**CHAPTER 1**

1. **INTRODUCTION**

**1.1 Definition and Classifications of statistics**

**Definition of Statistics:**

* Statistics is a collection of numerical facts and data.
* Statistics is a mathematical science dealing with the methods of collection, organizing the collected data, presentation, analysis and interpretation of the data.

We can define statistics in two ways.

1. **Plural sense (lay man definition).**

* It is an aggregate or collection of numerical facts.
* Statistics is defined as aggregates of numerical expressed facts (figures) collected in a systematic manner for a predetermined purpose

1. **Singular sense (formal definition):** as a science or a field of study or a body of knowledge.

Statistics is defined as the science of collecting, organizing, presenting, analyzing and interpreting numerical data for the purpose of assisting in making a more effective decision about population based on sample data.

* Now days, statistics is used almost in every field of study, such as natural science, social science engineering, medicine, agriculture, e t c. In sports, for example, a statistician may keep records of the number of yards a running back gains during a football. In education, a researcher might want to know if new methods of teaching are better than old ones. These are only a few examples of how statistics can be used in various occupations.
* Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.

**Classifications of Statistics**

Depending on how data can be used statistics is sometimes divided in to two main areas or branches.

1. **Descriptive Statistics**: is a statistical method which deals with the collection, organization and summarization of data by using graphs, charts and tables.

* deals with describing data without attempting to infer anything that goes beyond the given set of data,
* Consists of collection, organization, summarization and presentation of data.

Example: The average age of students in this class is 21.

1. **Inferential Statistics**: is a method used to generalize from a sample to a population.

* deals with making inferences and/or conclusions about a population based on data obtained from a limited sample of observations,
* Consists of performing hypothesis testing, determining relationships among variables and making predictions.

Examples:

1. The average age of a student in HU University is 20.1 years.
2. From past figures, it has been predicted that 31 of registered voters will vote in the November election.
3. To determine the most effective dose of a new me dication (on the basis of tests performed with volunteer patients from selected hospitals)
4. To compare the effectiveness of two reducing diets (based on the weights losses of persons who were taking the diets)
5. There is a relationship between smoking tobacco and an increased risk of developing cancer.
   1. **Stages in Statistical Investigation**

There are five stages or steps in any statistical investigation.

1. **Collection of data**: the process of measuring, gathering, assembling the raw data up on which the statistical investigation is to be based.
   * Data can be collected in a variety of ways; one of the most common methods is through the use of survey. Survey can also be done in different methods, three of the most common methods are:
     + Telephone survey
     + Mailed questionnaire
     + Personal interview.

**Exercise:** discuss the advantage and disadvantage of the above three methods with respect to each other.

1. **Organization of data**: Summarization of data in some meaningful way, e.g. table form
2. **Presentation of the data**: The process of re-organization, classification, compilation, and summarization of data to present it in a meaningful form.

Example: charts, graphs and tables.

1. **Analysis of data**: The process of extracting relevant information from the summarized data, mainly through the use of elementary mathematical operation.
2. **Inference of data**: It is a process of making interpretations or conclusions from sample data for the totality of the population.

* Statistical techniques based on probability theory are required.
  1. **Definitions of some terms**

1. **Population**: It is the collection of all possible observations of a specified characteristic of interest (possessing certain common property) and being under study.

Example: All of the students in HU those who take stat 2011 course.

All staff members of HU.

1. **Sample**: It is a subset of the population, selected using some sampling technique in such a way that they represent the population. A sample should be a representative of the population.
2. **Sampling:** The process or method of sample selection from the population.
3. **Sample size:** The number of elements or observation to be included in the sample.
4. **Census:** Complete enumeration or observation of the elements of the population. Or it is the collection of data from every element in a population.
5. **Parameter:** Characteristic or measure obtained from a population.

Examples: Average, Range, proportion, variance

1. **Statistic:** Characteristic or measure obtained from a sample.
2. **Variable** is a characteristic of an object that can have different possible values.

Example: a researcher wants to study the academic performance of fist year student in HU. But for several constraints he cannot enumerate the whole students. So he took randomly 500 students and obtained the average GPA to be 2.58.

* 1. Identify the population? b. Identify the sample? c. Identify the statistic?
  2. **Applications, Uses and Limitations of statistics**

**Applications of statistics:**

Statistics can be applied in any field of study which seeks quantitative evidence. For instance (in engineering)

* To compare the breaking strength of two types of materials
* To determine the probability of reliability of a product.
* To control the quality of products in a given production process.
* To compare the improvement f yield due to certain additives (fertilizer, herbicides, (wee decides), e t c

to see some applications

1. **agriculture:** which types of varieties gives the best yield

which feed mixtures should chickens be fed so that they will gain most weight

1. **medical and pharmaceutical research :**

**Uses of statistics:**

The main function of statistics is to enlarge our knowledge of complex phenomena. The following are some uses of statistics:

1. It presents facts in a definite and precise form.
2. Data reduction.
3. Measuring the magnitude of variations in data.
4. Furnishes a technique of comparison
5. Estimating unknown population characteristics.
6. Testing and formulating of hypothesis.
7. Studying the relationship between two or more variable.
8. Forecasting future events.

**Limitations of statistics**

As a science statistics has its own limitations. The following are some of the limitations:

* Deals with only quantitative information.
* Deals with only aggregate of facts and not with individual data items.
* Statistical data are only approximately and not mathematical correct.
* Statistics can be easily misused and therefore should be used be experts.
  1. **Types of variables and Scales of measurement**
     1. **Types of Variable**

There are two types of variables.

1. **Quantitative variables:** are variables which are numerical in nature and can be measured and counted. Example: height, weight, no of students, GPA etc.
2. **Qualitative variables:** are variables that cannot be quantified directly.

Are variables that can be placed into distinct category according to some characteristics. Examples: colour, beauty, sex, location.

**Quantitative variables can be further classified as**

* + Discrete variables, and
  + Continuous variables

1. **Discrete variables** are variables whose values are determined by counting.

*Examples:* number of students, number of households (family size), Number of pages of a book.

1. **Continuous variables** are variables whose values are determined by measuring rather than counting.

*Examples:* weight, Length, Volume, e t c.

* + 1. **Scales of Measurement**

**There are four measurement scales**

1. **Nominal scale: -** “Nominal “is a Latin word for “name”

* It is a scale of measurement grouping individuals into different categories in which no order or ranking can be imposed on the data.
* No arithmetic and relational operation can be applied.
* One is different from the other

Examples:

1. Political party preference (Republican, Democrat, or Other,)
2. Sex (Male or Female.)
3. Marital status(married, single, widow, divorce)
4. Blood type (A,B,AB,O)
5. Regional differentiation of Ethiopia.
6. **Ordinal scale: -** “ ordinal” is a Latin word, meaning “order”
   * It is a scale for grouping and ordering of individuals in to different categories.
   * Data consisting of an ordering of ranking of measurements are said to be on an ordinal scale of measurements.
   * Level of measurement which classifies data into categories that can be ranked.
   * Differences between the ranks do not exist.
   * Arithmetic operations are not applicable but relational operations are

applicable.

* + Ordering is the sole property of ordinal scale.
  + One is different from and grater /better/ less than the other

Examples:

#### Letter grades (A, B, C, D, F).

#### Rating scales (Excellent, Very good, Good, Fair, poor).

#### Military status

#### Economic status (low, medium, high)

#### Ordinal scales data contain and convey more information than the nominal scale data, for relative magnitudes are known, however, quantitative comparisons are impossible.

1. **Interval scale:**  is a measurement scale in which classifies data that can be ranked and

differences are meaningful.

* There is no true zero point (arbitrary zero point) zero does not shows a total absence of the quantity being measured.
* In this measurement scale one is different, better/greater and by a certain amount of difference than another (Possible to add and subtract but multiplication and division are not possible)

37Oc – 35oc = 2oc

45oc – 43 oc= 2oc

40oc = 2(20oc) But this does not imply that an object which is 40 oc is twice as hot as an object which is 20 oc (oF = 9/5, oc +32)

40 oc → 9/5 x 40 oc + 32 = 104 oF

20oc → 9/5 x20 oc + 32 = 68 oF

Oc = 5/9 (oF- 32)

60 oF→ 5/9 (60 – 32) = 15.56 oc

30 oF→ 5/9 (30 - 32) = -1.11 oc

*Example:*

* IQ
* Temperature in oF.
* Interval scale data convey better information than nominal and ordinal scale data.

1. **Ratio scale:** is a measurement scale in which

* There exists a zero point on the measurement scale and zero shows a total absence of the quantity being measured .
* One is different, larger /taller/ better/ less by a certain amount of difference and so much times than the other.
* (+, -, \*, / Are possible on this scale)
* This measurement scale provides better information than interval scale of measurement

*Examples:*

* Weight
* Height
* Number of students
* Age