CHAPTER 1

STATISTICS: AN INTRODUCTION

(1) STATISTICS:

Statistics deals with the methods of classification and analysis of data (numerical and non-numerical) for drawing valid conclusions and making reasonable decisions. It has meaningful applications in production engineering, in analysis of experimental data, in economics, in law, in medicines, in biology, etc. The importance of statistical methods whether it be in engineering, in social sciences, in biological sciences, in medical sciences, in health sciences, or, in physical sciences, is on the increase. As such we shall now study this interesting and important field and its applications. Depending on how data are used, the two major areas of statistics are descriptive statistics and inferential statistics.

(a) DESCRIPTIVE STATISTICS:

It consists of the collection, organization, summarization, and presentation of data. (It describes the situation as it is).

(b) INFERENTIAL STATISTICS:

It consists of making inferences from samples to populations, hypothesis testing, determining relationships among variables, and making predictions. (Inferential statistics is based on probability theory. It goes beyond what is known).

(**NOTE:** By probability, we mean the chance of an event occurring. For example, people who play cards, dice, bingo, and lotteries are using the concepts of probability theory. It is also used in the insurance industry and other areas such as genetics, etc.).

*** *** ***

(2) SOME USEFUL TERMINOLOGIES:

We shall now explain certain terminologies which will be useful in the study of various statistical techniques and their applications.

(A) In order to gain information about seemingly haphazard events, statisticians study random variables. These are defined as follows:

(i) VARIABLES:

A variable is a characteristic or an attribute that can assume different values. Height, weight, temperature, number of phone calls received, etc. are examples of variables.

(ii) RANDOM VARIABLES:

Variables whose values are determined by chance are called random variables.

(B) COLLECTION OF DATA:

The collection of data constitutes the starting point of any statistical investigation. It should be conducted systematically with a definite aim in view and with as much accuracy as is desired in the final results, for detailed analysis would not compensate for the bias and inaccuracies in the original data. The definition of data is given below.

(i) DATA:

The measurements or observations (values) for a variable are called data.

(ii) DATA SET:

A collection of data values forms a data set.

(iii) DATA VALUE OR DATUM:

Each value in the data set is called a data value or a datum**.**

Example: Suppose a researcher selects a specific day and records the number of calls received by a local office of the Internal Revenue Service each hour as follows: {8, 10, 12, 12, 15, 11, 13, 6}, where 8 is the number of calls received during the first hour, 10 the number of calls received during the second hour, and so on. The collection of these numbers is an example of a data set, and each number in the data set is a data value**.**

(C) Data may be collected for each and every unit of the whole lot (called population), for it would ensure greater accuracy. But, however, since in most cases the populations under study are usually very large, and it would be difficult and time-consuming to use all members, therefore statisticians use subgroups called samples to get the necessary data for their studies. The conclusions drawn on the basis of this sample are taken to hold for the population. The definitions of a population and a sample are given below.

(i) POPULATION:

A population is the totality of all subjects possessing certain common characteristics that are being studied.

(ii) SAMPLE:

A sample is defined as a subgroup or subset of the population.

(iii) RANDOM SAMPLE:

A sample obtained without bias or showing preferences in selecting items of the population is called a random sample.

(D) CLASSIFICATION OF VARIABLES (AND DATA):

(a) Random Variables (or Data) can be classified as qualitative or quantitative as follows:

(i) QUALITATIVE VARIABLES (OR DATA):

Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable "gender" is qualitative. Other examples of qualitative variables are religious preferences, geographic locations, grades of a student, etc.

(ii) QUANTITATIVE VARIABLES (OR DATA):

Quantitative variables are numerical in nature and can be ordered or ranked. For example, the variable "age" is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, body temperatures, etc.
(b) Quantitative random variables (or data) can be further classified either as discrete or continuous, depending on the values it can assume. These are defined as follows:

(i) DISCRETE VARIABLES (OR DATA):

Discrete variables assume values that can be counted (such as, 0, 1, 2, 3, etc.). They are obtained by counting. Examples of discrete variables are the number of children in a family, the number of students in a class-room, the number of calls received by a switchboard operator each day for one month, etc.

(ii) CONTINUOUS VARIABLES (OR DATA):

Continuous variables can assume all values between any two specific values. They are obtained by measuring. For Example, "temperature" is a continuous variable, since the variable can assume all values between any two given temperatures. Other examples of continuous variables are height, weight, length, time, etc.

(3) RECORDED VALUES OF A CONTINUOUS RANDOM VARIABLE AND ITS BOUNDARIES:

Since continuous data must be measured, rounding answers is necessary because of the limits of the measuring device. Usually, answers are rounded to the nearest given unit. For example, heights must be rounded to the nearest inch, weights to the nearest ounce, etc. Hence, a recorded height of 73 inches would mean any measure of 72.5 inches up to but not including 73.5 inches. Thus, the boundary of this measure is given as 72.5 – 73.5 inches. (We have taken 72.5 as one of the boundaries since it could be rounded to 73. But, we can not include 73.5 because it would be 74 when rounded). Sometimes 72.5 – 73.5 is called a class which will contain the recorded height of 73 inches. The concept of the boundaries of a continuous variable is illustrated in the following Table I:

TABLE I

Variable	Recorded Value	Boundaries (Class)
Length	15 cm	14.5 – 15.5 cm
Temperature	86^0 F	85.5 – 86.5^0 F
Time	0.43 sec	0.425 – 0.435 sec
Weight	1.6 gm	1.55 – 1.65 gm

Note: The boundaries of a continuous variable in the above table are given in one additional decimal place and always end with the digit 5. The concept of the class (or boundaries) of a continuous variable will be discussed again in Chapter2.

(4) MEASUREMENT SCALES OF A DATA: Data can also be measured by various scales. The four basic levels of measurements are nominal, ordinal, interval, and ratio. These are described below:

TABLE II

MEASUREMENT SCALES (DEFINITIONS AND EXAMPLES)

Nominal-level Data	Ordinal-Level Data	Interval-level Data	Ratio-level Data
Definition: The nominal-level of measurement classifies data into mutually exclusive (non-overlapping), exhaustive categories in which no ordering or ranking can be imposed on the data**.**	**Definition:** The ordinal-level of measurement classifies data into categories that can be ordered or ranked. However, precise differences between the ranks do not exist. (For example, when people are classified according to their build (small, medium, or large), or when students are classified according to their grades (A, B, C, or D), a large variation exists among the individuals in each class.	**Definition:** The interval-level of measurement ranks data, and precise differences between units of measure do exist**.** However, there is no meaningful zero (i.e., starting point). For example, many standardized psychological tests yield values measured on an interval scale. There is a meaningful difference of one point between an IQ of 109 and an IQ of 110. There is no true zero (i.e., no starting point) because IQ tests do not measure people who have no intelligence.	**Definition:** The ratio-level of measurement possesses all the characteristics of interval measurement (i.e., data can be ranked, and there exists a true zero or starting point)**.** In addition, true ratios exist between different units of measure. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. In other words, the first person can lift twice as much as the second person.
Examples: (i) Zip Code (ii) Gender (Male, Female) (iii) Eye Color (Blue, Brown, Green, Hazel) (iv) Political Affiliation (v) Religious Affiliation (vi) Major Field of Study (Math., Comp. Sc.) (vii) Nationality (viii) Marital Status	**Examples:** (i) Grade (A, B, C, D, F) (ii) Judging (1st place, 2nd place, etc.) (iii) Rating Scale (Poor, Good, Excellent) (iv) Ranking of Tennis Players	**Examples:** (i) SAT Score (ii) IQ (iii) Temperature	**Examples:** (i) Height (ii) Weight (iii) Time (iv) Salary (v) Age (vi) Number of Phone Calls

(5) BASIC METHODS OF SAMPLING:

When the population is large and diverse, a sampling method must be designed so that the sample is representative, unbiased and random, i.e. every subject (or element) in the population has an equal chance of being selected for the sample. The following sampling methods are commonly used for obtaining a random sample.

TABLE III

Random Sampling	**Stratified Sampling**	**Systematic Sampling**	**Cluster Sampling**	**Convenience Sampling**
This method requires that each member of the population be identified and assigned a number. Then a set of numbers drawn randomly from this list forms the required random sample. Note that each member of the population has an equal chance of being selected. For a large population, computers are used to generate random numbers which contain series of numbers arranged in random order.	This method requires that the population be classified into a number of smaller homogeneous strata or subgroups. A sample is drawn randomly from each stratum. For example, a population could be stratified by age, sex, marital status, education, religion, occupation, ethnic background or virtually any characteristic.	This method requires that every kth member (or item) of the population be selected to form the required random sample. For example, we might select every 10th house on a city block for the random sample.	The population area is first divided into a number of sections (or subpopulations) called clusters. A few of those clusters are randomly selected, and sampling is carried out only in those clusters. For example, a community can be divided into city blocks as its clusters. Several blocks are then randomly selected. After this, residents on the selected blocks are randomly chosen, providing a sampling of the entire community.	In convenience sampling, we use the results that are readily available.

*** *** ***

(6) STATISTICAL INFERENCE AND MEASUREMENT OF RELIABILITY:

Definition 1: A statistical inference is an estimate or prediction or some other generalization about a population based on information contained in a random sample of the population. That is, the information contained in the random sample is used to learn about the population.

Definition 2: A measure of reliability is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

(7) ELEMENTS OF DESCRIPTIVE AND INFERENTIAL STATISTICAL PROBLEMS:

TABLE IV

FOUR ELEMENTS OF DESCRIPTIVE STATISTICAL PROBLEMS	**FIVE ELEMENTS OF INFERENTIAL STATISTICAL PROBLEMS**
1. The population or sample of interest. **2.** One or more variables (characteristics of the population or sample units) that are to be investigated. **3.** Tables, graphs, numerical summary tools. **4.** Identification of patterns in the data.	**1.** The population of interest. **2.** One or more variables that are to be investigated. **3.** The sample of population units. **4.** The statistical inference about the population based on information contained in the random sample of the population. **5.** A measure of reliability for the statistical inference.

REFERENCES: