contains two limits upper and lower limit of each class interval.

**Example 1**

| Age of Diabetic Patients | Number of Patients |
|---|---|
| 0 – 10 | 3 |
| 10 - 20 | 20 |
| 20 - 30 | 70 |
| 30- 40 | 100 |
| 40 - 50 | 200 |
| 50 - 60 | 75 |
| 60 - 70 | 12 |

**Example 2**

| Weight interval (lb) | Frequency |
|---|---|
| 10 - 20 | 5 |
| 20 - 30 | 19 |
| 30- 40 | 10 |
| 40 - 50 | 13 |
| 50 - 60 | 4 |
| 60 - 70 | 4 |
| 70 - 80 | 2 |
| **Net** | **57** |

The number of intervals chosen should result in a table that considerably improved the readability of the data. The following rules of them are useful to select or make the intervals for a frequency distribution table

1. Form the intervals, such that the intervals will have significance in relation to the nature of the data
2. While forming intervals, we should not have too many empty intervals or intervals without any frequency
3. Maximum number of intervals in a distribution can be taken up to eight to twenty

The width of all intervals in general should be the same and it helps the reader to read the data easily and allows for simple computation of statistical data or values.

**Some Definition with Regard to Construction of Series**

**RANGE:** The range of a frequency distribution may be defined as the difference between the lower limit of the first interval and the upper limit of last class interval. In the above example the range is 80-10 = 70

**CLASS – INTERVAL:** The class interval may be defined as the size of grouping of data. According to the example which is written above 10-20, 20-30, 30-40 - - - - 70-80, are class – intervals.

**CLASS LIMITS**: The class limits of frequency distribution are defined as the upper and lower limits of each class interval. Since each class interval contains all possible values ranging from the lower limit of the given class interval and infinitely approaching the lower limit of next higher class interval. In general lower limit one class interval and the upper limit of the next succeeding class interval as class limits. In the example mentioned above 10, 20, 30 40, 60, 70 are lower limits and 20, 30, 40, 50, 60, 70, 80 upper limits of class intervals. Together they are called as class limits. Lower limits are denoted as l1 and upper limits are written as l2.

**MAGNITUDE OF A CLASS INTERVAL:** The difference between the upper and lower limit of each class interval is called as magnitude of that class. Here in the above example 20-10 = 10 is the magnitude of the first class – interval.

**MID-VALUE or MID-POINT:** The central point of each interval is termed as mid-value or central value of an interval. Mid-value is written as M and it is obtained by using the equation

$$M = (l1 + l2)/2.$$

**FREQUENCY:** The number of items or observations which are falling within a particular class interval is termed as frequency of that class/class interval.

**Methods of forming class Intervals:**

There are two methods to form class intervals

1. Exclusive method
2. Inclusive method

**Exclusive method:** In case exclusive method, the upper limit of one class interval is the lower limit of the next class interval.

| **Exclusive Method** | | |
|---|---|---|
| **Class-Interval** | | **Frequency** |
| 0 | 10 | 5 |
| 10 | 20 | 12 |
| 20 | 30 | 18 |
| 30 | 40 | 22 |
| 40 | 50 | 40 |
| 50 | 60 | 19 |
| 60 | 70 | 13 |
| 70 | 80 | 8 |
| 80 | 90 | 4 |
| **Total** | | **141** |

**Inclusive method:** In case inclusive method the ambiguity about items identical to a limit of the class interval is sought to be removed. Here upper limit need not be the lower limit of next interval and upper limit will be included in that interval.

| **Inclusive Method** | | |
|---|---|---|
| Class-Interval | | Frequency |
| 0 | 09 | 5 |
| 10 | 19 | 12 |
| 20 | 29 | 18 |
| 30 | 39 | 22 |
| 40 | 49 | 40 |
| 50 | 59 | 19 |
| 60 | 69 | 13 |
| 70 | 79 | 8 |
| 80 | 89 | 4 |
| **Total** | | **141** |

## Cumulative Series

Cumulative frequency distribution has a running total of the values. It is constructed by adding the frequencies of the first class interval to the frequency in the second class interval, the same totals is added to frequencies in the third class interval. This process continues until the final total appearing opposite the last class interval, that total will be equal to the total frequency of the frequency distribution. The cumulative frequency may be classified as downward and Upward.

The downward accumulation results in a list presenting the number of frequencies "less than" any given amount as revealed by the lower limit of succeeding class interval; and the upward accumulation results in presenting the number of frequencies "more than" and given amount as revealed by the upper limit of a preceding class interval.

**DOWNWARD CUMULATION**

| Less Than | Frequency | Cf |
|---|---|---|
| 10 | 5 | 5 |
| 20 | 12 | 17 |
| 30 | 18 | 35 |
| 40 | 22 | 57 |
| 50 | 40 | 97 |
| 60 | 19 | 116 |
| 70 | 13 | 129 |
| 80 | 8 | 137 |
| 90 | 4 | 141 |
| | 141 | |

**UPWARD CUMULATION**

| More Than | Frequency | Cf |
|---|---|---|
| 0 | 5 | 141 |
| 10 | 12 | 136 |
| 20 | 18 | 124 |
| 30 | 22 | 106 |
| 40 | 40 | 84 |
| 50 | 19 | 44 |
| 60 | 13 | 25 |
| 70 | 8 | 12 |
| 80 | 4 | 4 |
| | 141 | |

CHAPTER

# 2 Measure of Central Tendency (Statistical Averages)

## MEANING AND IMPORTANCE

An average reduces the whole distribution or large number of observation to a single figure or value. The average can be defined as **"Measures of central tendency"** because they describe the tendency of items to group around the Middle in a frequency distribution of numerical values or numbers.

This tendency of items to group around the middle is a characteristic which tends itself to measurement and the measurement of that tendency is called average. The averages plays an very important role in Biostatistics. Many statistical techniques which are applied in statistical analysis depend upon the average or central tendency. That may be the main reason for calling statistics as science of average

**Ex:**

1. The average B P of the population can be called as central tendency
2. Average weight of human being can also be taken as central tendency

## OBJECTS AND FUNCTIONS OF AVERAGES

An average is the precise and a simple indicator of the central tendency of whole distribution. The main functions of an average are

1. **To present the salient features of a mass of complex data: W**ith the help of an average, it will be more convenient to express the data or information in a very abbreviated numerical form, in such a way that the salient features of the data collected in table are clearly brought out.

   **Ex**: when we measure the average height of the sample of a particular population, that single figure enable one to draw a general conclusion about characteristics of the phenomena under study.

   The purpose of an average is to represent a group of individual values in a simple and concise manner.