LESSON TWO: DATA COLLECTION, ORGANIZATION AND PRESENTATION

2.1 Introduction

- Data refers to any information or facts collected for reference or analysis.
- There are two types of data: secondary data and primary data.

Secondary Data

- Its data that been gathered earlier for some other purpose. In contrast, the data that are collected first hand by someone specifically for the purpose of facilitating the study are known as primary data.
- E.G: the demographic statistics collected every ten years are the primary data with the registrar of persons but the same statistics used by anyone else would be secondary data with that individual.

Advantages of secondary data

i) It is far more economical as the cost of collecting original data is saved.

ii) Use of secondary data is time saving.

Disadvantages of secondary data

i) One does not always know how accurate the secondary data are.

ii) The secondary data might be out dated.

- Before using secondary data it is important to consider the following:

i) **Whether the data are suitable for the purpose of investigation**

- ✓ The suitability of the data can be judged in the light of the nature and scope of investigation.
- ✓ E.G: if the object of inquiry is to study the wage levels including allowances of workers and the data relate to basic wages alone, such data would not be suitable for the immediate purpose

ii) **Whether the data are adequate for the purpose of investigation**

- ✓ Adequacy of the data is to be judged in the light of the requirements of the study and the geographical area covered by the available.
- ✓ E.G: if the object is to study wage rates of the workers in the sugar industry in Kenya and if the available data cover only one region, it would not serve the purpose.

- ✓ The question of adequacy may also be considered in the in the light of the time period for which the data are available
- ✓ E.G: For studying trend of prices data for the last 8-10 years may be required but if from the sources known the data available is for the last 5-6 years only, this would not serve the object.

iii) **Whether the data are reliable**

- ✓ Reliability of the data has to do with the data collection procedures.
- ✓ To ensure reliability of the data one may need to determine the context in which the data were collected, the procedure followed and the level of accuracy exercised in the collection.
- ✓ Determination of the reliability of secondary data is perhaps the most important and at the same time most difficult job.

Primary Data

- Primary data are measurements observed and recorded as part of an original study.
- The work of collecting primary data is usually limited by time, money and manpower available.
- When the data to be collected are very large in volume, it is possible to draw reasonably accurate conclusions from a sample.
- There are two methods of obtaining primary data:
    a) Questioning
    b) Observation
- Questions may be asked in person or in writing. A formal list of such questions is called a **questionnaire.**
- When the data are collected by observation, the investigator asks no questions. Instead, he observes and records the desired information.
- Of the two methods named above, the questionnaire method is more widely used for calculating business data. Three different ways of communicating with questionnaires are available
    i) Personal interview
    ii) Mail
    iii) Telephone interview

- Personal interviews are those in which an interviewer obtains information from respondents in face-to-face meetings. The information obtained by this method is likely to be more accurate because the interviewer can clear-up doubts, can cross-examine the informants and thereby obtain correct information.
- In mail surveys, questionnaires are mailed to respondents who are supposed to fill them and return. They are appropriate where the field of investigation is very vast and the informants are spread over a wide geographical area.
- Telephone interviews are similar to personal interviews except that communication between interviewer and respondents is on telephone instead of direct personal contact.

2.3 Organization and Presentation of Data

- Data collected in an investigation and not organized systematically is called **raw data**. The arrangement of this data in ascending or descending order of magnitude is called an **array.**
- The difference between the largest and the smallest value is called the **range.**
- E.G: The table below records the heights, in inches, of eight students. Column I represents the raw data and column II illustrates the arrangement in an array.

Raw Data	Array
66	65
68	66
72	66
65	68
66	68
73	69
68	72
69	73

- The largest value is 73 and the smallest is 65. Hence, the range is 73 – 65 = 8 inches.

Frequency Distribution

Ungrouped data

- In forming an array a value is repeated as many times as it appears. The number of times a value appears in the listing is referred to as its **frequency.** In giving the frequency of a value, we answer the question, " How frequently does the value occur in the listing?"
- When the data is arranged in tabular form by giving its frequencies, the table is called a **frequency table.** The arrangement itself is called a **frequency distribution.**
- Quite often it is useful to give relative frequencies instead of actual frequencies. The **relative frequency** of any observation is obtained by dividing the actual frequency of the observation by the total frequency (sum of all frequencies).
- If the relative frequencies are multiplied by 100 and expressed as a percentage, we get the **percentage frequency distribution.**
- An advantage of expressing frequencies as percentages is that one can then compare frequency distributions of two sets of data.

  Example:

  The following data were obtained when a die was tossed 30 times. Construct a frequency table.

1	2	4	2	2	6	3	5	6	3
3	1	3	1	3	4	5	3	5	3
5	1	6	3	1	2	4	2	4	4

Grouped Data

- When dealing with a huge mass of data and when the observed values consist of too many distinct values, it is preferable to divide the entire range of values and group the data into classes.
- E.G: If we are interested in the distribution of ages of people, we could form the classes 0 – 19, 20 – 39, 40 – 59, 60 – 79 and 80 – 99. A class such as 40 – 59 represents all the people with ages between 40 and 59 years inclusive.
- When data are arranged in this way, they are called **grouped data.** The number of individuals in a class is called the **class frequency.**
- The following set of steps are suggested to form a frequency distribution from the raw data

i) **Range**

Scan through the raw data and find the smallest and the largest value. The largest value minus the smallest value gives the range.

ii) **Number of classes**

Decide on a suitable number of classes. This could be anywhere from six to twenty.

iii) **Class size**

Divide the range by the number of classes. Round this figure to a convenient value to obtain the class size and form the classes.

iv) **Frequency**

Find the number of observations in each class.

Example

The following data gives the amounts (in dollars) spent on groceries by 40 housewives during a week.

22	12	9	8	33	32	30	33	8	11
21	16	12	15	37	30	16	22	12	24
18	25	37	16	25	28	25	18	9	28
25	28	26	15	12	35	38	16	24	31

Construct a frequency distribution using seven classes.

Class Intervals, Class Marks and Class Boundaries

- The blocks 10 – 20, 20 – 30, 30 – 40, etc are called **class intervals.** The lower ends of the class intervals are called **lower limits** and their upper ends are called **upper limits.**
- The number of values specified in a given interval is called its **length** or **width** or **magnitude.**

  E.G: The class 1 – 3 has values 1, 2, 3 thus its length is 3.

  The class 5 – 9 has values 5, 6, 7, 8, 9; the length or magnitude is 5
- There are two types of classes
  - i) Inclusive type: These are of the type 5 – 9, 10 – 14, 15 – 19, … where both the upper and lower class limits are included in a given class.
  - ii) Exclusive type: These are of the type 5 – 10, 10 – 15, 15 – 20, … where the upper class limit of a given class is the lower class limit of the succeeding class.

The class 5 – 10 has values 5, 6, 7, 8, 9 and the class 10 – 15 has 10, 11, 12, 13, 14.

NB: The conversion of inclusive type of classes to exclusive type is useful in calculating certain measures such as mode and median.

- A point that represents the halfway or dividing point between successful classes is called a **class boundary**. If d is the difference between the lower class limit of a given class and the upper class limit of the succeeding class, then

  Upper Class Boundary (UCB) = Upper Class Limit (UCL) + $\frac{1}{2}d$

  Lower Class Boundary (LCB) = Lower Class Limit (LCL) - $\frac{1}{2}d$

- The **class mark** is defined as the mid point of a class interval. It is computed by adding the lower and upper class limits of a class and then dividing by 2.

  Mid point = $\frac{1}{2}(UCB + LCB)$

  = $\frac{1}{2}(UCB - LCB)$

Example

Class	L.C.L	U.C.L	L.B	U.B	Class mark (Midpoint)
10 – 19	10	19	9.5	19.5	14.5
20 – 29	20	29	19.5	29.5	24.5
30 – 39	30	39	29.5	39.5	34.5
40 – 49	30	49	39.5	49.5	44.5
50 – 59	50	59	49.5	59.5	54.5

NB: The upper boundary of one class is the lower boundary of the next.

Cumulative Frequency Distribution:

- If a frequency distribution is arranged in the "less than" form, it is called a **cumulative frequency distribution** which presents the accumulated values.
- When the data is not grouped, a cumulative frequency distribution will show the number of items less than or equal to a given value.

Example

The data below gives the weights of 30 people. Find the cumulative frequency distribution.

Weight	Frequency	Cumulative frequency (c.f)
140	3	3
150	5	8
160	6	14
170	7	21
180	6	27
190	3	30

- When the data is grouped, the cumulative frequency distribution gives the total frequency of all the values less than the upper boundary of a given class.

Example

Find the cumulative frequency distribution for the grouped data given below:

Class	Frequency	Cumulative frequency (cf)
5 – 19	4	4
20 – 34	12	16
35 – 49	15	31
50 – 64	16	47
65 – 79	22	69
80 – 94	11	80

2.4 Graphical Representation of a Frequency Distribution

The following types of graphical representation are usually used for frequency distribution.

a) Histogram: It is a graph in which classes boundaries are marked on the horizontal axis and the class frequencies on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars and the bars are drawn adjacent to each other.

b) Frequency polygons and Frequency Curve: A frequency polygon is a line graph where we plot the class marks or midpoints along the horizontal axis and the corresponding frequencies along the vertical axis. The class midpoints are connected with a line segment.

If the classes are very many and the class widths are so small that the midpoints are close together, the polygon can be formed by free hand to give a smooth curve known as a frequency curve.

c) Cumulative Frequency Curve or the Ogive. An ogive is a line graph obtained by representing the upper class boundaries along the horizontal axis and the corresponding cumulative frequency along the vertical axis.

2.5 Exercise

A random sample of 50 auto drivers insured with a company and having similar auto insurance policies was selected. The following data shows monthly auto insurance premium (in Kshs.000) paid by them.

54	40	45	20	60	30	35	40	55	70	20	15
45	60	45	25	15	30	25	18	35	25	45	56
59	25	27	39	50	56	20	25	30	30	41	25
56	48	45	25	35	60	55	48	38	34	60	60
60	64										

i) Group the above data starting with the class 10 -20 exclusive

ii) Represent the data using a Histogram and an Ogive.