1. Introduction

1.1 Nature of Statistics

What is statistics? Consider the following examples:

- A pollster interested in the outcome of an upcoming election interviews a certain number of voters and makes a prediction as to who will probably win the election.
- To get an idea of the economic status of a town, an economist obtains the salaries of all workers in the town and then computes the average.
- To estimate the average age of students at UNZA, a sociology student selects 1000 students, computes their average age and then uses this value as an estimate.

All these examples make use of data or observations. An observation might be numerical, such as a person's age or non-numerical such as a person's gender (male or female).

1.1.1 Definition (Defn)

Statistics is the science of collecting, simplifying and describing data as well as making inferences (drawing conclusions) based on the analysis of data.

There are two branches of statistics namely **descriptive** and **inferential** statistics.

Descriptive statistics deals with simplifying and giving properties of data. The objective of descriptive statistics is to organize, summarize or describe the data to make it more comprehensible e.g. A list of the salaries of workers in Lusaka would be incomprehensible because of its length but the average of these salaries would give an understanding of the economic status of Lusaka residents.

Inferential statistics involves drawing conclusions based on the analysis of data e.g. the sociology student who estimates the average age of UNZA students based on a knowledge of the average age of some of these students uses inferential statistics.

1.1.2 Definition

The entire collection of all the elements we are interested in is called a **population**. (These elements might be people, cars, data values, etc).
A collection of some of these elements obtained from the population is called a **sample**.

Note

Inferential statistics can be defined in terms of population and sample.

1.1.3 Defn

Inferential statistics is concerned with making judgements (or inferences) about a population based on the properties of some sample obtained from the population.

1.1.4 Defn

A numerical property of a population is called a **parameter**. A numerical property of a sample is called a **statistic**. (Numerical property means a property that can be expressed as a number.) e.g.
Problem: estimate the age of UNZA students.

Population: all UNZA students (specifically ages of all UNZA students).
Parameter: the average age of all UNZA students.
Statistic: the average of the ages for a sample of UNZA students.

1.2 Sampling Methods

In order to use samples to make judgements about a population, we need samples that are representative of the population (i.e. that resemble the population).

We now consider some of the methods of obtaining samples.

1.2.1 Defn

A **random sample** is a sample obtained from the population in such a manner that all samples of the same size have equal likelihood of being selected.
Any method of obtaining random samples is called **random sampling**.

Note

1. One method of random sampling is the lottery method. How does it work?
   - Elements in the population are identified by a name or number written on a tag.
   - The tags are placed in a container and are then will mixed.
   - A tag is drawn by chance from the container and this process is repeated until the desired number of tags is obtained.

   Another method of random sampling is the random number method. How does it work?
   - Elements in the population are identified by numbers.
   - A random number table or calculator or computer (generates random numbers) is used to obtain a random sample.

2. When the sampling is done in such a way that there are no repetitions in the sample (i.e. by ignoring numbers that repeat), the resulting sample is called a **simple random sample** (SRS).

1.2.2 Defn

If the population is divided into subpopulations called strata and we take a random sample from **each** stratum, the resulting sample is called a **stratified sample.**
e.g. Zambian voters could be divided into strata according to provinces: western, eastern, southern, northern, etc. A stratified sample is obtained by taking a random sample from each province (all the samples from the provinces put together constitute a stratified sample).

Note

1. The size of the sample from each stratum is often proportional to the size of the stratum in the population. The resulting sample is called a **proportional stratified sample** e.g. suppose 25% of Zambian voters are from Southern province, 20% from Eastern province, 10% from Western province, etc. To obtain a stratified sample of

size 1000, we would sample 250 voters from Southern, 200 from Eastern, 100 from Western, etc.
2. Stratified samples are often easier to obtain than true random samples.

Even though a population can be divided into strata, it may not be convenient to sample from each stratum.

1.2.3 Defn

A cluster sample is a sample obtained by selecting **some** of the strata and then sampling from each of these.

Note

1. In stratified sampling, the strata are frequently quite different from one another but are homogeneous internally. In cluster sampling, however, the strata are often defined in such a way that the clusters are as internally diverse as possible yet easy to access.
2. In practice, most large surveys use multistage cluster sampling. For example, to obtain a multistage cluster sample of Zambian families, we could start with a list of all provinces in Zambia. These are usually called primary sampling units (PSUs). The procedure is as follows:
   - Take a sample of PSUs.
   - Take a sample of districts in each sampled PSU.
   - Take a sample of streets in each selected district.
   - Take a sample of families on each selected street.

   At each stage, simple random sampling could be used.