# Chapter One

# 1.0 SCOPE FOR STATISTICAL METHODS

## 1.1Definitions and Meaning of Statistics

The term "Statistics" can be defined in various ways as given by different authors. Some of these definitions are:

1. Statistics is a discipline that deals with the scientific method of collection, organization, summarization, presentation, analysis and interpretation of data.
2. Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.
3. Statistics are the classified facts representing the conditions of the people in a state, specifically those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.
4. Statistics may be defined as the aggregate of facts affected to a marked extent by a multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other.
5. By Statistics, we mean quantitative data affected to a marked extent by a multiplicity of causes.

Generally, the term statistics is used to imply either statistical data or statistical method. When it is used in the sense of statistical data, it refers to its quantitative feature while the other aspect refers to the body of theories and techniques employed to analyze the numerical information in order to make wise decision. It is a branch of the scientific method that is used in dealing with those phenomena which can be described numerically, either by measurements or by code or counts.

Some of the examples of the quantitative feature of statistics are the yearly enrolment of students in Nigerian tertiary institutions, number of in-patients recorded daily in hospital, distribution of clerical workers' income, the ages of registered students in an undergraduate class etc. However, there are also some quantities that are not numerical, but can be made by counting and coding. The attributes or gender of set of respondents to a particular questionnaire are not quantitative nor numbers, but by coding each of the listed qualities as well as the gender (i.e. Male =1 and Female

= 0) or by counting them in numbers, one can associate a numerical description to the varying qualities and sex of all respondents. Examples of statistical data are shown in Tables 1.1 and 1.2 below.

**Table 1.1: Statistics of Students Enrolment in Tertiary Institutions by Gender**

| **Gender** | **Number** | **Percentage** |
|---|---|---|
| Male | 6,000 | 46 |
| Female | 7,000 | 54 |
| Total | 13,000 | 100 |

**Table 1.2: Statistics of Students Enrolment in Tertiary Institutions by Faculties**

| **Faculty** | **Number** | **Percentage** |
|---|---|---|
| Science | 2,500 | 19 |
| Management Studies | 5,000 | 38 |
| Environmental | 1,000 | 8 |
| Engineering | 1500 | 12 |
| Social Sciences | 3,000 | 23 |
| Total | 13,000 | 100 |

**Source: Extracts from Records**

## 1.2 Importance of Statistics

1. Statistical knowledge helps one to use the proper methods to collect data, employ the correct analyses and effectively present the results.
2. Statistics is a crucial process behind how to make discoveries in science, make decisions based on data and make predictions.
3. Statistical knowledge is relevant from ‘Anatomy’ to ‘Zoology’. There is no known field of research which does not requires statistical techniques for accomplishment.

## 1.3 Basic Functions of Statistics

The ideal function of statistics is to enlarge our knowledge of complex phenomena; and to bring precision to our ideas that would otherwise remain vague and indeterminate. Statistics is thus, able to widen our knowledge because of the following services it renders to humanity.

1. Statistics presents facts in a definite form.

2. It simplifies the unwieldy and complex mass of data and presents them in a more suitable manner.
3. Statistics classifies numerical facts into more understanding features.
4. Statistics furnishes varying techniques of comparison.
5. Statistics enlarge individual experience in various ways.
6. Statistics provide guidance in the formulation of policies.
7. Statistics enable measurement of the magnitude of a phenomenon.
8. Statistics endeavours to interpret phenomena under investigation.

## 1.4 Areas of Statistical Application

Statistical methods have become useful tools in all human endeavours. Statistics plays a vital role in every field of human activities. Nowadays statistics holds a central position in almost every field, including industry, commerce, business, physics, chemistry, psychology, economics, Mathematics, biology, botany, engineering, astronomy, etc., thus the application of statistics is very wide. The following are some of the major areas of statistics applications.

1. **Statistics in Business.** The need for statistical information in the smooth functioning of business undertaking increases along with its size. No business, big or small, public or private can flourish in this era of large-scale production and cut-throat competition without the help of statistics. Statistical information is needed from the inception of a new business till the time of its exist. All the factors that are likely to affect judgment in the course of business cycle are quantitatively weighted and statistically analyzed before taking any decisions.
2. **Actuarial Science** is the discipline that applies mathematical and statistical methods to access risk in the insurance and finance industries.
3. **Astrostatistics** is the discipline that applies statistical analysis to the understanding of astronomical data.
4. **Biostatistics** is a branch of biology that studies biological phenomena and observations by means of statistical analysis and it includes medical statistics.
5. **Business Analytics** is a rapidly developing business process that applies statistical method of data sets (often very large) to develop new insight and understanding of business performance & opportunities.

6. **Chemometrics** is the science of relating measurements made in chemistry practical or on a chemical process to the state of the system via application of mathematical or statistical methods.
7. **Econometrics** is the application of statistical methods to the empirical study of economic theories and relationships.
8. **Demography** is the statistical study of populations and its environs, as they change over time.
9. **Environmental statistics** is the application of statistical methods to environmental science, and this include weather, climate air and water quality.
10. **Epidemiology** is the statistical study of factors affecting the health and illness of populations, and serves as the foundation of interventions made in the interest of public health and preventive medicine.
11. **Geostatistics** is a branch of geography that deals with the analysis of data from disciplines such as petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography
12. **Jurimetrics** is the application of probability and statistics to law.
13. **Machine learning** is the subfield of computer science that formulates algorithms for making predictions from data.
14. **Operations research** is an interdisciplinary branch of applied mathematics and scientific technique that uses methods such as mathematical modeling, statistics and algorithms to arrive at optimal solutions to complex problems.
15. **Population ecology** is an aspect of ecology that deals with the dynamics of species populations and how these populations interact with the environment
16. **Quality control** is a statistical technique aimed at studying, controlling and improving the variations in quality of manufactured products.
17. **Psychometrics** is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes and personality traits.
18. **Reliability engineering** is the study of the ability of a system or component to perform its required functions under stated conditions for a specified period of time.

19. **Statistical mechanics** is the application of probability theory, which includes statistical tools for dealing with large populations that is concerned with the motion of particles or objects when subjected to eternal force.
20. **Statistical physics** is one of the fundamental theories of physics, which uses methods of probability theory in solving physical problems.
21. **Statistical thermodynamics** is the study of the microscopic behaviors of thermodynamic systems using probability theory to provide a molecular-level interpretation of thermodynamic quantities such as work, heat, free energy and entropy.

## 1.5 Features of Statistics

The definition of Statistics as numerical statement of facts, makes it quite clear that Statistics should possess the following characteristics.

1. Statistics are aggregate of facts. A single figure relating to production, sales, age, birth, death etc. is not statistics but aggregates or series of such figures would be statistics because of their comparability and relationship.
2. Statistics are affected to a marked extent by a multiplicity of causes. A number of causes affect statistics in relation to a particular field of enquiry, e.g., in Agricultural production, statistics are affected by climate, soil fertility, availability of raw materials, mode of transportation etc.
3. Statistics are numerically expressed, enumerated or estimated. The subject of statistics is concerned essentially with facts expressed in numerical form with their quantitative details. Therefore, facts indicated by terms such as ‘strongly agreed’, ‘agreed’, ‘disagreed’ are not statistics until a numerical equivalent is assigned to each expression in terms of coding.
4. Statistics are enumerated or estimated according to reasonable standard of accuracy. Personal bias and prejudices of the enumeration are not allowed in the counting or estimation of figures, otherwise conclusions from such figures would not be accurate.
5. Statistics should be collected in a systematic manner for a predetermined purpose. If there is no predetermined purpose, all the efforts in collecting the figures may prove to be wasteful.
6. Statistics should be capable of being placed in relation to each other. The collected figure should be capable and well-connected in the same department of inquiry.

## 1.6 Limitation of Statistics

The scope of the science of statistics is restricted by certain limitations among which are:

1. The use of statistics is limited to numerical studies. Statistics deal with only such phenomena as are capable of being quantitatively measured and numerically expressed.
2. Statistical methods deal with population or aggregate of individuals rather than with individuals.
3. Statistical measurements rely on estimates and approximation.
4. Statistical results might lead to fallacious conclusions by deliberate manipulation of figures and unscientific handling.

## Exercises

1. Define and explain the meaning of statistics.
2. 'Statistics are numerical statements of facts in any field of inquiry, placed in relation to each other' Discuss.
3. "Statistics is a science of counting" Comment and give a comprehensive definition of statistics,
4. Examine critically the important definitions of statistics, pointing out the one considered to be the best.

# Chapter Two

# 2.0 STATISTICAL INVESTIGATION

Statistical investigation has been crowned as a highly structural approach to the problem solving of various research questions. The aim of conducting a statistical investigation is to answer the many questions that are present in the environment, and it is a technique that has been commonly applied by statisticians. Most statistical investigations do not set out to obtain data about every item in a population but rely on a sample from the population. The first stage is to decide the size of the sample. A sample must be evenly spread over the population, and choosing a random sample helps to remove bias. A sample investigation is the process of learning about the population based on the sample drawn from it and from which conclusions are drawn.

The main steps utilized in a statistical investigation include four components which are:

1. Clarifying the problem and formulating questions or hypotheses that can be answered with the data.
2. Designing or creating an appropriate experiment that can collect the required data
3. Finding and using the appropriate techniques needed to accurately analyze the collected data
4. Interpreting the collected data and results so as to answer the questions and hypotheses proposed.

## 2.1 Population and Sample

Statistics is basically a science that studies collection and interpretation of numerical of data. A researcher might be interested in studying the intelligence quotient of science students in a tertiary institution. The generality of science students in such institution might be too numerous to comprehend with for effective survey, hence the needs to adopt a few fractional part of the larger group that would have been seen to represent the entire students under consideration.

## Population

A population is a large group of objects about which inferences are to be made. However, definition of population can be stated with respect to different phenomena.

In demography, **population** is the number of living people that live together in the same place. According to the United States Census Bureau, the world's population was about 7.55 billion in the year 2019.

In biology, a population is all the organisms of the same group or species which live in a particular geographical area, and have the capability of interbreeding.

In some cases, the population is obvious. For instance, the population of all registered students in a department of tertiary institution or in all departments of the tertiary institution. Whereas, population is not obvious in some cases. The outcome of a statistical experiment is called an **observation,** and this is not obvious prior to the conduct of the experiment. Thus, the size of a population is the maximum number of observations that can be made in the population. A population can be finite or infinite. A **finite** population has a limited or countable number of individuals or objects while an **infinite** population has an unlimited number. For example, the number of registered students at Federal Polytechnic Ilaro for the 2018/2019 academic session is a finite population while the number of attendants (both market women and customers) at a particular Agricultural market is an infinite population. The task of collecting data from a small finite population is relatively simple while that of large finite population requires much efforts or sometimes impossible. Whereas, collection of complete data from an infinite population is practically impossible.

Generally, population can be classified into four categories

- **Finite:** Countable fixed numbers of units or items.
- **Infinite:** Uncountable numbers of items e.g. grains of rice
- **Real:** items in the population are all physically present.
- **Hypothetical:** a population that results from repeated trials e.g. observations from coin tossing and throwing of die.

## Sample

A sample refers to a small, manageable size of a larger group, which is expected to contain the characteristics of a larger population where it is originated. It is a few or subset of a population carefully selected and studied in order to deduce or make inference about the whole population.

A sample may be large or small, and the results obtained from such a sample may be used to draw inference about the population. A sample must be representative of the population if it is to lead to valid inferences. Sample becomes necessary especially when the population of interest is infinite or where there are no enough resources available to cover the whole finite population.

Sample can be random or purposive. A **random** sample is the one selected in such a way that every member of the population has equal chance of being included in the sample. This is achievable

through scientific procedures such as using Microsoft excel or other statistical packages. Manually conducted statistical experiments, such as tossing of coin, throwing of die, urn method, using a table of random numbers, etc. are equally capable of generating random sample.

## Procedures for Generating Random Sample using Microsoft Excel

1. Create spreadsheets for the population values in as many columns as required
2. Create a new column and title it 'Random'
3. Enter the argument $'=rand()'$ in the first row of the new column and press enter to generate the first random value.
4. Click on the random value generated in step 3, place cursor on the right bottom side of it and drag down to generate random values for the remaining rows.
5. Highlight the new column and copy. Then click on paste and select paste values.
6. Highlight the population values and click on data. Then click on sort to link up with a dialog box.
7. In the dialog box, sort by random and click 'ok'
8. The population column automatically sorts itself randomly.
9. From the population column, highlight the required number of sample for selection.

Figure 2.1 below is a demo of what is expected at the end of the exercise.

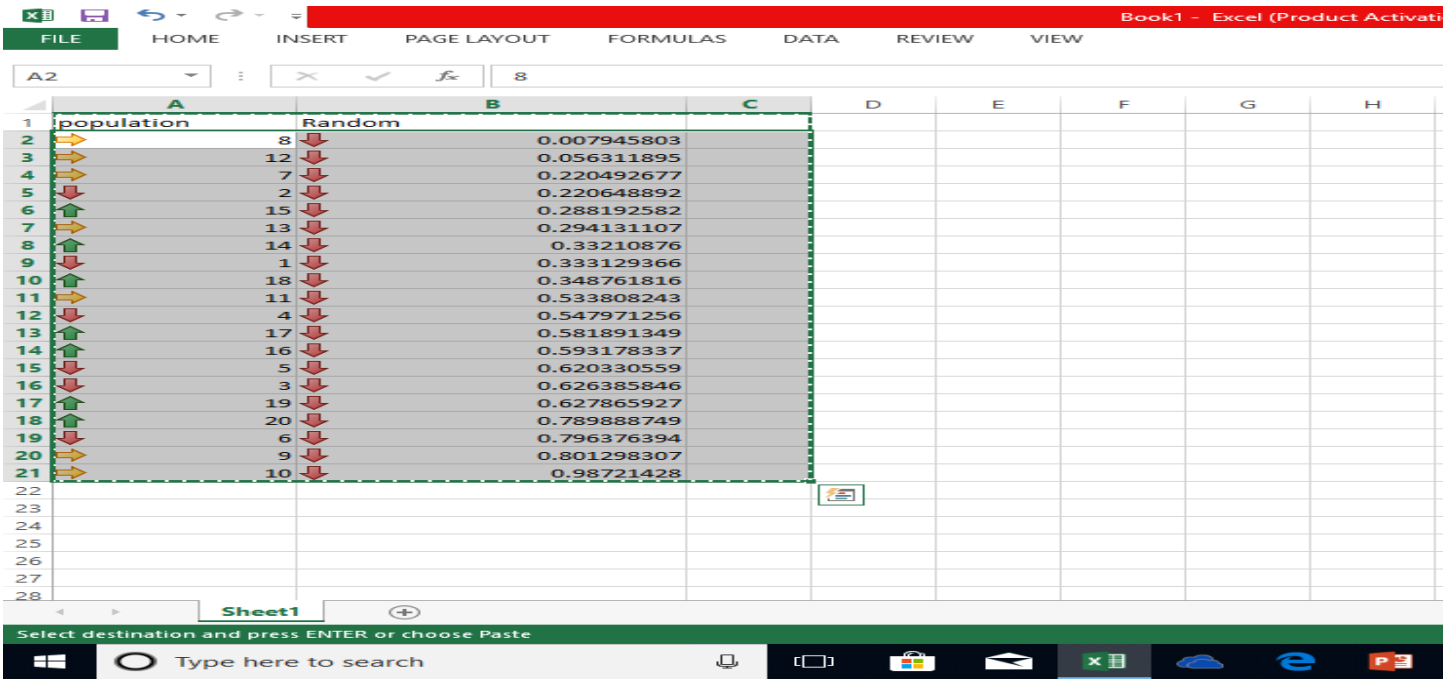


**Figure 2.1: Random number output using Microsoft Excel**

## 2.2 Variables

Variable can be defined as a real-value function defined on the sample space. It is a property with respect to which individuals in a sample differ in some ascertainable way i.e. a quantity changing in value. If the value of a variable cannot be predetermined before observation, then the variable is said to be a random variable.

Random variables fall into two broad categories: Qualitative or Quantitative variables.

**Qualitative Variables** are variables that cannot be expressed in numerical form. They cannot be measured but must be expressed quantitatively in terms of codes for analytical purpose. They are attributes such as skin colour (black or white), gender (male or female), marital status (married, single, divorce), etc.

**Quantitative Variables** are variables that can be expressed in numerical form e.g. age, weight. Length, height, temperature, volumes, number of schools, etc. They can be **Discrete** or **Continuous.**

**A discrete variable** is a variable that can assume only finite or countable number of values between any two fixed points. That is, the variable can only take distinct values. For example, number of senior staff in a departmental office, number of students in a class etc.

**A continuous variable** is any variable that can assume infinite number of values within a given interval. For example age, weight, length, areas, distance, heights, if $x \in [2,4]$**,** then $x$ can take values $2.1, 2.2, 2.3, \cdots, 4$. In fact, there is no limit to the number of values $x$ can take.

## 2.3 Sampling and Sampling Techniques

**Sampling** is the process of sample selection and the manner or scheme of doing it is called sampling techniques or method.

The inference drawn from sample data is extended to the entire population when information about the entire population is not available. As the population in most inquiries becomes quite large, the cost of such census enumeration in time and money will be substantial. Thus, we can estimate the parameters of population without measuring every item of the population provided we can determine the reasonable bounds of errors of such estimates from the true values of the population parameters.

## Reasons/Purpose of Sampling

- Foremost purpose is to obtain maximum information about the population under consideration at minimum cost, time and resources.
- When population is infinite

- When the item or unit is destroyed under investigation.
- When the results are required in a short time.
- When available resources are limited.
- When population under consideration is either constantly changing or in a state of movement.
- When the items or units are scattered.

## Sampling Errors

This occurs as a result of drawing inference from sample results as against the entire population. This is because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population. This is the measure of the difference between a sample statistic and the real population value. For example, the difference between the $\bar{x}$ and μ is measured by the $\sigma_{\bar{x}}$ $(standard\ error)$.

## Non-Sampling Errors

This is caused by other factors which cannot be attributed to sampling. They arise during census as well as sampling surveys because of biases and mistakes such as

- Faulty planning
- Non-response
- Non-random selection of samples
- Incompleteness and inaccuracy of returns
- Compilation errors.
- Coding errors
- Questionnaire errors
- Editing errors

## Parameter

This is some unknown but fixed quantity, computed from population observation or values describing the characteristics of such population. It is the value that summarizes some characteristic of a population e.g. $\mu, \sigma$. It is usually denoted by θ or $K(\theta)$ in the parameter space Ω from a known density function $f(x: \theta)$.

## Statistic

This is the measurement or characteristic obtained from sample e.g. $\bar{X}$, S. It is a quantity computed from sample observations describing T(x) is a function of the sample $X_1 \ldots . . X_n$

Note that the value of statistic varies randomly from one sample to another whereas the value of a parameter is constant.

## Sampling with or without Replacement

When an object is selected from a finite population, we have a choice to replace or not to replace the object before making the second selection. When we replaced, each object can be picked more than once but in case we do not replace, the object can only occur once. Sampling which allowed each object to appear more than once is called **sampling with replacement**, otherwise it is called **sampling without replacement**.

## Construction of Sampling Distribution

The following are the procedures required to construct sampling distribution with or without replacement:

- Given a finite population of size $N$, select all possible sample of size $n$
- The total possible sample (TPS) size for sampling with replacement is $N^n$ while that of without replacement is $C_n^N$
- Compute the statistic of interest in each sample and tabulate the results.
- The resulting distribution is called the sampling distribution of the computed statistic.
- Compute the mean of the sampling of the mean as:

$$\boldsymbol{\mu_{\bar{x}} = \frac{1}{N^n}\sum_{i=1}^{N^n}\bar{x}_i} \qquad \boldsymbol{for\ sampling\ with\ replacement}$$

$$\boldsymbol{\mu_{\bar{x}} = \frac{1}{C_n^N}\sum_{i=1}^{C_n^N}\bar{x}_i} \qquad \boldsymbol{for\ sampling\ without\ replacement}$$

- Compute the standard deviation of the sampling distribution of the mean as:

***For sampling with replacement***

$$\boldsymbol{\sigma_{\bar{x}_i} = \sqrt{\frac{\sum_{i=1}^{N^n}(\bar{x}_i-\mu_i)^2}{N^n}}} \ \textbf{or} \ \boldsymbol{\sqrt{\frac{\sum_{i=1}^{N^n}\bar{x}_i^{\,2}}{N^n} - \mu_{\bar{x}}^2}} \ \textbf{or} \ \boldsymbol{\frac{\sigma}{\sqrt{n}}}$$

***For sampling without replacement***

$$\boldsymbol{\sigma_{\bar{x}_i} = \sqrt{\frac{\sum_{i=1}^{C_n^N}(\bar{x}_i-\mu_i)^2}{C_n^N}}} \ \textbf{or} \ \boldsymbol{\sqrt{\frac{\sum_{i=1}^{C_n^N}\bar{x}_i^{\,2}}{C_n^N} - \mu_{\bar{x}}^2}} \ \textbf{or} \ \boldsymbol{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

**Where** $\boldsymbol{\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2}$

## Example 2.1

A population consists of five numbers 1, 2, 4, 5 and 8. Consider all possible samples of size two which can be drawn with and without replacement from the population. Compute

(a) The mean of the population

(b) The standard deviation of the population

(c) The mean of the sampling distribution of mean

(d) Standard error of the mean.

**Solution- Sampling with Replacement: N = 5, n = 2**

(a) The mean of the population ($\mu$)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+2+4+5+8}{5} = \frac{20}{5} = 4$$

(b) The standard deviation of the population

$$\boldsymbol{\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2} = \sqrt{\frac{1^2+2^2+4^2+5^2+8^2}{5} - 4^2} = \sqrt{\frac{110}{5} - 16} = \sqrt{6} = 2.4495}$$

(c) The possible samples and their corresponding means (sampling means) are:

| TPS = $N^n = 25$ | **1** | | **2** | | **4** | | **5** | | **8** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 1,1 | (1.0) | 1,2 | (1.5) | 1,4 | (2.5) | 1,5 | (3.0) | 1,8 | (4.5) |
| **2** | 2,1 | (1.5) | 2,2 | (2.0) | 2,4 | (3.0) | 2,5 | (3.5) | 2,8 | (5.0) |
| **4** | 4,1 | (2.5) | 4,2 | (3.0) | 4,4 | (4.0) | 4,5 | (4.5) | 4,8 | (6.0) |
| **5** | 5,1 | (3.0) | 5,2 | (3.5) | 5,4 | (4.5) | 5,5 | (5.0) | 5,8 | (6.5) |
| **8** | 8,1 | (4.5) | 8,2 | (5.0) | 8,4 | (6.0) | 8,5 | (6.5) | 8,8 | (8.0) |

**Thus, the mean of sampling of the means is given as:**

$$\boldsymbol{\mu_{\bar{x}} = \frac{1}{N^n}\sum_{i=1}^{N^n} \bar{x}_i = \frac{1+1.5+2.5+\cdots+6+6.5+8}{25} = \frac{100}{25} = 4.0}$$

(d) Standard error of the mean.

$$\boldsymbol{\sigma_{\bar{x}_i} = \sqrt{\frac{\sum_{i=1}^{N^n} \bar{x}_i^{\,2}}{N^n} - \mu_{\bar{x}}^2} = \sqrt{\frac{1^2+1.5^2+2.5^2+\cdots+6^2+6.5^2+8^2}{25} - 4^2}}$$

$$\boldsymbol{= \sqrt{\frac{475}{25} - 16} = \sqrt{3} = 1.7321}$$

$$\boldsymbol{Alternatively,} \qquad \boldsymbol{\sigma_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}} = \frac{2.4495}{\sqrt{2}} = 1.7321}$$

**Sampling without Replacement: N = 5, n = 2**

(c) The possible samples and their corresponding means (sampling means) are:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **TPS=** $C_n^N$ **=10** | **1,2** | **1,4** | **1,5** | **1,8** | **2,4** | **2,5** | **2,8** | **4,5** | **4,8** | **5,8** |
| $\bar{x}$ | **1.5** | **2.5** | **3.0** | **4.5** | **3.0** | **3.5** | **5.0** | **4.5** | **6.0** | **6.5** |

**Thus, the mean of sampling of the means is given as:**

$$\mu_{\bar{x}} = \frac{1}{C_n^N}\sum_{i=1}^{C_n^N}\bar{x}_i = \frac{1.5 + 2.0 + 3.0 + \cdots + 4.5 + 6.0 + 6.5}{10} = \frac{40}{10} = 4$$

(d) Standard error of the mean.

$$\sigma_{\bar{x}_i} = \sqrt{\frac{\sum_{i=1}^{C_n^N}\bar{x}_i^2}{C_n^N} - \mu_{\bar{x}}^2} = \sqrt{\frac{1.5^2+2.0^2+3.0^2+\cdots+4.5^2+6.0^2+6.5^2}{10} - 4^2}$$

$$= \sqrt{\frac{182.5}{10} - 16} = \sqrt{2.25} = 1.5$$

$$\textit{Alternatively},\ \sigma_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2.4495}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = 1.7321\ x\ 0.8660 = 1.5000$$

## Sampling Techniques

Sampling techniques are procedures or scheme adopted to select sample from the population. The choice of technique to adopt is informed by the type of research and characteristics of population. Sampling techniques is broadly divided into two, namely: **Random or Probability sampling and Non-Probability or Non-random sampling.**

## Random sampling

In this technique, Sample is drawn from the entire population according to some laws of chance in which each element in the population has a known chance or probability of being selected in the sample. Some of the random sampling techniques include;

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling
- Multi-stage Sampling

## Simple random sampling (SRS)

In Simple random sampling, each and every item in the population has an equal chance of being sampled and each one has the same probability of being selected. This can be with replacement or without replacement. For example, if we have to select a sample of 400 employees from a universe of 10,000 employees, then lottery can be conducted on all the 10,000 employees. The procedures involved assigning numbers to units of the population from 00001 to N (10,000). A series of number between 00001 and N is then selected randomly. The use of random number tables is another method of selecting samples in this type of sampling technique. The result obtained is then used to make inference about the entire population.

Thus, in selecting $n$ samples from population of size $N$ with replacement, the first to the last sample unit has $\frac{n}{N}$ chance of being selected. Hence, the probability of selecting $n$ specified sample is $\frac{n}{N}, \frac{n}{N} . \cdots, \frac{n}{N} (n\ times)$. In case of without replacement, the first sample has $\frac{n}{N}$ chance of being selected, the second sample has $\frac{n-1}{N-1}$ chance of being selected, the third has $\frac{n-2}{N-2}$ chance of being selected while the last unit ($n^{th}$ sample) has $\frac{1}{N-n+1}$ chance of being selected. Hence, the probability of selecting $n$ specified units is $\frac{n}{N}, \frac{n-1}{N-1}, \frac{n-2}{N-2}, \cdots, \frac{1}{N-n+1}$.

## Basic Condition for the use of SRS

- Homogeneous/ Uniform Population i.e. same characteristics.
- Availability of complete Sampling frame i.e. (complete list which serves as a guide to the population being covered).
- Finite Population i.e. countable.
- The sample size is also determined in advance

**Example 2.2**

Student population in a Polytechnic is 1000. Take a random sample of 100 students from the population.

## Solution

Assign number 0001 to 1000 to each student. Wrap the numbers and place the wrapped numbers in a bag or container. Mixed the numbers thoroughly and select a sample of 100 slips from the bag one after the other. Persons with the chosen slips will constitute the sample. If the sampling is with replacement, each selected slip is replaced back in the bag before the next selection. In

sampling without replacement each selected slip is not replaced back in the bag before the next selection.

## Stratified sampling

This type of sampling design is used for heterogeneous population. The population is first divided into homogenous groups called strata and each group is referred to as stratum, samples are then drawn from each group using the simple random sampling technique. This could be done using **proportional allocation** (proportional to the size of each stratum) or **Optimum allocation** (based on variability within each stratum). The stratification factors could be state, town, group of school etc. Conditions for the use of Stratified Sampling include:

- Availability of complete Sampling frame
- Finite population
- Heterogeneous population but capable of being divided into mutually exclusive groups called strata.
- Sample size is determined in advance

Given a population of size $N$ and a stratum of size $N_h$, where ($h$ is the number of strata), a sample of size $n$ is selected using proportional allocation formula given as

$$n_h = \frac{nN_h}{N}$$

If the sample is to be selected using optimum allocation, the required formula becomes

$$n_h = \frac{nN_h\sigma_h}{\sum N_h\,\sigma_h}$$

$where\ \sigma_h\ is\ the\ variability\ within\ each\ stratum.$

## Cluster sampling

This is used when sampling frame is not available and there is availability of uniform population capable of being divided into groups called cluster. In this type of sampling design, population is divided into small groups called clusters and then selecting the clusters rather than individual elements for inclusion in the sample. A random sample of the clusters is taken and for any cluster selected, all the units constitute numbers of sample.

Conditions for the use of Cluster Sampling include:

- Incomplete or non-availability of sampling frame
- Uniform population capable of being divided into groups called cluster

- Finite Population i.e. countable.
- The sample size in each cluster should be the same and determined in advance

## Systematic sampling

This is a method in which every $k^{th}$ member of the population is selected after the first member of the sample has been selected randomly. Systematic sampling is assumed to be random because the first sample obtained is selected by Simple Random Sampling. Example of Systematic sampling is selecting every $10^{th}$ name on a list. Systematic Sampling can be described as follows:

Let Population size $= N$, Sample Size $= n$, then sampling interval $(K) = \frac{N}{n}$

The first sample is selected randomly between 1 and $k$. Suppose $k$ =5, if the first element selected randomly is 3, then sample consists of those population units occurring as $3,\ 3+5,\ 3+2(5),\ 3+3(5), \cdots, 3+(n-1)5$.

## Multi-stage sampling

This refers to a sampling technique, in which the population is distributed into a number of first-stage sampling units and a sample is taken of these first stage units by some suitable method. This is the first stage of sampling process called the Primary sampling units (PSU). Each of these selected sample from the PSU is further sub-divided into second stage units, and from these again a sample is taking by some suitable method. If the process terminates at this stage, then the scheme is called a Two-stage sampling while the sample taken is called Secondary sampling units (SSU). When further stages are added, the scheme becomes a Multi-stage sampling technique. For instance, in a socio-economic survey that involved West Africa, the first stage of the sampling process is to divide the continent into countries and select in a suitable manner a sample of these countries. The selected countries will again be divided into states and select in a suitable manner a sample from these states. The selected states will again be divided into Local government areas from which sample are further selected. The selected Local government areas will then be divided into smaller sampling units (say towns) and from out of these will be selected the sample units that are to serve as sample for the purpose of investigation.

## Non-random Sampling

This does not involve probability but solely depend on the judgment of the enumerator. Not all elements have an equal chance of being selected in the sample. This include but not limited to the following:

Quota sampling

Judgmental sampling

Convenience Sampling

Snowball Sampling

## Quota sampling

Quota sampling involve general breakdown of the sample into proportion. It is a form of stratified sampling where selection within strata is non-random and interviewers are simply given quota to be filled from different strata, actual selection of sample is based on interviewer's judgment. This is ideal for characteristics of subgroups. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population.

## Judgmental sampling (Purposive sampling)

This sampling method involves deliberate selection of particular units (experts) of the population because of certain characteristics the sample possess. This is ideal when population is very difficult to locate. An expert uses personal judgment to select what a truly representative sample will be. It involves human judgment. There could be bias in this method. This is used when only a small number of sampling units are in the population, and when studying some unknown traits of a population.

## Convenience Sampling

In this sampling method, the researcher attempts to obtain a sample of convenient elements by selecting convenient sample units. The respondents are selected because they happen to be in the right place at the right time, examples are street interviews, journalists interviewing people on the street.

## Snowball Sampling:

This technique begins with the selection of the initial respondents on random basis perhaps through the use of judgmental sampling technique based on the required population characteristics. After respondents on this initial selection list have been interviewed, they are requested to give the researcher referral on the other possible respondents with similar features. The researcher then moves from referral to referrals thus creating a snowball effect, hence the name of this technique.

## 2.5 Scale of Measurements

Measurements is the process of assigning values or scores to the observed variable or phenomenon. Measurement exists in several levels depending on what is to be measured, the instrument to be employed, degree of accuracy desired and the method of measurement. To understand variables in used, it is pertinent to know their level of measurement. For example, figure 2 might indicate a mark of two or a situation that the object is ranked second in the class. For proper understanding of these differences, types of levels of variables have been identified as follows:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

**Nominal Scale** Is the weakest form of measurement which involved merely classifying observed data into various categories without ordering e.g classification of politician into parties' affiliation.

**Ordinal Scale:** Here, the observed value classified into various categories are ordered e.g class of degree classified into 1st class, 2nd class upper, 2nd class lower, 3rd class and Pass.

**Interval Scale:** It is an ordered scale in which the difference between measurements is a meaningful quality. For instance, the two (2) inches difference in the height of 2 individuals with heights 6.5inches and 6.7 inches has a meaning in the scale. Example of interval scaling are Temperature, calendar time.

**Ratio Scale:** If in addition to differences being meaningful and equal at all points on scale, scale also has a meaningful zero point, then we refer to it as a ratio scale. For instance, a person who is 76 inches tall is twice as tall as someone with height 38 inches. Ratio scales are the most sophisticated of scales, since it incorporates all the characteristics of nominal, ordinal and interval scales.

## Exercises

1. What is statistical enquiry? Describe the main steps in a statistical enquiry.
2. A population consists of five numbers 2, 3, 4, 7 and 9. Consider all possible samples of size three which can be drawn with and without replacement from the population. Compute

(a) The mean of the population

(b) The standard deviation of the population

(c) The mean of the sampling distribution of mean

(d) Standard error of the mean.

3. Sampling is necessary under certain conditions. Discuss.
4. Compare the various methods of sample selection.
5. Distinguish between probability and non-probability sampling
6. Explain the law of large numbers, giving instances in which the law is relevant in statistical investigation.
7. State the of statistical regularity and give suitable illustrations.
8. State clearly the important features of all the methods of probability sampling.

# Chapter Three

## 3.0 DATA COLLECTION DATA CLASSIFICATION

Data collection is the biggest task in statistical inquiry. To achieve a seamless and useful data collection, it will be necessary to give consideration to the following preliminary activities:

- Statement of purpose
- Scope of inquiry
- Choice of statistical units
- Choice of data collection technique
- Standard of accuracy

The bedrock of every statistical investigation prior to data collection is the preparation of a statement of purpose for the statistical inquiry at hand. This will give the investigator some knowledge of the technical requirements for the survey. Failure to work out a meaningful statement of purpose can lead to the choice of wrong field for the inquiry/survey and subsequently, collection of meaningless data that will not be relevant to the purpose of inquiry. The scope of a particular statistical inquiry will be decided with reference to the available space, time and the number of items to be covered.

The choice of statistical units is essential for the correct solution of any statistical problem. It is not enough that the data are collected with utmost accuracy, but it is essential too that the unit employed for expressing the numerical information is appropriate, definite, stable and specific.

Having determined the scope of inquiry and suitable units of measurement, the next step is to determine the most suitable method of data collection. In this case, the investigator can either look into records of institutions that engages in data collections and publications for public consumption or carry out a special survey using a suitable data collection technique. The selection of a particular source is dependent upon variety of factors such as time availability, desired accuracy, available funds, investigators' expertise etc.

The determination of statistical accuracy that is to be observed is equally of essence before embarking on data collection. The degree of accuracy that is necessary and the accuracy level which is actually attainable need to be put into consideration. Though, it is pertinent to bear in mind that absolute accuracy is neither attainable nor highly necessary in data collection.

## 3.1Primary and Secondary Data

**Primary data** refers to the statistical observations which the investigator originates for the purpose of research being embarked upon. Thus, if it is desired to conduct a research into the market acceptability of a newly launched product, and the opinion pertaining to this research are collected by the investigator or his agents directly from field of inquiry, such data would be termed as primary data.

The term **Secondary data** on the other hand refers to that statistical facts which Is not originated by the investigator himself, but which he obtains from some established records such as periodic publications of statistical agencies, journals, newspapers, textbooks, internet sources etc.

The difference between primary and secondary data is largely that of degree. Data which are primary in the possession of an individual may be secondary in the hands of another. Thus, the data collected during population census are primary to the National Population Commission (NPC), but to a person who makes use of these data for further research, they will be termed secondary data.

## 3.2Primary Methods of Data Collection

These methods that aim at collecting primary data are referred to as primary methods. These comprise of:

- Personal Interview
- Telephone Interview
- Mailed Questionnaire
- Questionnaire in charge of enumerators
- Direct Observation
- Results of Experiment
- Statutory Registration

**Personal Interview**

This method required the investigators to personally contact the respondents and explain clearly in form of interview, the objectives of the survey and data requirements. This approach will give room for persuasion in case of unwilling respondents and thereby allowed required information to be sourced genuinely.

**Advantages**

I. It is best suited to situations where the problems cannot be completely understood by the targeted audience.
II. This method is suitable in social anthropological research where the questions cannot be formulated beforehand and one question leads to another.
III. It is equally useful in situations where great depth in study is required.
IV. It provides reliable and genuine response especially when experienced interviewers are involved.

**Disadvantages**

I. The method is time consuming
II. It is not suitable for large groups of respondents.
III. It is very expensive
IV. Errors may be committed while recording responses of respondents during interview.

**Telephone Interview**

This method has become very prominent in collecting data globally. With the increasing number of mobile phone users around the globe, it is possible for enumerators to interview the identified respondents through phone calls and record their responses.

**Advantages**

I. The method is not expensive to manage, since the only cost involved is that of telephone.
II. It is practically easy to get a wider coverage using this method, since majority of the populace, both educated and illiterates now uses mobile phone.
III. The time taken to obtain responses is usually fast, especially when the targeted respondents are in the right frame of mind.
IV. It is possible for the investigator to monitor the recruited interviewers for efficiency.
V. It is the most convenient method of data collection. The investigator can carry out the interview at his/her comfort zone at any convenient time provided it agrees with that of the respondents.

**Disadvantages**

I. It may be difficult to get elaborate answers to questions asked through telephone.
II. Responses received is strongly dependent on the mood of the respondents.

III. Occasional delays may occur in getting responses, especially in developing countries where network problem is still an issue of concern.

**Mailed Questionnaire**

In this method, one drafts a detailed questionnaire and mail it to the targeted respondents for filling and subsequent returning to the investigator. This method is suitable in the environments where the service of postal agency is superb, and it requires the inclusion of postal stamp in the questionnaire to enable the respondent mail their responses at no cost on their part. However, with the global usage of internet services, the use of postal agency has been almost relegated to the background. Investigators now mail structured questionnaire through internet based medium where respondents can easily access and respond through the same medium.

**Advantages**

I. The costs of running this method are relatively less.
II. It allows for consultation on the part of respondents to enunciate valid opinion.
III. It is a very useful method of getting information that cannot be sourced conveniently from respondent on face to face with the interviewers.
IV. It is convenient to use.

**Disadvantages**

I. One of the drawbacks of this method is that it is very difficult to design a questionnaire that can be understood by all and fill appropriately. Often the questions intent is misunderstood leading to inaccurate or even irrelevant responses.
II. The method increases the effort a respondent has to put in filling the questionnaire as a result of too many detailed instructions, and this reduces the percentage of response.
III. Another fault of this method is that, there is no way of ensuring that questionnaire are filled and returned. This often results in very poor percentage of return of which those responding may not form a fair sample of the population.
IV. There may not be enough motivation with the targeted respondents and which may cast serious doubts on the validity of responses.
V. Respondents may lack the knowledge of the required facts.

**Questionnaire in charge of Enumerators**

In this method, one drafts a detailed questionnaire and put in charge of enumerators who go around and fill them after obtaining the desired information or better still allow the respondents to fill the

questionnaire themselves if they are literate. These enumerators may be employed either on a pay basis or may offer services free. The enumerators are usually well trained and instructions are given to them regarding the way in which the schedules are to be filled and the information elicited.

**Advantages**

I. It is a suitable method where respondents are illiterate
II. The enumerators can see to it that only relevant answers are obtained from the respondents
III. The method ensures great reliability as the accuracy can be checked by supplementary questions
IV. The method is very useful in collecting information on exceptional difficult items on a prepared questionnaire.

**Disadvantages**

I. The method can be expensive especially when training of enumerators is involved.
II. Errors may be committed during the recording of respondents' responses by the enumerators.

**Drafting of Questionnaire**

A questionnaire is a document or form (printed or electronic) which contains structured questions designed to collect information for the purpose of statistical enquiry. The questionnaire is designed to facilitate response and makes data processing easy. The following are some of the points that should be kept in mind while drafting the questionnaire:

1. **Clarity**- The researcher must be clear about the various aspects of his research problem and should be able to ask questions that are as simple and as clear as possible. Thus a question on marital status scheduled in the following manner is simple and quite clear to solicit genuine response.

   **Marital Status (**please tick as appropriate**).**

   **Married** ☐ **Single** ☐ **Separated** ☐ **Divorced** ☐ **Widowed** ☐

2. **Avoid open-ended questions**. The pattern of questions depends on the nature of information being sought, the sampled respondents and the nature of proposed analysis. Questions can either be open-ended or close-ended, the choice of which must be decided by the researcher. Usually open-ended questions are discouraged in questionnaires because of the complications usually encountered in harnessing its information. Close-ended

questions are well-structured and responses of many respondents can be easily harnessed and collated for analysis.

3. **Ambiguous questions should be avoided**. The researcher should prepare a draft of questions that are not vague and unclear to the respondents. Questionnaire must contain straight forward questions that the respondents would not experience any difficulty in answering. Questions asked should ordinarily solicit unambiguous response without too much interference from the enumerators.
4. **Avoid certain types of questions**. Those questions should be avoided which are likely to among others
    a) Arouse the resentment of the respondent
    b) Frighten the respondents
    c) Insult the personality of the respondents
5. **Number of questions.** The number of questions asked should be consistent with the scope of the research.
6. **Units of measurements.** The unit in terms of which information is to be given must be clearly instructed in the questionnaire.
7. **Structure of questionnaire**. Depending on the type of study, questionnaires are usually structured into sections. The first section of a questionnaire is usually the demographic part, and information such as marital status, sex, age, academic qualification, etc. is sought in this part. Questions in the remaining part of the Questionnaire are expected to conform to the objective of the study. For example, assuming the objective of the study is to research into the "impact of ICT on the academic performance of students of tertiary institutions". After structured demographic questions, questions that should follow are expected to relate to the stated objectives.
8. **Pilot study** should be conducted so as to pre-test the questionnaire for its suitability in conducting the survey. In view of the reliability test conducted using Cronbach's Alpha value, the questionnaire may be edited following results of the pilot study.

**Direct Observation**

This method require that each unit of the sample is examined or observed physically for data collection. It may involve using of measuring instruments or other useful devices to aid data

collection. For example, census of vehicular traffic intensity will require the enumerators to be physically present to take count of vehicular movement at the point of data collection.

**Advantages**

I. Data collected is devoid of exaggeration
II. It provides reliable result especially when the survey is being supervised by experienced enumerators.
III. Supervision by a superior officer is relatively easy during survey.

**Disadvantages**

I. It is laborious
II. The method is time consuming
III. It is not suitable for large groups of respondents.
IV. It is very expensive

**Results of Experiment**

This method requires the conduct of scientific experiment to generate data. Such experiments are usually conducted in the laboratory under strict scientific procedures, and quality of data collected is largely dependents on the expertise of the person in charge as well as the apparatus used in conducting the experiments.

This method is largely used by researchers in the fields of Engineering, Biological sciences, Biostatisticians, Chemical sciences, Microbiologists, Agronomists, Medical practitioners etc.

**Advantages**

I. Data collected is mostly reliable since it is based on scientific approach.
II. Supervision by a superior officer is relatively easy during the conduct of experiments.

**Disadvantage**

I. The accuracy of data collected is largely dependent on the quality of apparatus used.

**Statutory Registration**

This method involved data collection through the statutory registration of the populace vital statistics. Every respondent is expected to have reported and registered as required by law, certain events in their lives. According to vital registration system, events such as births, deaths, marriage, divorce, immigration and emigration are regarded as vital statistics and are expected upon their occurrence, to be reported and registered with the appropriate authority.

**Advantages**

I. There is likely to be high rate of response since it is backed by law.

II. It is not expensive.

**Disadvantage**

I. The level of compliance to vital registration system is usually very low especially in developing countries, thereby resulting into high rate of non-response.

## 3.3 Possible Sources of Errors in Data Collection

The following are some of the errors associated with primary data collection:

- **Non response error –** This arises from the inability to obtain a useful response to all survey items from the targeted respondents. A critical concern is when that nonresponse leads to biased estimates.
- **Specification error –** This may occur when there is a mismatch between what the survey is measuring and what it is intended to measure.
- **Measurement error –** This includes a large family of errors that may occur when response on a survey results in the collection of inaccurate or incomplete information
- **Faulty planning**
- **Non-random selection of samples (Sampling error)**
- **Incompleteness and inaccuracy of returns**
- **Compilation errors**
- **Coding error**
- **Questionnaire errors**
- **False response error**

## 3.4 Sources of Secondary Data

Data collection is not always necessary to be conducted through survey. Data can equally be obtained from records of institutions that collect and publish statistics as part of their routine duties. The most prominent routine compilers and publishers of statistics are government agencies. Few data are usually under the purview of quasi-government and private establishments. Statistical data also appears in trade journals, magazines, market reports and other periodicals. With reference to the manner of its publication, secondary data may be divided into three groups:

- **Regular Publication.** These are statistical data published at known intervals. Examples of such are daily All share index published by Nigerian Stock Exchange (N.S.E), monthly market indices published by state department of statistics, monthly consumer price index (CPI) published by National Bureau of Statistics (NBS), etc.
- **Periodical Publication.** Example of such data is Nigeria population census figure published every ten years.
- **Irregular Publication**. This consist of special survey of statistical phenomenon, with no regular dates of publication, e.g., the reports of National minimum wage committee.

## 3.5 Associated Problems of Data Collection in Developing Countries

1. Lack of adequate funding
2. Lack of cooperation on the part of the respondents
3. High level of illiteracy on the part of the respondents
4. Lack of trained personnel
5. Lack of modern equipment to assist in data collection
6. Non availability of adequate logistic support
7. Political factors

## 3.6 Data Editing

This involved the scrutiny of duly filled returned questionnaires and it is usually done at the early stage with a view to detect errors, misinformation, inconsistencies and omissions. The work of editing requires skill and scientific understanding of survey at hand, and the questionnaires must be edited for consistency, completeness, uniformity and accuracy. In this case, defectives questionnaires would be returned for amendment or rejected completely. Once the returned questionnaires are found to be consistent, data should necessarily be condensed into a few manageable groups and tables for further analysis. Data collected through survey are usually coded at this stage through which the categories of data are transformed into quantitative variables that may be tabulated. Computer application is of great importance in data management, it not only saves time but also make it possible to study large number of variables affecting a problem simultaneously. Analyzing from a questionnaire for instance may require the researcher to number all the questionnaires, so as to get them ready for coding. At this stage coding can be done directly from questionnaire into the computer either in Microsoft excel sheets or any other statistical packages such as

Statistical Package for Social Sciences (SPSS). Once the codes have been entered into the computer package, suitable analysis can then be performed for inferences.

## 3.7 Data Classification

Classification is the process of arranging data in groups or classes according to their affinities, and gives expression to the unity of attributes that may subsist amongst a diversity of information obtained during survey.

The data collected for the purpose of a statistical inquiry sometimes consist of a few simple values which can be easily understood without any kind of special treatment. However more often, there is an overwhelming mass of responses which are too detailed without any form of structure. Data obtained from primary source are obviously in a raw state while secondary data also are not better either, inasmuch they are not in the form that is suited for the purpose of inquiry.

The set of characteristics one choses as the basis of classification depends on the nature of study under consideration. Thus, data can be classified according to the following attributes:

- **Quantitative Classification**- This involves classification of data according to some variables that are measurable. Attributes such as age, heights, weights, number of items etc. are quantitative in nature. Examples of such classification are given in Tables 3.1 and 3.2 below:

**Table 3.1: Classification of Undergraduates Students by Age**

| Age | Number |
|---|---|
| 16 - 18 | 1,500 |
| 19 – 21 | 5,000 |
| 22 – 24 | 2,000 |
| 25 – 27 | 2,500 |
| 28 – 30 | 2,000 |
| **Total** | **13,000** |

**Source: Extracts from Records**

**Table 3.2: Classification of University Students Enrolment by Faculties**

| Faculty | Number |
|---|---|
| Science | 2,500 |
| Management Studies | 5,000 |
| Environmental | 1,000 |
| Engineering | 1,500 |
| Social Sciences | 3,000 |
| **Total** | **13,000** |

**Source: Extracts from Records**

- **Qualitative Classification**- This involves classification according to some attributes which cannot be measured numerically. Attributes such as gender, marital status, religion, state of origin, nationality, academic qualification, etc. are qualitative in nature. Example of such classification is given in Table 3.3 below:

**Table 3.3: Classification of Polytechnic Students Enrolment by Gender**

| Gender | Number |
|---|---|
| Male | 16,000 |
| Female | 17,000 |
| Total | 33,000 |

**Source: Extract from Records**

- **Periodical Classification**- Data can also be classified into periods such as days, weeks, months, quarters and years. Example of such classification is given in Table 3.4 below.

**Table 3.4: Classification of Annual University Students Enrolment**

| Years | Number |
|---|---|
| 2015 | 229,500 |
| 2016 | 235,000 |
| 2017 | 251,000 |
| 2018 | 261,500 |
| 2019 | 2 63,000 |
| **Total** | **1,240,000** |

**Source: Extract from Records**

- **Occupational Classification –** Data can also be classified according to set of tasks performed by individual. Major occupational classification is industry, commercial and services. For instance, the entire workforce of a country can be classified as shown in the table below:

**Table 3.5: Classification of Country's Workforce by Occupation**

| Job Categories | Number (Millions) |
|---|---|
| Industry | 36 |
| Commercial | 47 |
| Services | 73 |
| Total | 166 |

**Source: Extract from Records**

- **Geographical Classification –** This involves data classification according to physical distribution. Such classifications are in terms of regions, geo-political zones, states, local government, towns etc. For instance, the classification of states in Nigeria into the six geopolitical zones is presented in Table 4.6 given below:

**Table 3.6: Classification of States into Geopolitical Zones**

| Zones | States |
|---|---|
| North West | Jigawa state, Kano state, Kaduna state, Katshina state, Kebbi state, Sokoto state and Zamfara state. |
| North East | Borno state, Bauchi state, Yobe state, Adamawa state, Yobe state, Taraba state and Gombe state |
| North Central | Federal Capital Territory, Niger state, Kwara state, Kogi state, Benue state, Nasarawa state and Plateau state. |
| South West | Lagos State, Oyo state, Oyo state, Ogun state, Ekiti state and Ondo state. |
| South East | Abia state, Enugu state, Anambra state, Ebonyi state and Imo state. |
| South South | Akwa Ibom state, Delta state, Rivers state, Bayelsa state, Edo state and Calabar state. |

**Source: Extract from Internet**

## 3.8 Frequency Distribution

The aftermath of every survey usually bring forth a mass of shapeless data not capable of being readily assimilated or interpreted. Thus, a pertinent need for the data to be organized properly.

The first task to be carried out in the organization of collected data is to prepare an array by arranging the values of the data in ascending order (i.e. from smallest to the highest). This will enable the range over which the data items are spread to be known, as well as the data general distribution.

Consider the original data obtained during survey of the ages of 100 students among the registered undergraduates of a tertiary institution in Nigeria, as presented in the table below.

**Table 3.6: Original Data of Ages of Undergraduate Students Collected**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 18 | 17 | 15 | 15 | 16 | 16 | 18 | 20 | 21 |
| 15 | 15 | 16 | 17 | 21 | 22 | 23 | 23 | 24 | 17 |
| 18 | 16 | 19 | 19 | 20 | 20 | 18 | 19 | 19 | 18 |
| 21 | 22 | 23 | 22 | 16 | 16 | 17 | 18 | 19 | 21 |
| 26 | 26 | 25 | 23 | 17 | 17 | 16 | 19 | 20 | 22 |
| 21 | 23 | 18 | 17 | 17 | 16 | 16 | 23 | 16 | 19 |
| 18 | 19 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| 24 | 28 | 28 | 27 | 26 | 25 | 24 | 23 | 24 | 18 |
| 17 | 20 | 21 | 21 | 29 | 19 | 18 | 21 | 22 | 23 |
| 19 | 20 | 20 | 21 | 23 | 24 | 25 | 19 | 21 | 22 |

**Source: Institution Records**

By organizing the data of Table 3.6 into the form given in Table 3.7, much more information becomes apparent. A proper scrutiny of Table 3.7 reveals the number of students of a particular age. It can be observed that students of ages 15 years and 24 years old are 5 in numbers each; ages 16, 17 and 23 years are 10 in numbers each; ages 18 and 19 are 12 in numbers each; 20 years of age are 8 in numbers; 21 years of age are 11; 22 years old are 7; ages 25 and 26 years are 3; ages 27 and 29 are 1-year-old and age 28 years are 2 in numbers. Thus, figures 5, 10, 12, 8, 11, 7, 3, 1 and 2 are the frequencies of the listed ages respectively. **Frequency** thus means the number of times a certain observation is repeated in a given data. The Table 4.8 so formed is known as frequency distribution of an ungrouped data while a grouped frequency distribution is one where all classes and their corresponding frequencies are listed.

**Table 3.7: Frequency Distribution of the Ages of 100 Undergraduate Students**

| Ages | Tallies | Frequency | Ages | Tallies | Frequency |
|---|---|---|---|---|---|
| 15 | 𝍸 | 5 | 25 | III | 3 |
| 16 | 𝍸 𝍸 | 10 | 26 | III | 3 |
| 17 | 𝍸 𝍸 | 10 | 27 | I | 1 |
| 18 | 𝍸 𝍸 II | 12 | 28 | II | 2 |
| 19 | 𝍸 𝍸 II | 12 | 29 | I | 1 |
| 20 | 𝍸 III | 8 | | | |
| 21 | 𝍸 𝍸 I | 11 | | | |
| 22 | 𝍸 II | 7 | | | |
| 23 | 𝍸 𝍸 | 10 | | | |
| 24 | 𝍸 | 5 | | | |

The major challenge in preparing a grouped frequency distribution is that of selecting class intervals. This could be resolved if one is sure of the number of classes to be used. However, there are no straightforward rules for determining this, save for a rule of thumb which suggests between 5 and 20 classes, the exact number being determined by other considerations listed below:

- Calculate the range (maximum value – minimum value)
- Determine the approximate number of classes in which the data are to be grouped. Sturges, H. A provides a formula for computing approximate number of classes as:
  $K = 1 + 3.322 \log N$
  (where $K = Number\ of\ classes\ and\ logN = Natural\ logarithm\ of\ observations$)
- Determine the approximate class interval size. This is also known as class width (difference between upper and lower class boundaries). This is obtained by dividing the range of data by the number of classes. i.e.

$$Class\ width = \frac{Range}{K}$$

- Decide the starting point. The lower class limit or class boundary should cover the smallest value in the raw data. Then add the class width to the lower class boundary to compute the upper class boundary.
- Determining the remaining class limits (boundary). When the lowest class boundary has been decided, the remaining lower and upper class limits may be determined by adding the class interval size repeatedly till the largest value of the data is observed in the class.
- Each data entry must belong to one and only one class.

**Illustration 3.1**

Using the original data of table 3.6, construct a grouped frequency (discrete) distribution table.

**Step 1:** $\boldsymbol{Range = 29 - 15 = 14}$

**Step 2:** $\boldsymbol{Number\ of\ Classes}\ (K) = 1 + 3.322 \log N = 1 + 3.322 log 100 = 7.644 \approx 8$

**Step 3:** $\boldsymbol{Class\ width = \frac{14}{8} = 1.75 \approx 2}$

The remaining steps are executed in the given frequency distribution of Table 4.9 and Table 4.10 for discrete and continuous classes respectively.

**Table 3.7: Grouped (Discrete) Frequency Distribution of Students' Ages**

| Ages | Tallies | Frequency |
|---|---|---|
| 15-16 | 𝍸 𝍸 𝍸 I | 16 |
| 17-18 | 𝍸 𝍸 𝍸 𝍸 II | 21 |
| 19-20 | 𝍸 𝍸 𝍸 𝍸 | 20 |
| 21-22 | 𝍸 𝍸 𝍸 III | 18 |
| 23-24 | 𝍸 𝍸 𝍸 | 15 |
| 25-26 | 𝍸 I | 6 |
| 27-28 | III | 3 |
| 29-30 | I | 1 |

**Table 3.8: Grouped (Continuous) Frequency Distribution of Students' Ages**

| Ages | Tallies | Frequency |
|---|---|---|
| 15-17 | 𝍸 𝍸 𝍸 | 15 |
| 17-19 | 𝍸 𝍸 𝍸 𝍸 II | 22 |
| 19-21 | 𝍸 𝍸 𝍸 𝍸 | 20 |
| 21-23 | 𝍸 𝍸 𝍸 III | 18 |
| 23-25 | 𝍸 𝍸 𝍸 | 15 |
| 25-27 | 𝍸 I | 6 |
| 27-29 | III | 3 |
| 29-31 | I | 1 |

### 3.9 Features of Grouped Frequency Data

- **Class Limits**- This can be explained in terms of lower and upper class limits. The lower class limit of a class is the lowest value that can be classified into a class while the upper class limit is the highest value that a class can accommodate. For example, in class 17-18 of table 4.9, the lower class limit is 17 while the upper class limit is 18.
- **Class Boundary**- This is equally of lower and upper class. The lower class boundary of a class is the average of the lower class limit of that class and upper class limit of the preceding class while the upper class boundary of a class is the average of the upper class limit of that class and lower class limit of the succeeding class. For instance, the

lower class boundary of class 17-18 is computed as $\frac{17+16}{2} = 16.5$ while its upper class limit is computed as $\frac{18+19}{2} = 18.5$.

- **Class Midpoint-** This is the average of the lower and upper class limits of a class. For example, the class midpoint of class 17-18 is computed as $\frac{17+18}{2} = 17.5$.
- **Class Width-** This is the difference between upper class boundary and lower class boundary. Thus, the class width of class 17-18 is computed as $18.5 - 16.5 = 2.0$. It is pertinent to note that class width is expected to be uniform for a particular data set.

## 3. 10 Relative Frequencies

Relative frequency is the frequency of a given class divided by total number of observations. It is generally expressed as a percentage. When not reported in percentage, the sum of relative frequencies should be 1.

**Relative Frequency** for a class is computed as $\frac{frequency\ in\ a\ given\ class}{Total\ number\ of\ frequecy} \times 100$

Relative frequency curve (histogram) is the graph of the relative frequency table while frequency curve is the graph of the frequency distribution table. Frequency curve shows the variability of the population and it takes on certain characteristic shapes as indicated in the Figures below.

1. **Normal Curve**

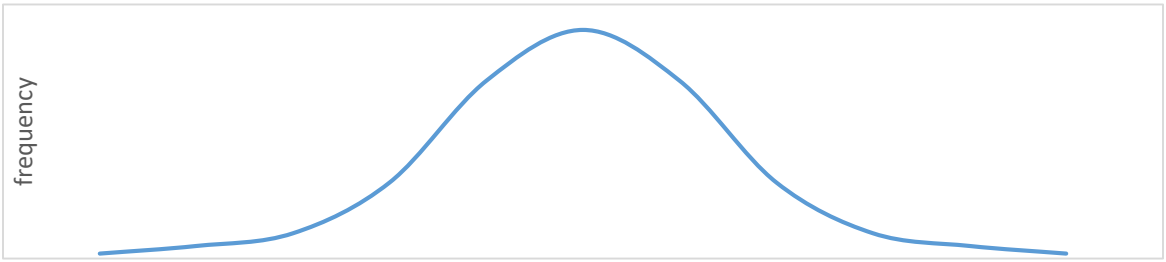


Figure 3.1: Normal curve

2. **Bimodal Curve**

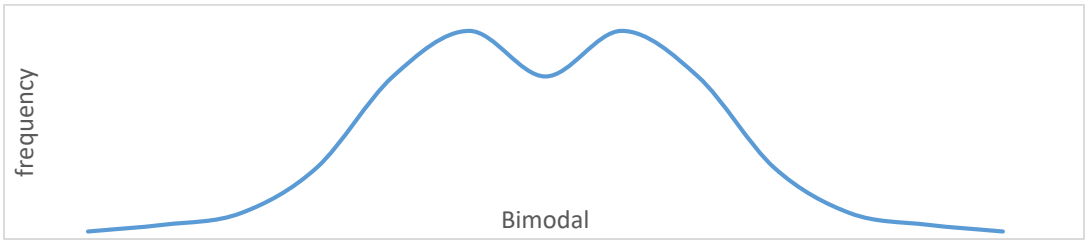


Figure 3.2: Bimodal curve

### 3. Multimodal Curve

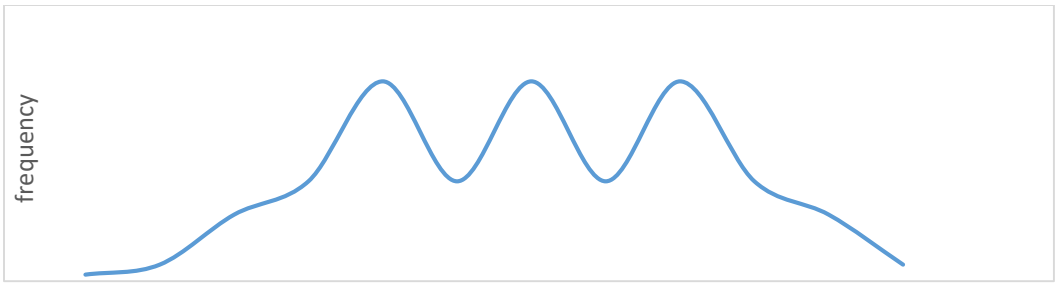


Figure 3.3: Multimodal curve

### 4. Exponential curve

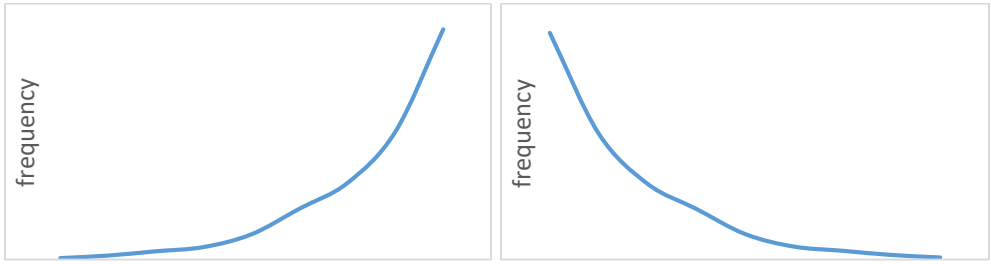


Figure 3.4: The Exponential curve

### 5. Skewed Distribution

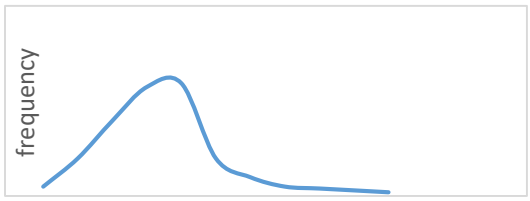


Figure 3.5(a): Positively skewed

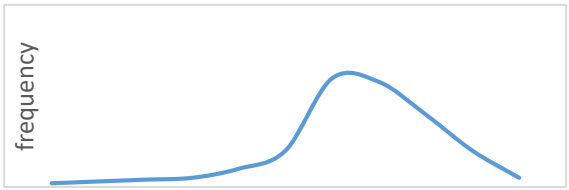


Figure 3.5(b): Negatively skewed

**Table 3.9: Relative Frequency of Students' Ages**

| Ages | Frequency | Relative Frequency |
|---|---|---|
| 15-17 | 15 | 15% |
| 17-19 | 22 | 22% |
| 19-21 | 20 | 20% |
| 21-23 | 18 | 18% |
| 23-25 | 15 | 15% |
| 25-27 | 6 | 6% |
| 27-29 | 3 | 3% |
| 29-31 | 1 | 1% |
| **TOTAL** | **100** | **100%** |

When frequencies of two or more classes are added up, such totals are called **Cumulative Frequency.**

**Relative Cumulative Frequency** is sometimes called percentage Cumulative Frequency. Relative cumulative frequency for a particular class is computed as $\frac{Cumulative\ frequency\ in\ a\ given\ class}{Total\ number\ of\ frequecy} \times 100$

**Table 3.12: Cumulative Frequency of Students' Ages**

| Ages | Frequency | Cumulative Frequency | Relative Cumulative Frequency |
|---|---|---|---|
| 15-17 | 15 | 15 | 15% |
| 17-19 | 22 | 37 | 37% |
| 19-21 | 20 | 57 | 57% |
| 21-23 | 18 | 75 | 75% |
| 23-25 | 15 | 90 | 90% |
| 25-27 | 6 | 96 | 96% |
| 27-29 | 3 | 99 | 99% |
| 29-31 | 1 | 100 | 100% |

**Illustration 3.2: Construction of Frequency Distribution Table in SPSS**

Using the data of example 4.7, the following steps are required in constructing a grouped frequency table in SPSS:

1. Enter the **data name** in the SPSS variable view platform.
2. Enter the **data** in the SPSS data view platform
3. Click on **Analyze→ Descriptive Statistics→Frequencies**
4. Move the variable of interest into the right-hand column.
5. Click on **OK** to generate the frequency distribution table
6. For grouped frequency table, choose **Transform** and select **Recode into different variables** after steps 1 and 2.
7. Select the variable name **Ages** into **the Input Variable → Output Variable** box.
8. Type **Age_cat** in the box titled **Name** and fill the **Label** box with **Ages in categories.** Then click on **Change to reflect** a dialog box as shown below.

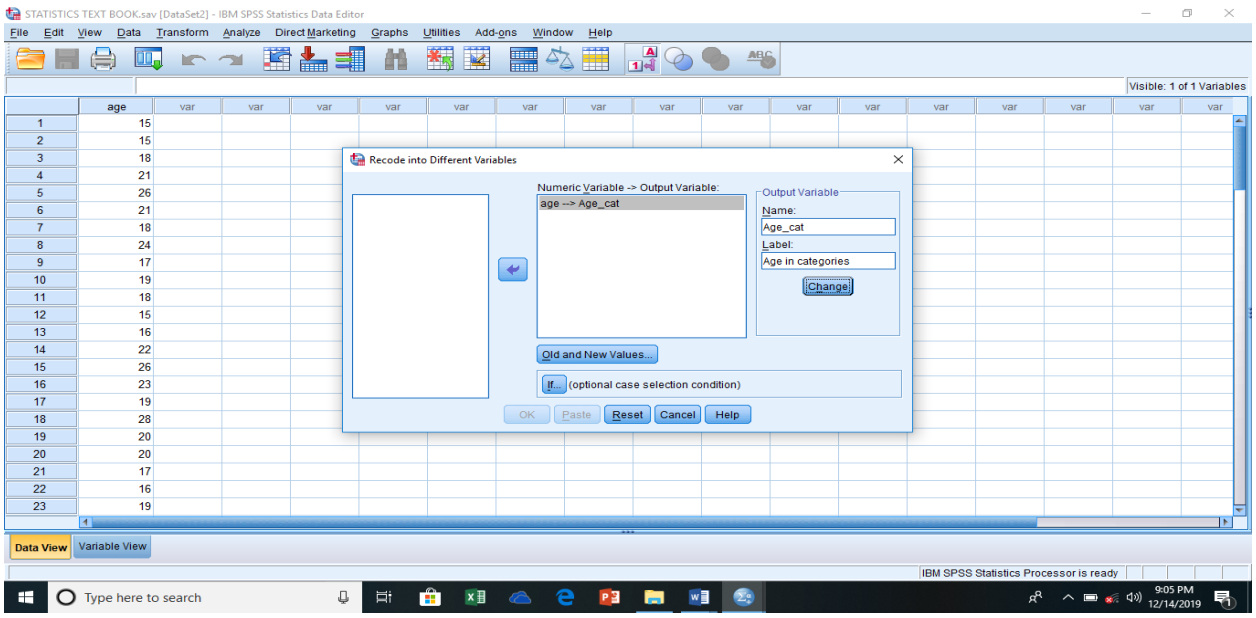


**Fig. 3.6: SPSS Screen Print**

9. Then click **Old and New values** to open a new dialog box titled **Recode into different variables: Old and new values.**
10. In the **Recode into different variables: Old and new values,** Select the **Range** box and type ‘15’ on top of **Through** and ’16’ (‘17’ for continuous interval) below it. Click on the **Output variables are strings** box and the **Width** box will display the appropriate value. Also type ’15-16’ (’15-17’ for continuous interval) in the **Value** box and click **Add**.
11. Repeat step 10 until all the ranges are set as shown in the dialog box below.

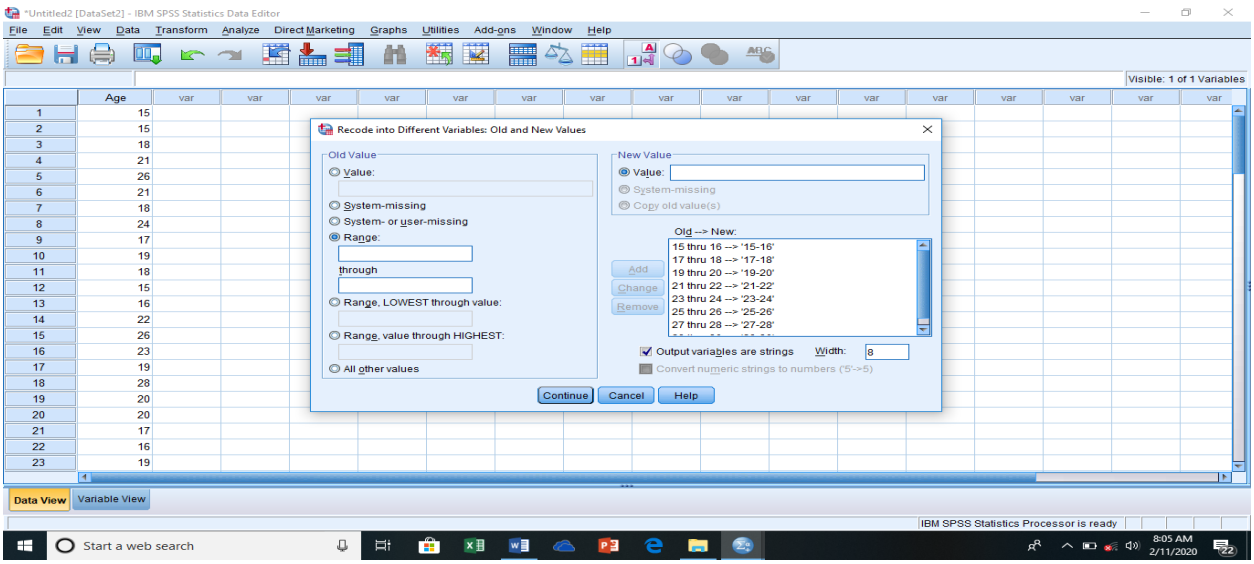


**Fig. 3.7: SPSS Screen Print**

12. Click on **Continue** to return to the **Recode into Different variables** dialog box, and then click **OK** to return to the Data Editor window.
13. In the Data Editor window, choose Analyze → Descriptive Statistics→ Frequencies. In the Frequencies dialog box, select the variable name Age_cat and click the arrow to transfer it into the **Variable(s)** box. Check the **Display frequency table** box and click **OK.**

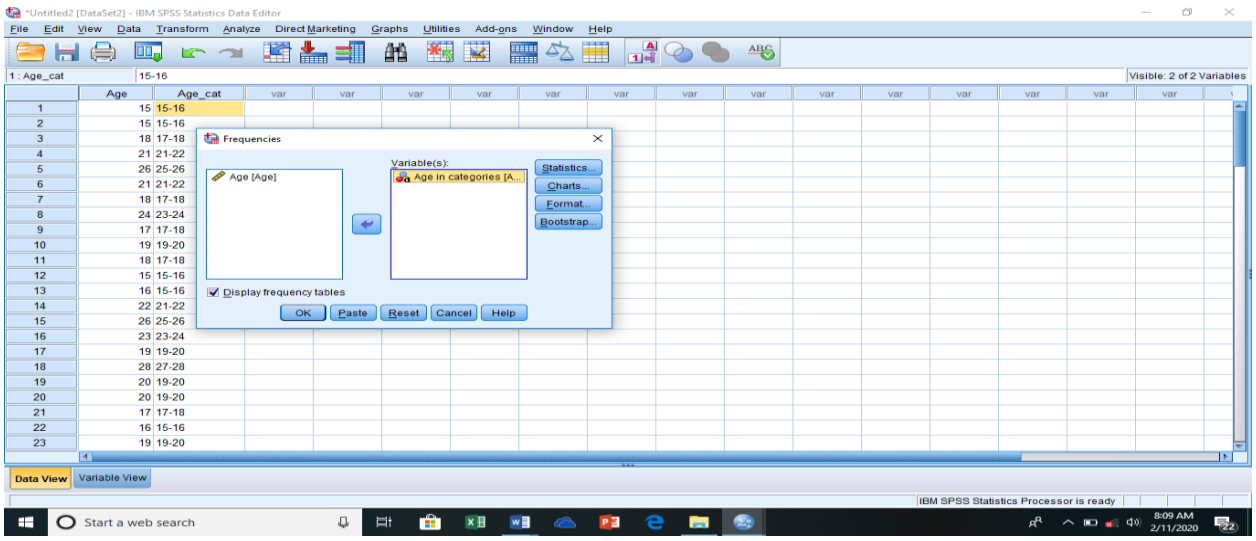


**Fig. 3.8: SPSS Screen Print**

The output table below is generated.

**Table 3.13 Discrete Class Interval : Age in categories**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 15-16 | 16 | 16.0 | 16.0 | 16.0 |
| | 17-18 | 21 | 21.0 | 21.0 | 37.0 |
| | 19-20 | 20 | 20.0 | 20.0 | 57.0 |
| | 21-22 | 18 | 18.0 | 18.0 | 75.0 |
| | 23-24 | 15 | 15.0 | 15.0 | 90.0 |
| | 25-26 | 6 | 6.0 | 6.0 | 96.0 |
| | 27-28 | 3 | 3.0 | 3.0 | 99.0 |
| | 29-30 | 1 | 1.0 | 1.0 | 100.0 |
| | Total | 100 | 100.0 | 100.0 | |

**Exercises**

1. Discuss the primary and secondary methods of data collection. In what special circumstances are the two methods suitable?
2. Explain briefly, three methods of data collection and distinguished them in terms of cost, time and consistency of information.
3. Explain the precautions that must be taken while drafting a questionnaire.
4. Design a questionnaire to research into the acceptability of your course of study.
5. Which method of data collection is suitable to the following type of inquiries?
   (a) Research into the living standard of civil servants in Nigeria.
   (b) Inquiry into the attitudinal behaviors of science and management based students in tertiary institutions.
   (c) Inquiry by a research organization into the acceptability of local rice consumption in Nigeria.
   (d) Market survey of agricultural products in a local market
6. What is data editing and why is it important in data processing?
7. Discuss the forms of classification of a raw mass of collected data and give suitable examples.
8. Distinguish between a relative frequency and cumulative frequency tables.
9. The following are the ages in years of 40 randomly selected students in a class. Tabulate the data in the form of frequency distribution and calculate the relative frequencies.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 20 | 18 | 19 | 20 | 21 | 18 | 17 | 22 | 18 |
| 18 | 17 | 20 | 22 | 19 | 18 | 17 | 19 | 20 | 21 |
| 18 | 19 | 20 | 21 | 22 | 19 | 21 | 23 | 21 | 18 |
| 17 | 19 | 19 | 18 | 20 | 19 | 21 | 20 | 17 | 19 |

10. Following are the salaries of 70 construction workers in thousands of naira

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 70 | 67 | 65 | 75 | 80 | 90 | 100 | 105 | 95 | 69 |
| 100 | 105 | 95 | 75 | 78 | 98 | 89 | 80 | 86 | 85 |
| 86 | 87 | 90 | 95 | 91 | 78 | 79 | 90 | 89 | 100 |
| 79 | 90 | 98 | 90 | 89 | 79 | 80 | 89 | 90 | 99 |
| 98 | 110 | 95 | 120 | 95 | 96 | 120 | 115 | 110 | 120 |
| 95 | 100 | 95 | 95 | 96 | 97 | 96 | 95 | 100 | 95 |
| 110 | 95 | 96 | 100 | 96 | 95 | 79 | 95 | 100 | 96 |

(a) Classify the data in the form of discrete and continuous frequency distribution taking the lowest class as 65 – 74.

# Chapter Four

# DATA REPRESENTATION

Visual representation of information is simpler and more easily understandable. As the size of figures and tables increase, they become confusing and uninteresting to such an extent that no one would be interested to take keener looks into them unless one is specially interested, hence the needs for diagrammatic (visual) representation of information for easy comprehension. Diagrams and charts have become widely acceptable and useful means of communicating statistical information in a concise and precise manner.

The major benefit of diagrams lies in the fact that they facilitate quick comparisons of basic information for meaningful inference.

## 4.1 Line Graph

Line graphs are simple, easy to construct and present. The graph often allows one to detect visually the trends, patterns, outliers or relationships that otherwise might have gone unnoticed. They are mostly used when the number of classes is few.

**Example 4.1:** The table below shows the number of children in families in a particular environment for a year. A random sample of 320 families was surveyed, having 0-9 number of children per family.

**Table 4.1: Survey of Children per Family**

| Number of Children | Number of Families |
|---|---|
| 0 | 10 |
| 1 | 3 |
| 2 | 5 |
| 3 | 83 |
| 4 | 63 |
| 5 | 54 |
| 6 | 25 |
| 7 | 18 |
| 8 | 39 |
| 9 | 20 |

**Source: Field Work**

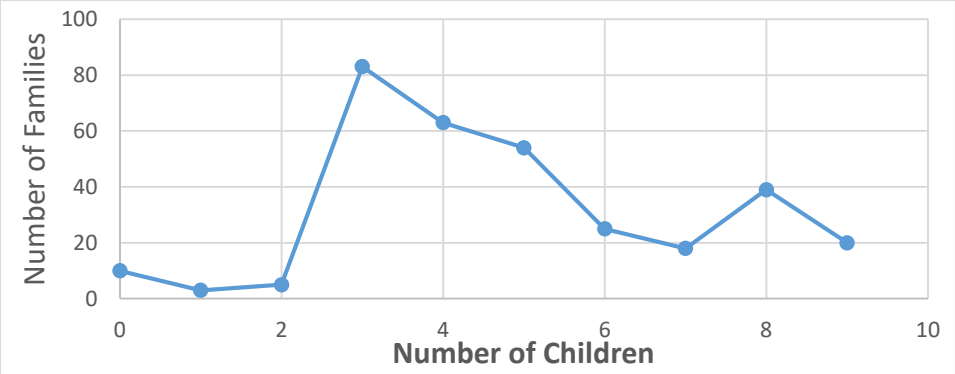


**Fig. 4.1: Line Graph of Children per Family**

It is pertinent to note that the oscillatory movement in the data is more strikingly shown in the diagram than the tabular presentation.

## 4.2 Bar Diagrams

Bar diagram is one of the simplest method of presenting a numerical data. In bar diagrams, the magnitude of numerical data is represented by bars. The easiest of the bar diagrams is the Simple bar chart. Others are Component and Multiple bar charts.

## Simple Bar Chart

A simple bar chart is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis. In a simple bar chart, we make bars consisting of a set of equally spaced bars of equal width, but of varying length which represent the magnitude of a quantity being represented.

**Example 4.2**

Represent the information in table 5.1 on a simple bar chart.

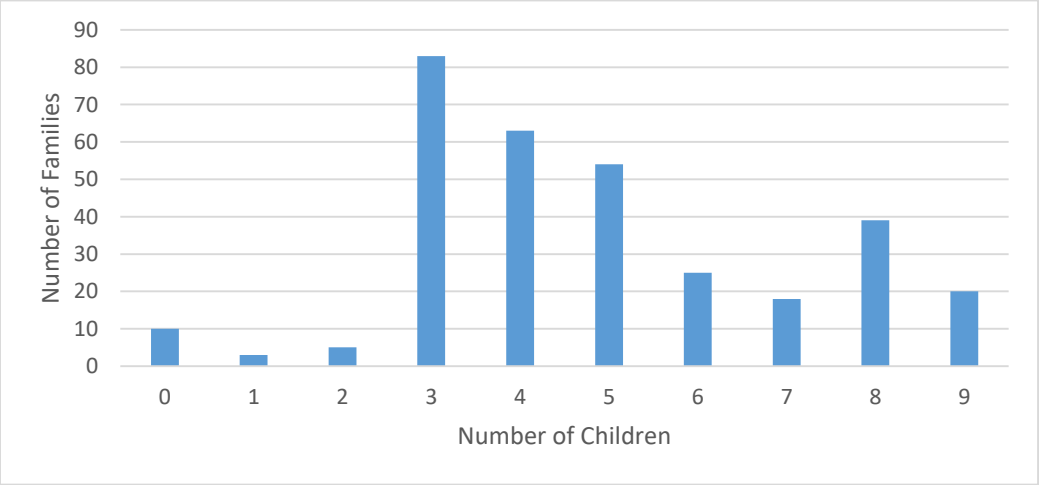


**Fig. 4.2: Simple Bar Chart of Children per Family**

## Component Bar Chart

A component bar chart is used to represent data in which the total magnitude is divided into different components. In this bar diagram, first we make simple bars for each class taking the total magnitude of that class into consideration and then divide these simple bars into parts in the magnitude of required components.

**Example 4.3:** The table below represent the UTME results of four candidates randomly selected from a CBT center.

**Table 4.2: UTME Results of Selected Candidates**

| Candidates | UTME Subjects | | | | |
|---|---|---|---|---|---|
| | **English** | **Mathematics** | **Chemistry** | **Physics** | **Total** |
| **A** | 66 | 82 | 78 | 75 | 301 |
| **B** | 60 | 72 | 67 | 71 | 270 |
| **C** | 70 | 79 | 76 | 75 | 300 |
| **D** | 60 | 65 | 70 | 55 | 250 |

**Source: CBT Centre**

Represent the above information on a component bar chart.

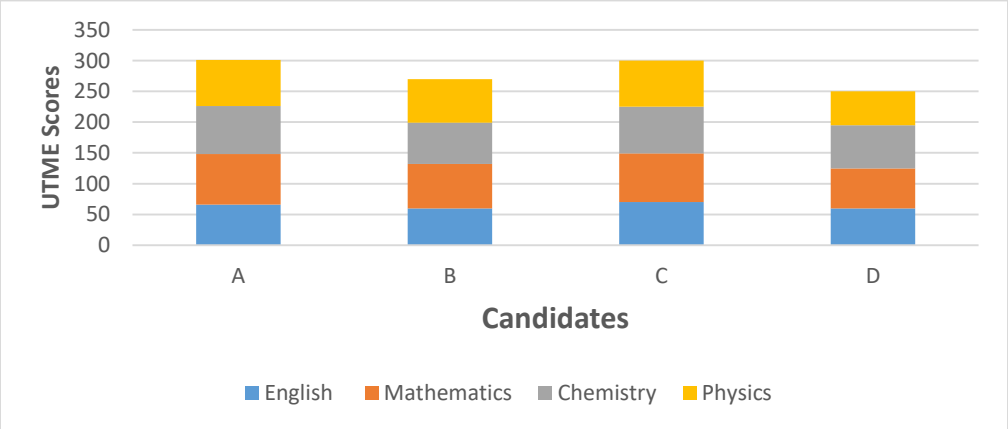


**Fig. 4.3: Component bar Chart of UTME Results**

## Multiple Bar Chart

In a multiple bars diagram, two or more sets of inter-related data are represented. Multiple bar diagram facilitates comparison between more than one phenomena. It is a chart depicting two or more characteristics in the form of bars of length proportional in magnitude of the characteristics. For example, a chart comparing the scores of three students in UTME examination may be drawn with sets of bars, one bar of each pair for each subject's scores, and one pair for each student.

**Example 4.4:** Present the Multiple bar chart of the information in table 4.2

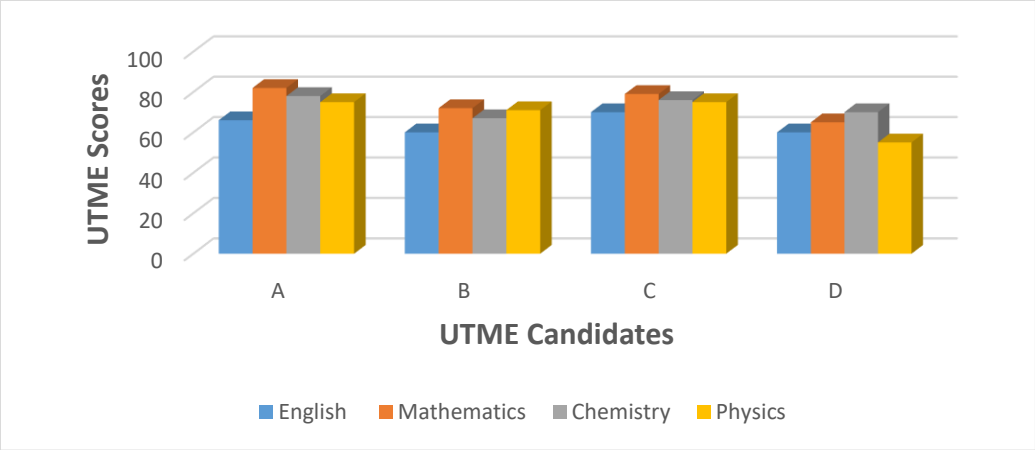


**Fig. 4.4: Multiple Bar Chart of UTME Candidate Results**

## 4.3 Pie Chart

A pie chart is a circle diagram showing how the aggregate of an item is divided into its main components. It is a radius chart divided into sectors with the id of protractor, illustrating relative magnitudes in degrees. It is constructed by dividing 360 degrees proportionately among the components.

**Example 4.5:** The table below represent the percentage of Nigeria's population by geopolitical zones according to the Nigeria Demographic and Health Survey published in 2004.

**Table 4.3: Percentage of Nigeria's Population by Geopolitical Zone, 2003**

| Region | Percentage | |
|---|---|---|
| | **Women** | **Men** |
| North Central | 14.7 | 14.9 |
| North East | 17.9 | 17.9 |
| North West | 27.5 | 25.7 |
| South East | 9.7 | 8.8 |
| South-South | 17.6 | 19.0 |
| South West | 12.6 | 13.7 |

**Source: Nigeria National Population Commission and ORC Macro, *Nigeria Demographic and Health Survey 2003* (2004).**

Represent the above information on a Pie chart.

**Solution**

Women Population (Sector angles) $= \frac{Region\ size}{Total} x\ 360$

North central = $\frac{14.7}{100} x\ 360 = 52.92°$

North East = $\frac{17.9}{100} x\ 360 = 64.44°$

North West = $\frac{27.5}{100} x\ 360 = 99°$

South East = $\frac{9.7}{100} x\ 360 = 34.92°$

South-South = $\frac{17.6}{100} x\ 360 = 63.36°$

South West = $\frac{12.6}{100} x\ 360 = 45.36°$

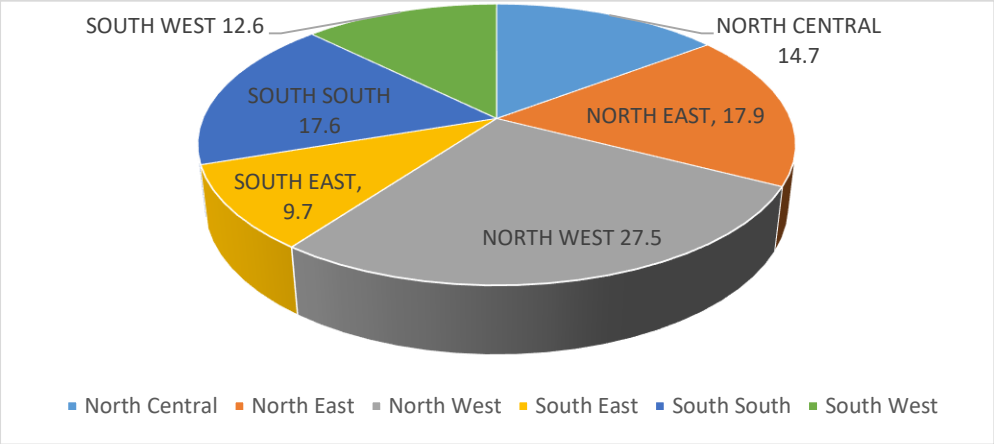


**Fig. 4.5: Pie chart of Women Population**

Men Population (Sector angles) $= \frac{Region\ size}{Total} x\ 360$

North central = $\frac{14.9}{100} x\ 360 = 53.64°$

North East = $\frac{17.9}{100} x\ 360 = 64.44°$

North West = $\frac{25.7}{100} x\ 360 = 92.52°$

South East = $\frac{8.8}{100} x\ 360 = 31.68°$

South-South = $\frac{17.6}{100} x\ 360 = 68.40°$

South West = $\frac{13.7}{100} x\ 360 = 49.32°$

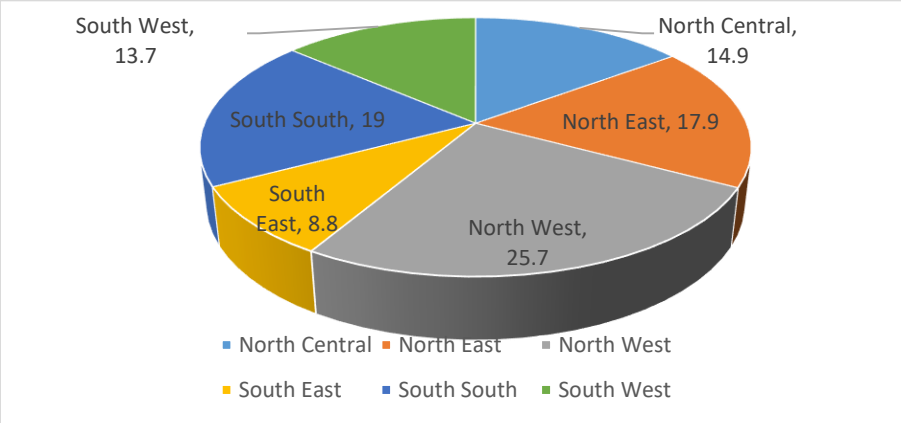


**Fig. 4.6: Pie chart of Men Population**

## 4.4 Pictograms

Presentation of information in pictorial form help in quick visualization of comparison of magnitudes of numerical values of given items. A pictogram might be represented by pictures of Orange, ball or human being as a matter of choice, in which the numbers of pictures correspond to the numerical values of given items.

**Example 4.6:** The following represent the estimated Total Fertility Rates (TFR) in Nigeria's Six Major Geopolitical Zones, 2003 (Ages 15-49). Present the information on Pictogram.

**Table 4.4: Total Fertility Rates in Nigeria's Geopolitical Zones**

| Region | TFR | Approximated Rates |
|---|---|---|
| North Central | 5.7 | 6.0 |
| North East | 7.0 | 7.0 |
| North West | 6.7 | 7.0 |
| South East | 4.1 | 4.0 |
| South-South | 4.6 | 5.0 |
| South West | 4.1 | 4.0 |

**Source: Nigeria National Population Commission and ORC Macro, *Nigeria Demographic and Health Survey 2003* (2004).**

**Solution**

North Central: ⊙ ⊙ ⊙ ⊙ ⊙ ⊙
North East: ⊙ ⊙ ⊙ ⊙ ⊙ ⊙ ⊙
North West: ⊙ ⊙ ⊙ ⊙ ⊙ ⊙ ⊙
South East: ⊙ ⊙ ⊙ ⊙
South South: ⊙ ⊙ ⊙ ⊙ ⊙
South West: ⊙ ⊙ ⊙ ⊙
**Key:** ⊙= 1

## 4.5 Stem-and-Leaf Plot

A Stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram to assist in visualizing the shape of a distribution. It is the combination of graphical and sorting technique. The sorting is based on the data where the stem occupying the left side of a vertical line, represents the leading digit(s) of the data and the leaf occupying the right side is the trailing digit(s). For examples, Stem "3" leaf "1" implies 31 and Stem "1" leaf "7" implies 17.

Stem-and-leaf plot is a useful tool in Exploratory Data Analysis (EDA) and it is particularly useful when the data are not too numerically large. A sample size of less than 200, would be ideal for the construction of Stem-and-leaf.

**Example 4.6**

The following figures represent the number of patients attended to by a consultant of a specialist hospital in 25 days. Construct a stem and leaf plot for the observations.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 26 | 16 | 17 | 72 | 21 | 70 | 31 | 23 | 41 | 23 | 42 | 52 |
| 12 | 18 | 26 | 37 | 17 | 53 | 62 | 61 | 25 | 19 | 20 | 45 | |

**Solution**

First, we sort the observations in ascending order as follows:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 15 | 16 | 17 | 17 | 18 | 19 | 20 | 21 | 23 | 23 | 25 | 26 |
| | 26 | 31 | 37 | 41 | 42 | 45 | 52 | 53 | 61 | 62 | 70 | 72 |

**Table 4.5: Number of Patients Consulted in a Specialist Hospital**

| **Stem** | **Leaves** |
|---|---|
| 1 | 2 5 6 7 7 8 9 |
| 2 | 0 1 3 3 5 6 6 |
| 3 | 1 7 |
| 4 | 1 2 5 |
| 5 | 2 3 |
| 6 | 1 2 |
| 7 | 0 2 |

## 4.6 Box-and-Whisker Diagram

Box-and-whisker plot (also known as Boxplot) is a graphical representation used for describing important features of a distribution such as central value, spread and departure from symmetry through their quartiles. The five values that are practically important in the construction of Boxplots are minimum, $Q_1$, $Q_2$, $Q_3$ and maximum. The three quartiles ($Q_1$, $Q_2$, $Q_3$) are sometimes called hinges of the plot. Boxplots is often used in exploratory data analysis (EDA) and it is a useful tool in identifying outlying values (outliers) from a given set of data. An outlier is an observation that lies in an abnormal distance from other values in a random sample from a population.
A box plot is constructed using the following steps:

- Arrange the data in ascending order
- Find the median of the data set. Note that the median is also called the second quartile ($Q_2$)
- Find the first ($Q_1$) and third ($Q_3$) quartiles. The first quartile would be the median of the numbers to the left of $Q_2$ while the third quartile is the median of the numbers to the right of $Q_2$
- Draw a plot line long enough to contain all the data points plus a little extra on either side.
- Mark the first, second and third quartiles on the plot line. The mark should be a vertical line starting slightly above the plot line.
- Make a box by drawing horizontal lines connecting the quartiles.
- Mark the outliers by placing a small dot on the smallest and largest numbers in the data set.
- Connect the outliers to the box with a horizontal line called the "whiskers" of the box or whiskers plot.

**Illustration 4.1: Construction of** Boxplot **in SPSS**

Using the data of example 5.7, the following steps are required

- Enter the data
- Click on **Graphs → Legacy Dialogs→ Boxplot** to reflect a dialog box shown below:

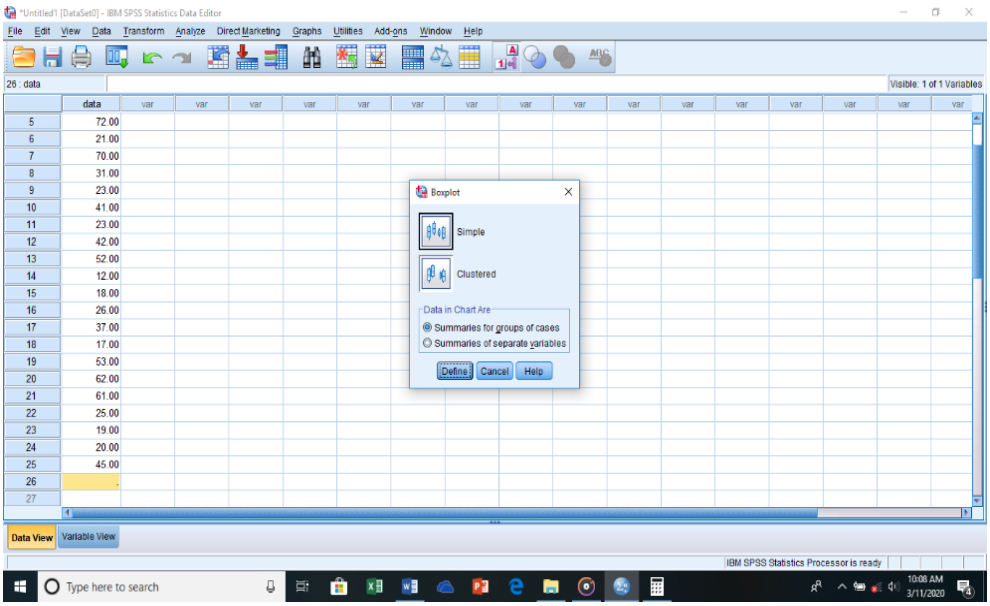


**Fig 4.8: SPSS Screen Print**

- Click on **Samples →Summaries of separate variables** to produce a dialog box shown below:

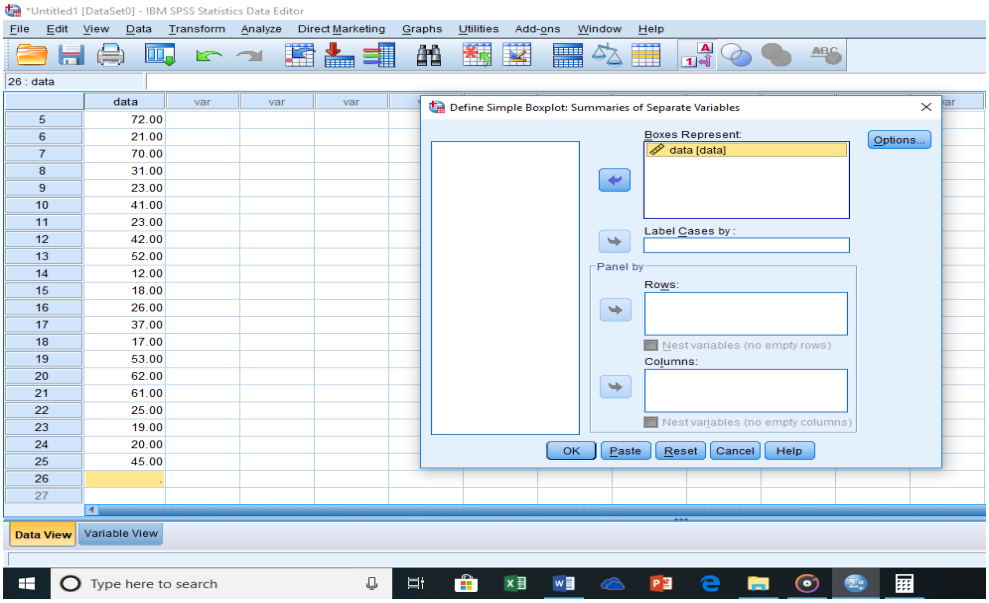


**Fig 4.9: SPSS Screen Print**

- Select the variable name **data** into **the Boxes Represent and click on OK to generate a boxplot shown below**

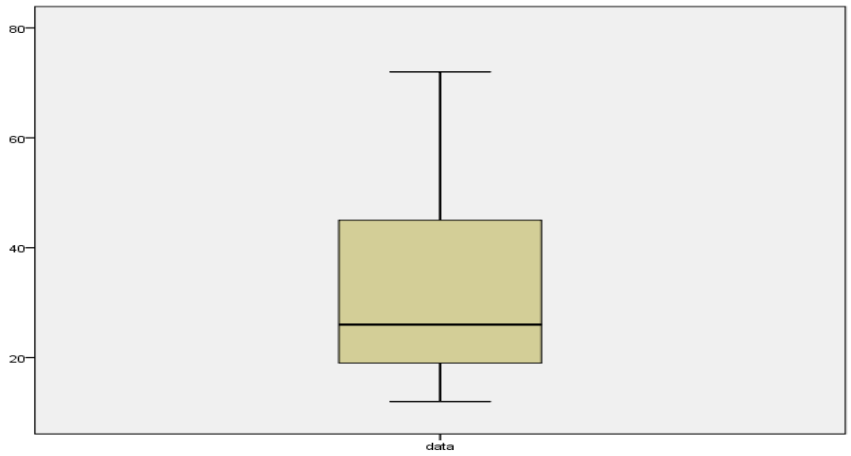


**Fig. 4.10: Boxplot of Patients Consulted in a Specialist Hospital**

From Fig. 4.10 above, the five values that are practically important in the construction of Boxplots are observed in ascending order as stated below:

$Minimum = 12$

$First\ Quartile\ (Q_1) = 18.5\ (25^{th}\ Percentile)$

$Median = 26$

$Third\ Quartile\ (Q_3) = 48.5\ (75^{th}\ Percentile$

$Maximum = 72$

The boxplot clearly shows that there are no outliers in the given data, as there are no dotted points outside the maximum and minimum values as indicated by the two extreme lines (The "whiskers").

## 4.7 Histogram

Histogram is a graph of frequency distribution. In histogram, a rectangle bar is drawn above each class interval such that the area of each rectangle is proportional to the frequency of observations falling in the corresponding interval. Unlike the bar chart, the bars of histogram are joined together with the base representing the width of the classes. A histogram gives the following pieces of information about any given observations:

- The spread of the observations. The histogram will have the shape of a bell (i.e. Symmetric) if the observations are evenly spread, otherwise the histogram will be skewed.
- It gives easy comparison of two frequency distributions
- It forms a basis for the approximation of frequency curve.
- It useful for the extrapolation of modal score.

It should be noted that a histogram is a graph of frequency distribution and should be drawn on a graph paper with adequate scales.

**Example 4.7:** Using the data of example 4.7, the histogram for the data is constructed from SPSS, using the class boundaries given below.

| Class Boundaries | Tallies | Frequency |
|---|---|---|
| 14.5 - 16.5 | IIII IIII IIII I | 16 |
| 16.5 - 18.5 | IIII IIII IIII IIII I | 21 |
| 18.5 - 20.5 | IIII IIII IIII IIII | 20 |
| 20.5 – 22.5 | IIII IIII IIII III | 18 |
| 22.5 – 24.5 | IIII IIII IIII | 15 |
| 24.5 – 26.5 | IIII I | 6 |
| 26.5 – 28.5 | III | 3 |
| 28.5 – 30.5 | I | 1 |

From the above table, it can be observed that the initial discrete classes have been made continuous through the computation of class boundaries before drawing the histogram. However, this action will not be necessary if the observations have been classified using continuous classes.

## 4.8 Frequency Polygon

The frequency polygon is prepared by locating the midpoint of class interval at the top of the histogram. These points are then connected by straight lines to create the polygon.

A smooth continuous curve drawn very close to the frequency polygon is called the frequency curve. In case of large sample, frequency curve is a good approximation to the population frequency curve called probability density function curve.

**Example 4.8**

The following are the salaries of 70 construction workers in thousands of naira

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 70 | 67 | 65 | 75 | 80 | 90 | 100 | 105 | 95 | 69 |
| 100 | 105 | 95 | 75 | 78 | 98 | 89 | 80 | 86 | 85 |
| 86 | 87 | 90 | 95 | 91 | 78 | 79 | 90 | 89 | 100 |
| 79 | 90 | 98 | 90 | 89 | 79 | 80 | 89 | 90 | 99 |
| 98 | 110 | 95 | 120 | 95 | 96 | 120 | 115 | 110 | 120 |
| 95 | 100 | 95 | 95 | 96 | 97 | 96 | 95 | 100 | 95 |
| 110 | 95 | 96 | 100 | 96 | 95 | 79 | 95 | 100 | 96 |

(a) Classify the data in the form of continuous frequency distribution taking the lowest class as 65 – 74.

(b) Draw the histogram of the frequency distribution and plot its frequency polygon.

**Solution:**

| Class | Tallies | Frequency |
|---|---|---|
| 65 – 74 | IIII | 4 |
| 74 – 83 | IIII IIII I | 11 |
| 83 – 92 | IIII IIII IIII | 14 |
| 92 – 101 | IIII IIII IIII IIII IIII IIII I | 31 |
| 101 – 110 | IIII | 4 |
| 110 – 119 | III | 3 |
| 119 – 128 | III | 3 |

(b) Histogram

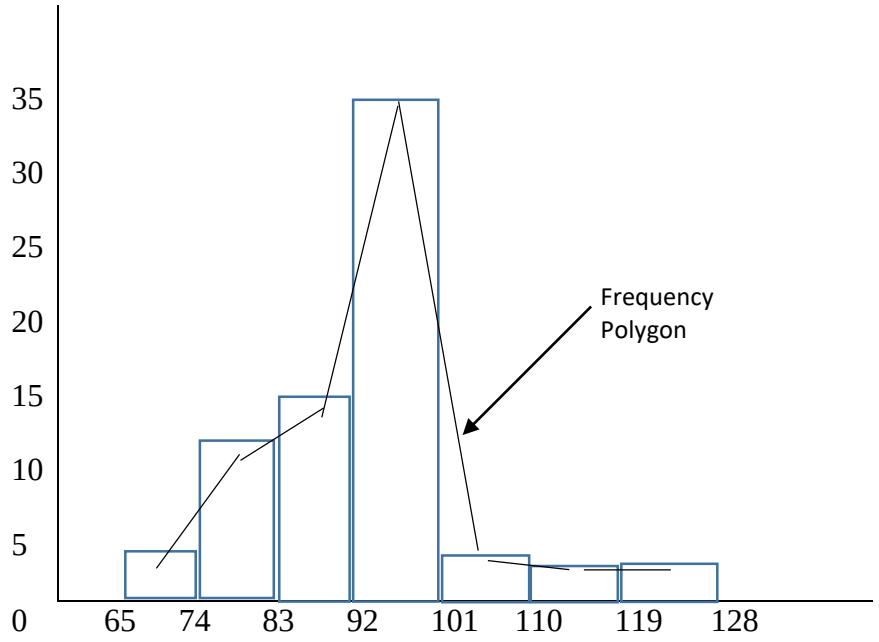


**Fig 4. 11: Histogram and Frequency Polygon of Construction workers Salaries**

## 4.9 Cumulative Frequency Curve (Ogive)

A way to further explore data is replacing a frequency distribution with a cumulative frequency distribution. In graphical term, the vertical scale (axis) represents the cumulative frequencies or relative cumulative frequencies while the horizontal scale represents the lower class boundary. Every Ogive starts on the left with relative or cumulative frequency of zero at the lower class boundary of the first class and ends on the right with a relative frequency of 100% or cumulative frequency equivalent to the total sum of the given frequencies.

Ogive can be used to calculate the First Quartile ($Q_1\ as\ \frac{N}{4}th\ demarcation\ of\ the\ ogive$), second quartile ($Q_2\ as\ \frac{N}{2}th\ demarcation\ of\ the\ ogive$) also known as the mean, Third quartile ($Q_3\ as\ \frac{3N}{4}th\ demarcation\ of\ the\ ogive$), inter-quartile range ($Q_3$ - $Q_1$) and semi inter-quartile range.

**Example 4.10:** Draw the cumulative frequency curve of the frequency distribution in example 4.9

**Solution:** below is the table of lower class boundaries (LCB) and cumulative frequency (CF) for Ogive

| LCB | 65 | 74 | 83 | 92 | 101 | 110 | 119 | 128 |
|---|---|---|---|---|---|---|---|---|
| CF | 0 | 4 | 15 | 29 | 60 | 64 | 67 | 70 |

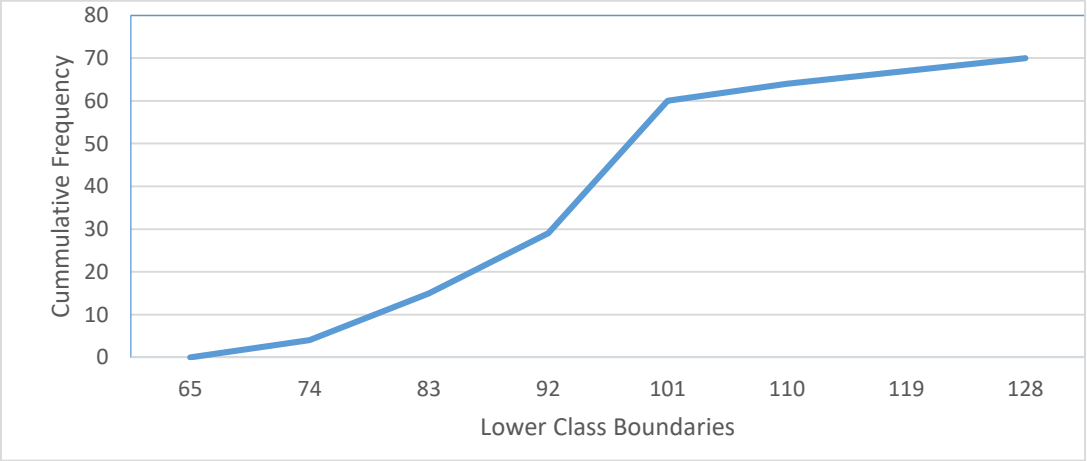


**Fig 4. 12: Ogive of Construction Workers**

**Exercises**

1. Diagrams are handy tools in the hands in the hands of a marketer. Discuss.
2. Explain the usefulness of diagrams in presenting statistical data.
3. List and state reasons, the kind of diagram(s) that would be more appropriate for representing the following categories of statistical information:
   (a) Number of academic staffs in the faculties of a University
   (b) Number of children by gender per few randomly selected family.
   (c) Monthly sales of three different items in a supermarket store.
   (d) Annual production in tones of few selected manufacturing companies.
   (e) Examination scores of the entire students in a class.

4. The table below shows the number of academic staffs in the five Schools at Federal Polytechnic Ilaro, Ogun State Nigeria as at October, 2019.

| Schools | Number of Academic Staffs |
|---|---|
| Pure and Applied Sciences | **84** |
| Management | **101** |
| Engineering | **88** |
| Environmental | **76** |
| Communication and Information Technology | **43** |

**Source: Academic and Physical Planning Unit, Federal Polytechnic Ilaro**

Represent the above information on the following diagrams

(a) Pie chart

(b) Bar chart

(c) Pictogram

5. Construct stem and leave for the following sets of data

(a)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 46 | 32 | 56 | 54 | 61 | 70 | 87 | 45 | 34 |
| 62 | 81 | 53 | 48 | 82 | 29 | 38 | 36 | 75 |
| 45 | 66 | 61 | 56 | 73 | 68 | 51 | 50 | 46 |

(b)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.9 | 1.2 | 1.5 | 2.4 | 2.5 | 3.7 | 2.7 | 4.5 |
| 5.5 | 2.7 | 2.3 | 1.4 | 0.6 | 2.8 | 3.4 | 4.2 |
| 5.8 | 5.1 | 3.2 | 5.1 | 1.9 | 1.7 | 3.8 | 5.4 |

6. Construct a boxplot for the observations listed below.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 126 | 16 | 17 | 72 | 21 | 70 | 31 | 23 | 41 | 23 | 42 | 52 |
| 12 | 18 | 26 | 37 | 17 | 53 | 62 | 61 | 25 | 19 | 20 | 45 | 164 |

7. Given the scores of 50 students in semester examination as follows

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 46 | 32 | 56 | 54 | 61 | 70 | 87 | 45 | 34 | 72 |
| 62 | 81 | 53 | 48 | 82 | 29 | 38 | 36 | 75 | 36 |
| 45 | 66 | 61 | 56 | 73 | 68 | 51 | 50 | 46 | 43 |
| 45 | 52 | 61 | 80 | 51 | 52 | 49 | 56 | 45 | 67 |
| 54 | 45 | 34 | 74 | 45 | 67 | 56 | 45 | 56 | 75 |

(a) Classify the data in the form of discrete frequency distribution taking the lowest class as 32 – 37.

(b) Draw the histogram of the frequency distribution and use the chart to extrapolate the modal score.

(c) Plot the distribution frequency polygon on the histogram drawn in (b).

(d) Draw the cumulative frequency curve and use it to determine $Q_1$, $Q_2$, $Q_3$, Quartile Deviation, Median score and $60^{th}$ percentile.

# Chapter Five

## 5.0 MEASURES OF LOCATION, PARTITION AND DISPERSION

### 5.1 Measures of Location

### 5.1.1 Arithmetic Mean

The aggregate (sum) of a set observations divided by the number of observations is the Arithmetic Mean. It is denoted by $\bar{X}$, read as X bar. $\bar{X}$ is also called the sample mean while the population mean (the average of all possible values) is denoted by Greek letter $\mu$ (miu). The population mean is a constant while the sample mean varies from sample to sample.

Let $x_1, x_2, x_3, \cdots, x_n$ be a set of observations on a variable $X$. The arithmetic mean of these observations can be expressed as

$$\bar{X} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (5.1)$$

For example, given the set of students' test scores as 6, 4, 5, 3, 7 and 5. The arithmetic mean is computed as

$$\bar{X} = \frac{6+4+5+3+7+5}{6} = \frac{30}{6} = 5$$

If the data set is such that $x_1, x_2, x_3, \cdots, x_n$ occur with frequencies $f_1, f_2, f_3, \cdots, f_n$, the sample mean will be expressed as

$$\bar{X} = \frac{f_1x_1+f_2x_2+f_3x_3+\cdots+f_n,x_n}{f_1+f_2+\cdots+f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=i}^{n} f_i} \qquad (5.2)$$

**Example 5.1**

The text scores of 10 undergraduates' students in a college are given as:

9, 14, 15, 10, 11, 12, 15, 16, 17, 18

Find the average score of the college students.

**Solution**

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{9+14+15+10+11+12+15+16+17+18}{10} = \frac{137}{10} = 13.7$$

**Example 5.2**

If the number of undergraduates in example 5.1 are increased to 25 and additional text scores given as follows: 14, 10, 16, 9, 12, 15, 9, 18, 17, 17, 14, 14, 15, 11, 11

Find the average score of the college students.

**Solution**

In the above example, the $x - values$ and $f - values$ are rewritten as shown in the frequency table given below:

**Table 5.1**

| $x$ | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 17 | 18 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $f$ | 3 | 2 | 3 | 2 | 4 | 4 | 2 | 3 | 2 | $\sum f = 25$ |
| $f\,x$ | 27 | 20 | 33 | 24 | 56 | 60 | 32 | 51 | 36 | $\sum_{i=1}^{n} f_i\, x_i = 339$ |

Thus,

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i\, x_i}{\sum_{i=i}^{n} f_i} = \frac{339}{25} = 13.56$$

For grouped data, in which the $x$ values have been classified into intervals without having the knowledge of the original data, the mean is computed as shown in example 5.3 given below.

**Example 5.3**

Compute the average salary of the construction workers in example 4.9

**Solution**

**Table 5.2**

| **Class** | **Frequency (f)** | **Mid-value ($x_i$)** | $\boldsymbol{x_i f_i}$ |
|---|---|---|---|
| 65 – 74 | 4 | 69.5 | 278 |
| 74 – 83 | 11 | 78.5 | 863.5 |
| 83 – 92 | 14 | 87.5 | 1225 |
| 92 – 101 | 31 | 96.5 | 2991.5 |
| 101 – 110 | 4 | 105.5 | 422 |
| 110 – 119 | 3 | 114.5 | 343.5 |
| 119 – 128 | 3 | 123.5 | 370.5 |
| **TOTAL** | **70** | | **6494** |

Thus,

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=i}^{n} f_i} = \frac{6494}{70} = 92.7714$$

## Assumed Mean Method of Computing Arithmetic Mean

The size of each $x - values$ can be reduced through the use of a value called assumed mean, to ease the efforts in arithmetic mean computation. In general, any number between the smallest and the largest mid-values can be used for assumed mean. It is however the best practice to use a number somewhere half way the mid-values. The assumed mean denoted by $'A'$ is then subtracted from every set of $x - values$ to obtain a set of coded values $'y'$. The mean of the $x - values$ is then obtained by adding $A$ to the mean of the coded values.
Thus, equation (5.3) is modified as

$$\bar{X} = A + \frac{\sum fy}{\sum f} \qquad (5.4)$$

$where\ y = x_i - A$

**Example 5.4**
Compute the average salary of the construction workers in example 4.9 using the assumed mean method.

**Solution**
Using assumed mean of 96.5, we compute $y = x_i - A$ as shown in the table below:

**Table 5.3**

| **Class** | $f$ | $x_i$ | $y = x_i - A$ | $fy$ |
|---|---|---|---|---|
| 65 – 74 | 4 | 69.5 | -27 | -108 |
| 74 – 83 | 11 | 78.5 | -18 | -198 |
| 83 – 92 | 14 | 87.5 | -9 | -126 |
| 92 – 101 | 31 | 96.5 | 0 | 0 |
| 101 – 110 | 4 | 105.5 | 9 | 36 |
| 110 – 119 | 3 | 114.5 | 18 | 54 |
| 119 – 128 | 3 | 123.5 | 27 | 81 |
| **TOTAL** | **70** | | | **-261** |

Thus,

$$\bar{X} = A + \frac{\sum fy}{\sum f} = 96.5 + \frac{(-261)}{70}$$

$$= 96.5 - 3.728571428571429 = 92.7714$$

## Points to Note about Arithmetic Mean

1. It has a smaller standard error than other measures of location.

2. It provides a good representation for all the data points.
3. The basic calculation is straightforward and it is easier to work with mathematically
4. It is very useful in further statistical work.
5. The distribution of sample mean will tend to be normally distributed even if the original data is not normal.
6. The sample mean is markedly affected by extreme observations (outliers) unlike other measures of location
7. It cannot be obtained graphically unlike other measures of location.
8. For nominal or ordinal data (qualitatively classified data), the computed mean value is of no statistical relevance.

### 5.1.2 Geometric Mean

The geometric mean of $n$ observations is the $n^{th}$ root of the product of all the observations. Given$x_1, x_2, x_3, \cdots, x_n$, the geometric mean is obtained by multiplying all the numbers together and obtain the $n^{th}$ root of the product.
Thus, Geometric mean ($\bar{X}_G$) is expressed as:

$$\bar{X}_G = \sqrt[n]{x_1.x_2.\cdots.x_n} = (x_1.x_2.x_3\cdots x_n)^{\frac{1}{n}} \quad (5.5)$$

For example, given the set of students' test scores as 6, 4, 5, 3, 7 and 5. The Geometric mean is

$$\bar{X}_G = \sqrt[6]{6.4.5.3.7.5} = \sqrt[6]{12600} = 4.823864396363730 \approx 4.82$$

Mathematically, if we take natural logarithm of (5.5), the geometric mean becomes

$$log\bar{X}_G = \frac{1}{n}\{logx_1 + logx_2 + \cdots + logx_n\} = \frac{\sum_{i=1}^{n} logX_i}{n} \quad (5.6)$$

If the data set $x_1, x_2, x_3, \cdots, x_n$ occur with frequencies $f_1, f_2, f_3, \cdots, f_n$, the geometric mean will be expressed as

$$\bar{X}_G = ({x_1}^{f_1}.{x_2}^{f_2}.{x_3}^{f_3}\cdots {x_n}^{f_n})^{\frac{1}{N}} \quad (5.7)$$

Where $N = f_1 + f_2 + f_3 + \cdots + f_n$

**Example 5.5**

1. The price of 50kg bag of rice has been on the increase consistently over a four year periods as follows:

| **Years** | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|
| **Price (#)** | 15,000 | 19,000 | 23,000 | 25,000 |

What is the average annual percentage increase in price?

**Solution**

$$Percentage\ Increase\ (P_{year}) = \frac{New-Old}{Old}$$

$$P_{2018} = \frac{19000-15000}{15000} = 0.2667$$

$$P_{2019} = \frac{23000-19000}{19000} = 0.2105$$

$$P_{2020} = \frac{25000-23000}{23000} = 0.0870$$

$$\bar{X}_G = [(1+P_{2018})x(1+P_{2019})x(1+P_{2020})]^{\frac{1}{3}}$$

$$= [(1+0.2667)x(1+0.2105)x(1+0.0870)]^{\frac{1}{3}}$$

$$= [(1.2667)x(1.2105)x(1.0870)]^{\frac{1}{3}} = \sqrt[3]{1.6667} = 1.1856$$

$$Average\ annual\ percentage\ increase = 1.1856 - 1$$

$$= 0.1856x100 = 18.56\%$$

## Points to Note about Geometric Mean

1. The geometric mean is usually less than the arithmetic mean
2. The value of geometric mean is not influenced by extreme values to the same extent as the arithmetic mean
3. Geometric mean is only useful for positive numbers
4. It is most useful for averaging ratios and determination of percentages increase/decrease.
5. The concept of geometric mean is used in the construction of index number and population growth rates.

## 5.1.3 Harmonic Mean

Harmonic mean is the reciprocal of the arithmetic mean of reciprocals of a set of observations. Given set of observations $x_1, x_2, x_3, \cdots, x_n$, the harmonic mean of $x$ denoted by $\bar{X}_H$ is expressed as

$$\bar{X}_H = \frac{1}{\frac{1}{n}\left[\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}\cdots+\frac{1}{x_n}\right]} \quad (5.8)$$

**Example 5.6**

Find the Harmonic mean of a set of students' test scores given as 6, 4, 5, 3, 7 and 5

**Solution**

$$\bar{X}_H = \frac{1}{\frac{1}{6}\left[\frac{1}{6}+\frac{1}{4}+\frac{1}{5}+\frac{1}{3}+\frac{1}{7}+\frac{1}{5}\right]} = \frac{840}{181} = 4.641$$

## Points to Note about Harmonic Mean

1. The harmonic mean is at most the value of both the geometric and arithmetic mean i.e $\bar{X}_H \leq \bar{X}_G \leq \bar{X}$
2. The value of geometric mean is not influenced by extreme values to a significant extent.
3. The harmonic mean of any data set cannot be calculated if it has zero value

## 5.1.4 The Median

Median is the value that lies half way amidst all the given observations arranged in ascending order. If the number of ordered observations is even, median is the average of the two middle values and when the number of ordered observations $(n)$ is odd, the median is the $\frac{1}{2}(n+1)th$ observation, which is simply the middle value.

**Example 5.8**

Find the median of a set of students' test scores given as 6, 4, 5, 3, 7 and 5.

**Solution**

First, we arrange the test scores in ascending order as: 3, 4, 5, 5, 6, 7. Then median is the average value of the two middle observations.

$$Median = \frac{5+5}{2} = 5$$

Example of median when data point is odd: 3, 7, 4, 5, 9, 7, 6, 2, 11

**Solution**

First, we arrange the test scores in ascending order as: 2, 3, 4, 5, 6, 7, 7, 9, 11

Thus, $\frac{1}{2}(n+1)th = \frac{1}{2}(9+1)th = 5th\ observation = 6$

For grouped data, in which the $x$ values have been classified into intervals without having the knowledge of the original data, the median is obtained as the $\frac{N}{2}th$ observation irrespective of whether the number of observations is even or odd. The first step is to identify the class in which the median falls, after which the exact median value is computed using the expression given as:

$$Median = L_c + \left(\frac{\frac{N}{2} - CF_b}{f_m}\right) x\ C \qquad (5.11)$$

Where

$L_c$ is the Lower Class boundary of the median class
$N$ is the Total frequency
$CF_b$ is the Cumulative frequency just before the median class
$f_m$ is the frequency of the median class
$C$ is the class size of the median class

**Example 5.9**
Compute the median salary of the construction workers in example 4.9.

**Solution**

**Table 5.4**

| **Class** | $f$ | $x_i$ | **Class Boundary** | CF |
|---|---|---|---|---|
| 65 – 74 | 4 | 69.5 | 65 – 74 | 4 |
| 74 – 83 | 11 | 78.5 | 74 – 83 | 15 |
| 83 – 92 | 14 | 87.5 | 83 – 92 | 29 |
| 92 – 101 | 31 | 96.5 | 92 – 101 | 60 |
| 101 – 110 | 4 | 105.5 | 101 – 110 | 64 |
| 110 – 119 | 3 | 114.5 | 110 – 119 | 67 |
| 119 – 128 | 3 | 123.5 | 119 – 128 | 70 |
| **TOTAL** | **70** | | | |

The median class is $\frac{70}{2}th = 35th\ observation = (92 - 101)\ class$

Thus,

$L_c = 92 \quad N = 70 \quad CF_b = 29 \quad f_m = 31 \quad C = 9$

$$Median = L_c + \left(\frac{\frac{N}{2} - CF_b}{f_m}\right) x\ C = 92 + \left(\frac{35 - 29}{31}\right) x\ 9 = 93.7419$$

**Example 5.10**
Find the median mark of observations classified in the table below.

| **Class** | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 | 80-89 |
|---|---|---|---|---|---|---|---|
| **Frequency** | 3 | 4 | 8 | 11 | 9 | 6 | 4 |

**Solution**

**Table 5.5**

| Class | $f$ | Class Boundary | CF |
|---|---|---|---|
| 20-29 | 3 | 19.5-29.5 | 3 |
| 30-39 | 4 | 29.5-39.5 | 7 |
| 40-49 | 8 | 39.5-49.5 | 15 |
| 50-59 | 11 | 49.5-59.5 | 26 |
| 60-69 | 9 | 59.5-69.5 | 35 |
| 70-79 | 6 | 69.5-79.5 | 41 |
| 80-89 | 4 | 79.5-89.5 | 45 |

The median class is $\frac{45}{2}th = 22.5th\ observation = (50-59)\ class$

Thus,

$L_c = 49.5 \quad N = 45 \quad CF_b = 15 \quad f_m = 11 \quad C = 10$

$Median = 49.5 + \left(\frac{22.5-15}{11}\right) x\ 10 = 56.3182$

## Estimating the Median from Histogram and Ogive

The process of estimating median from histogram amounts to finding the value of $x$-axis that corresponds to 50% of the total area of the histogram. The total area of histogram is obtained as a product of total frequency and class interval. That is, Total area of histogram = Total frequency x class interval.

Ogive allows for the estimation of median much more readily. All that is required is to simply read the abscissa-value corresponding to the ordinate value of $\frac{N}{2}$.

## Illustration

Using the histogram and Ogive of example 5.9, to estimate the median salary.

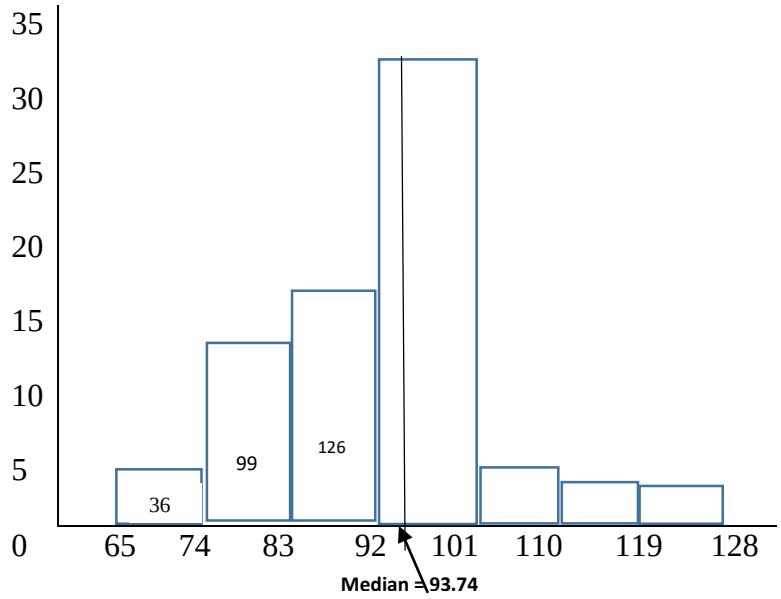


**Figure 5.1: Histogram of Students marks**

**Median Estimation from Histogram**:

Total Area of Histogram = Total frequency x Class interval $= 70x9 = 630$.

And $50\%\ of\ the\ total\ area = 315$.

$$Thus, the\ sum\ of\ areas\ before\ the\ median\ class$$

$$= 36 + 99 + 126 = 261 square\ unit.$$

The area to be added to 261 to make it 315 is 54 which is far less than the area of the next bar. That is, $315 - 261 = 54$.

$Thus, the\ remaining\ area\ to\ estimate\ the\ Median = \frac{54}{31} = 1.74.$

$Therefore, Median = 92 + 1.74 = 93.74$

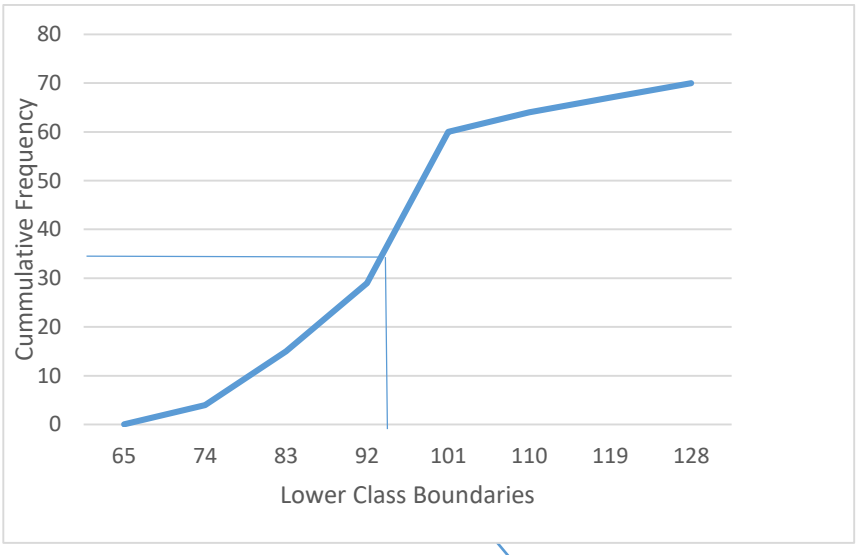


**Fig 5.2: Ogive of Students marks**

## Points to Note about Median

1. Median is not influenced by extreme values
2. Median result is not useful for further statistical work except in Non-parametric statistical analysis.
3. Median can be read from either the histogram or the cumulative frequency curve.
4. In population showing skewed distribution, median is a more representative measure of location than the arithmetic mean.
5. The median will always be equal to the mean for a symmetrical data.
6. The median will be greater than the arithmetic mean if the data is negatively skewed and for positively skewed data, it will be less than the mean.
7. Median value is equal to that of $2^{nd}$ quartile $(Q_2)$, $50^{th}$ Percentile$(P_{50})$ and $5^{th}$ Deciles $(D_5)$

## 5..1.5 The Mode

Mode is the value which occurs most in a given set of observations. That is, the value(s) with the highest frequency. It is a measure of relative standing which can be used for both qualitative and quantitative data. For frequency distributed data, the modal class is the class containing the highest frequency.

For example, the mode of a set of students' test scores given as 6, 4, 5, 3, 7 and 5 is 5. This is because it appears twice, more than any other values in the data set.

For grouped data, the first step is to obtain the modal class. Let this modal class be designated as $1\ to\ 2$ and its frequency denoted as $f$, thus the frequency for the class just before the modal class is $f_1$ and the one for the class immediately after is $f_2$. Then mode is calculated using the expression:

$$Mode = L_c + \left(\frac{f - f_1}{(f - f_1) + (f - f_2)}\right) C$$

If we denote $f - f_1$ by $\Delta_1$ and $f - f_2$ by $\Delta_2$, then

$$Mode = L_c + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) C \qquad (5.12)$$

Where

$L_c$ is the Lower Class boundary of the modal class

$C$ is the class size of the modal class

**Example 5.11**

Compute the modal salary of the construction workers in example 5.9

The modal class is $(92 - 101)\ class$.

Thus,

$L_c = 92\ \ f = 31 \quad f_1 = 14 \quad f_2 = 4 \quad C = 9 \quad \Delta_1 = 31 - 14 = 17 \quad \Delta_2 = 31 - 4 = 27$

$$Mode = 92 + \left(\frac{17}{17+27}\right) x\ 9 = 95.4773$$

**Example 5.12**

Find the modal mark of observations in example 5.10.

**Solution**

The modal class $= (50 - 59)\ class.$ Thus,

$L_c = 49.5\ \ f = 11 \quad f_1 = 8 \quad f_2 = 9 \quad C = 10 \quad \Delta_1 = 11 - 8 = 3 \quad \Delta_2 = 11 - 9 = 2$

$$Mode = 49.5 + \left(\frac{3}{3+2}\right) x\ 10 = 55.5$$

## Estimating the Mode from Histogram

The procedures are as follows:

1. Locate the tallest triangle i.e. the bar belonging to the modal class
2. Draw two diagonal lines from the tallest bar to the edges of the bars immediately to the right and left of the former.
3. Draw a vertical line through the intersection.
4. The point where the vertical line meets the horizontal axis is the mode.

**Illustration:** Draw the histogram of example 5.9 and use it to estimate the modal mark.

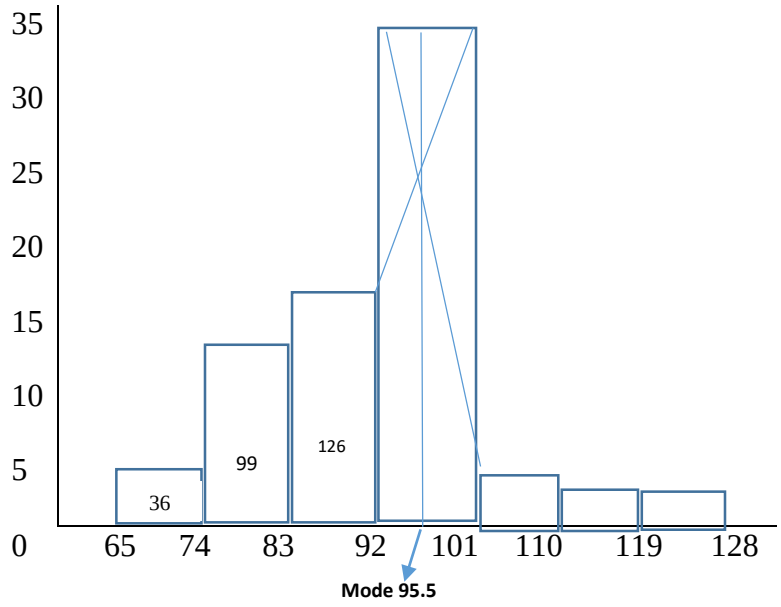


**Fig 5.3: Estimation of Mode from Histogram**

### Points to Note about Mode

1. Mode can be read accurately from the histogram
2. Mode is often an actual value and may therefore appear to be realistic.
3. Mode is not affected by extreme values
4. It is the only measure of location that can be used with nominal data.
5. It is not useful for further statistical work
6. There may be more than one mode in a given set of observation.

## 5.2 Measure of Partition

### 5.2.1 Deciles

Deciles of an ordered statistics is the required position multiply by total number observations and divided by 10. Thus, Deciles divide a given set of observations into ten parts with nine boundaries which partitioned the observations into $D_1,\ D_2, \cdots, D_9$.

For example 5th Deciles ($D_5$) which is also the same as the median would be taken as $\frac{5}{10}N^{th}$ observation of the ordered data set and 6th Deciles ($D_6$) as $\frac{6}{10}N^{th}$ observation of the ordered data set. The Deciles range or Inter-Deciles is computed as $D_9 - D_1$.

For grouped data, in which the $x$ values have been classified into intervals without having the knowledge of the original data, the $i^{th}$ Deciles is calculated using the expression:

$$D_i = L_{D_i} + \left(\frac{\frac{i}{10}N - CF_b}{f_{D_i}}\right) C \qquad i = 1, 2, 3, \cdots, 9 \tag{5.13}$$

Where

$L_{D_i}$ is the Lower Class boundary of the $i^{th}$ Deciles class

$N$ is the Total frequency

$CF_b$ is the Cumulative frequency just before the $i^{th}$ Deciles class

$f_{D_i}$ is the frequency of the $i^{th}$ Deciles class

$C$ is the class size of the $i^{th}$ Deciles class

**Example 5.13**

Find the 6th Deciles ($D_6$) of observations in example 5.10.

**Solution**

**Table 5.5 (Recalled)**

| Class | $f$ | Class Boundary | CF |
|---|---|---|---|
| 20-29 | 3 | 19.5-29.5 | 3 |
| 30-39 | 4 | 29.5-39.5 | 7 |
| 40-49 | 8 | 39.5-49.5 | 15 |
| 50-59 | 11 | 49.5-59.5 | 26 |
| 60-69 | 9 | 59.5-69.5 | 35 |
| 70-79 | 6 | 69.5-79.5 | 41 |
| 80-89 | 4 | 79.5-89.5 | 45 |

The 6th Deciles class = $\frac{6}{10} x\ 45th = 27th\ observation = (60-69)\ class$

Thus,

$$D_6 = L_{D_6} + \left(\frac{\frac{6}{10}N - CF_b}{f_{D_6}}\right) C$$

$L_{D_6} = 59.5 \qquad N = 45 \qquad CF_b = 26 \qquad f_{D_6} = 9 \qquad C = 10$

$$D_6 = 59.5 + \left(\frac{27-26}{9}\right) x\ 10 = 60.6111$$

## 5.2.2 Quartiles

A Quartile involved portioning a total frequency or a given set of observation into four equal parts with four boundaries, which partitioned the observations into first quartile $(Q_1)$, $second\ quartile\ (Q_2)\ and\ third\ quartile\ (Q_3)$.

The first quartile $(Q_1)$ is the $\frac{N}{4}th$ observation, second quartile $(Q_2)$ is the $\frac{N}{2}th$ observation while the third quartile $(Q_3)$ is the $\frac{3N}{4}th$ observation of the ordered data set. The inter-quartile range is computed as $Q_3 - Q_1$.

For grouped data, in which the $x$ values have been classified into intervals without having the knowledge of the original data, the $i^{th}(i = 1, 2, 3)$, Quartiles is calculated using the expression:

$$Q_i = L_{Q_i} + \left(\frac{\frac{i}{4}N - CF_b}{f_{Q_i}}\right) C \quad i = 1, 2, 3 \qquad (5.14)$$

Where

$L_{Q_i}$ is the Lower Class boundary of the $i^{th}$ quartile class

$N$ is the Total frequency

$CF_b$ is the Cumulative frequency just before the $i^{th}$ quartile class

$f_{Q_i}$ is the frequency of the $i^{th}$ quartile class
$C$ is the class size of the $i^{th}$ quartile class

**Example 5.14**

**(a)** Find the 1st, 2nd and 3rd Quartiles of observations in example 6.10.

**(b)** Use the results in (a), find the Quartile Deviation (Interquartile Range) and Semi-interquartile Range.

**Solution**

**Table 5.5 (Recalled)**

| Class | $f$ | Class Boundary | CF |
|---|---|---|---|
| 20-29 | 3 | 19.5-29.5 | 3 |
| 30-39 | 4 | 29.5-39.5 | 7 |
| 40-49 | 8 | 39.5-49.5 | 15 |
| 50-59 | 11 | 49.5-59.5 | 26 |
| 60-69 | 9 | 59.5-69.5 | 35 |
| 70-79 | 6 | 69.5-79.5 | 41 |
| 80-89 | 4 | 79.5-89.5 | 45 |

(a) The 1st Quartile ($Q_1$) class = $\frac{45}{4}th = 11.25th\ observation = (40-49)\ class$. Thus,

$$Q_1 = L_{Q_1} + \left(\frac{\frac{1}{4}N - CF_b}{f_{Q_1}}\right) C$$

Where $L_{Q_1} = 39.5 \quad N = 45 \quad CF_b = 7 \quad f_{Q_1} = 8 \quad C = 10$

$$Q_1 = 39.5 + \left(\frac{11.25 - 7}{8}\right) x\ 10 = 44.8125$$

The 2nd Quartile ($Q_2$) class = $\frac{45}{2}th = 22.5th\ observation$

$$= (50-59)\ class$$

Thus,

$$Q_2 = L_{Q_2} + \left(\frac{\frac{2}{4}N - CF_b}{f_{Q_2}}\right) C$$

Where $L_{Q_2} = 49.5 \quad N = 45 \quad CF_b = 15 \quad f_{Q_2} = 11 \quad C = 10$

$$Q_2 = 49.5 + \left(\frac{22.5 - 15}{11}\right) x\ 10 = 56.3182$$

The 3rd Quartile ($Q_3$) class = $\frac{3x45}{4} th = 33.75th\ observation$

$$= (60 - 69)\ class$$

Thus,

$$Q_3 = L_{Q_3} + \left(\frac{\frac{3}{4}N - CF_b}{f_{Q_3}}\right) x\ C$$

Where $L_{Q_3} = 59.5 \quad N = 45 \quad CF_b = 26 \quad f_{Q_3} = 9 \quad C = 10$

$$Q_3 = 59.5 + \left(\frac{33.75 - 26}{9}\right)\ x\ 10 = 68.1111$$

(b) Quartile Deviation (Interquartile Range) and Semi-interquartile Range.

$$Quartile\ Deviation = \ Q_3 - Q_1 = 68.1111 - 44.8125 = 23.2986$$

$$Semi - interquartile\ range\ = \frac{1}{2}\ (Q_3 - Q_1) = \frac{1}{2}(68.1111 - 44.8125) = \frac{23.2986}{2}$$
$$= 11.6493$$

### 5.2.3 Percentiles

Percentiles of an ordered statistics is the required position multiply by total number observations and divided by 100. Thus, Percentiles divide a given set of observations into hundred parts with ninety-nine boundaries which partitioned the observations into $P_1,\ P_2,\ P_3 \cdots, P_{99}$.
For example 50th Percentiles ($P_{50}$) which is also the same as the median would be taken as $\frac{50}{100} N^{th}$ observation of the ordered data set and 90th Percentiles ($P_{90}$) as $\frac{90}{100} N^{th}$ observation of the ordered data set. The Inter-percentile range is computed simply as $P_{90} - P_{10}$.

For grouped data, in which the $x$ values have been classified into intervals without having the knowledge of the original data, the $i^{th}$ Percentiles is calculated using the expression:

$$P_i = L_{P_i} + \left(\frac{\frac{i}{100}N - CF_b}{f_{P_i}}\right)\ C \quad i = 1, 2, 3, \cdots, 99 \qquad (5.15)$$

Where

$L_{P_i}$ is the Lower Class boundary of the $i^{th}$ Percentile class
$N$ is the Total frequency
$CF_b$ is the Cumulative frequency just before the $i^{th}$ Percentile class
$f_{P_i}$ is the frequency of the $i^{th}$ Percentile class

$C$ is the class size of the $i^{th}$ Percentile class

**Example 5.15**
Find the 25th and 75th Percentiles of observations in example 5.10.

**Solution**

**Table 5.5 (Recalled)**

| Class | $f$ | Class Boundary | CF |
|---|---|---|---|
| 20-29 | 3 | 19.5-29.5 | 3 |
| 30-39 | 4 | 29.5-39.5 | 7 |
| 40-49 | 8 | 39.5-49.5 | 15 |
| 50-59 | 11 | 49.5-59.5 | 26 |
| 60-69 | 9 | 59.5-69.5 | 35 |
| 70-79 | 6 | 69.5-79.5 | 41 |
| 80-89 | 4 | 79.5-89.5 | 45 |

The 25th Percentile class = $\frac{25}{100} x\ 45th = 11.25th\ observation = (40-49)\ class$.Thus,

$$P_{25} = L_{P_{25}} + \left(\frac{\frac{25}{100}N - CF_b}{f_{P_{25}}}\right) x\ C$$

Where $L_{P_{25}} = 39.5 \quad N = 45 \quad CF_b = 7 \quad f_{P_{25}} = 8 \quad C = 10$

$$P_{25} = 39.5 + \left(\frac{11.25-7}{8}\right) x\ 10 = 44.8125$$

The 75th Percentile class = $\frac{75}{100} x\ 45th = 33.75th\ observation = (60-69)\ class$

Thus,

$$P_{75} = L_{P_{75}} + \left(\frac{\frac{75}{100}N - CF_b}{f_{P_{75}}}\right) x\ C$$

Where $L_{P_{75}} = 59.5 \quad N = 45 \quad CF_b = 26 \quad f_{P_{75}} = 9 \quad C = 10$

$$P_{75} = 59.5 + \left(\frac{33.75-26}{9}\right) x\ 10 = 68.1111$$

## Estimating the Deciles, Quartiles and Percentiles from Histogram and Ogive

The process of estimating the three measures from histogram amounts to finding the value of $x$-axis which corresponds respectively to the areas covered by $D_i$, $Q_i$ and $P_i$ on the histogram total

area. The total area of histogram is obtained as a product of total frequency and class interval. That is, Total area of histogram = Total frequency x class interval.
Ogive allows for the estimation of the three measures much more readily. All that is required is to simply read the abscissa-value corresponding to the ordinate value of $D_i$, $Q_i$ and $P_i$ respectively.

## Illustration

Draw the Ogive and histogram of example 5.10 and use them to estimate the following:

(a) 6th Deciles ($D_6$)

(b) 3rd Quartile

(c) 25th Percentile

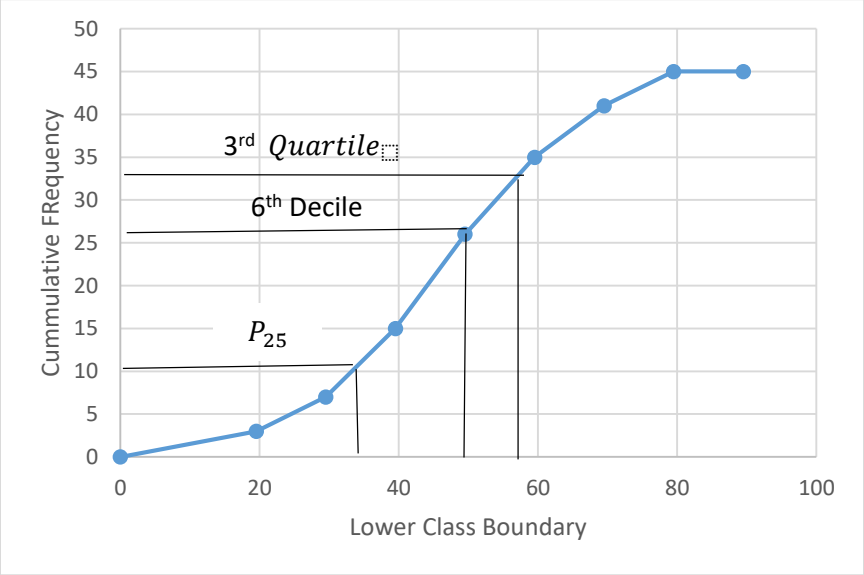


Fig. 5.4: Estimation of Deciles, Quartile and Percentile from Ogive

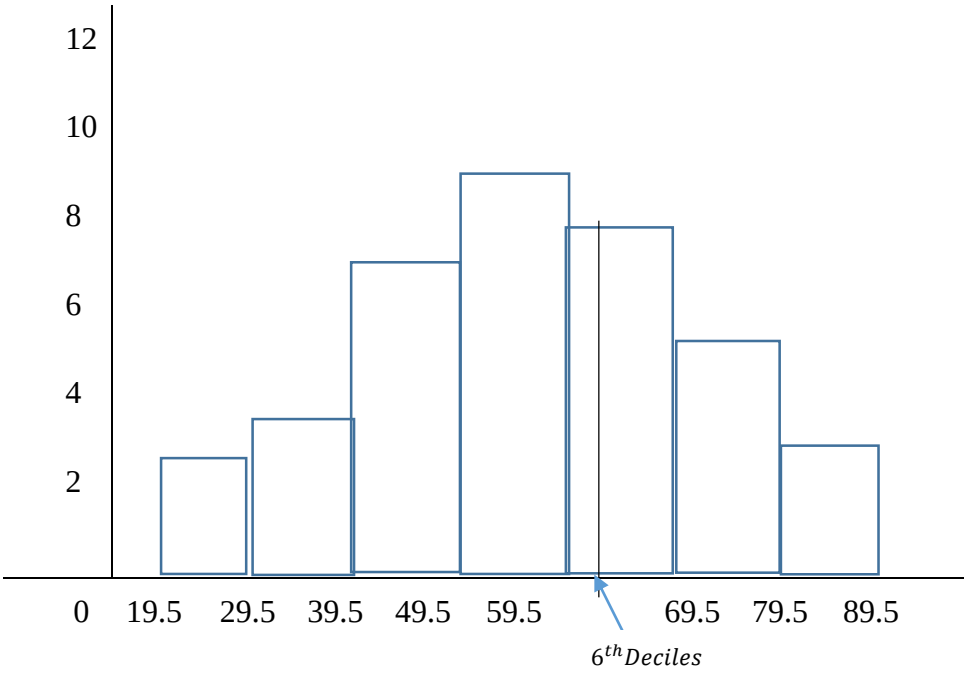


**Fig. 5.5: Estimation of Deciles, Quartile and Percentile from Histogram**

**Deciles Estimation from Histogram**: Total Area of Histogram = Total frequency x Class interval $= 45x10 = 450.$

And $\frac{6}{10}$ $of\ the\ total\ area = 270.$ Thus,

$Sum\ of\ areas\ of\ the\ classes\ before\ the\ \ 6^{th}\ Deciles\ class$

$= 30 + 40 + 80 + 110 = 260 square\ unit.$

The area to be added to $260$ to make it $270$ is $10$ which less than the area of the next bar. That is, $270 - 260 = 10.$

Thus, the remaining area to estimate $6^{th}\ Deciles = \frac{10}{9} = 1.11$

$Therefore, 6^{th}\ Deciles = 59.5 + 1.11 = 60.61$

Using the above procedures executed for the estimation of 6th Deciles, we can estimate the 3rd Quartile and 25th percentile as well.

## 5. 3 Measures of Dispersion

## 5.3.1Range

The simplest measure of variation is the range. It is the difference between the largest and the smallest observations in a sample. That is

$Range = maximum\ value - minimum\ value$ (5.16)

For example, given the set of students' test scores as 6, 4, 5, 3, 7 and 5. The range is $7 - 3 = 4.$

For another set of scores given as 7, 8, 1, 2, 5 and 6. The range is $8 - 1 = 7$. The results implied that the latter set of scores is more dispersed than the former.

### Points to Note about Range

1. Range may be considerably changed if either of the extreme values happens to drop out, while the removal of any other value would not be of any effect since it does not take into account the entire observations.
2. Range does not give any clues about the distribution of values in the observations relative to measures of location.
3. Range cannot be computed in case of open-ended classes of observations.

## 5.3.2 Mean Deviation

Mean deviation is obtained by regarding all the deviations as positive irrespective of their signs. Thus, Mean Deviation of a set of observations $x_1,\ x_2, \cdots, x_n$ is the mean of the absolute deviations from the mean as denoted by

$$Mean\ Deviation = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{5.17}$$

For grouped data,

$$Mean\ Deviation = \frac{\sum_{i=1}^{n} f|x_i - \bar{x}|}{\sum f} \tag{5.18}$$

**Example 5.16**
The text scores of 10 undergraduates' students in a college are given as:
9, 14, 15, 10, 11, 12, 15, 16, 17, 18

Find the mean deviation of the text scores.

**Solution**

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{9 + 14 + 15 + 10 + 11 + 12 + 15 + 16 + 17 + 18}{10} = \frac{137}{10} = 13.7$$

| $x_i$ | 9 | 14 | 15 | 10 | 11 | 12 | 15 | 16 | 17 | 18 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i - \bar{x}$ | -4.7 | 0.3 | 1.3 | -3.7 | -2.7 | -1.7 | 1.3 | 2.3 | 3.3 | 4.3 | 0 |
| $\|x_i - \bar{x}\|$ | 4.7 | 0.3 | 1.3 | 3.7 | 2.7 | 1.7 | 1.3 | 2.3 | 3.3 | 4.3 | 25.6 |

Thus,

$$Mean\ Deviation\ (MD) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{25.6}{10} = 2.56$$

### Example 5.17

Compute the average salary of the construction workers in example 4.9

**Solution**

**Table 5.2 (Recalled)**

| **Class** | **F** | $x_i$ | $\boldsymbol{x_i f_i}$ | $x_i - \bar{x}$ | $\lvert x_i - \bar{x}\rvert$ | $f\lvert x_i - \bar{x}\rvert$ |
|---|---|---|---|---|---|---|
| 65 – 74 | 4 | 69.5 | 278 | -23.3 | 23.3 | 93.2 |
| 74 – 83 | 11 | 78.5 | 863.5 | -14.3 | 14.3 | 157.3 |
| 83 – 92 | 14 | 87.5 | 1225 | -5.3 | 5.3 | 74.2 |
| 92 – 101 | 31 | 96.5 | 2991.5 | 3.7 | 3.7 | 114.7 |
| 101 – 110 | 4 | 105.5 | 422 | 12.7 | 12.7 | 50.8 |
| 110 – 119 | 3 | 114.5 | 343.5 | 21.7 | 21.7 | 65.1 |
| 119 – 128 | 3 | 123.5 | 370.5 | 30.7 | 30.7 | 92.1 |
| **TOTAL** | **70** | | **6494** | | | 647.4 |

Thus,

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=i}^{n} f_i} = \frac{6494}{70} = 92.7714 \approx 92.8$$

$$Mean\ Deviation = \frac{\sum_{i=1}^{n} f|x_i - \bar{x}|}{\sum f} = \frac{647.4}{70} = 9.2486 \approx 9.3$$

## 5.3.3 Median Deviation

In the computation of median deviation, mean ($\bar{x}$) is replaced by median (M).Thus, the median deviation is defined by

$$Median\ Deviation = \frac{\sum_{i=1}^{n} |x_i - M|}{n} \qquad (5.19)$$

## 5.3.4 Coefficient of Mean Deviation

The coefficient of mean deviation otherwise known as relative dispersion is computed by dividing the mean deviation by that measure of central tendency about which deviations are recorded. Thus,

$$Coefficient\ of\ Mean\ Deviation = \frac{Mean\ Deviation}{Mean} x100 \qquad (5.20)$$

$$Coefficient\ of\ Median\ Deviation = \frac{Median\ Deviation}{Median} x100 \quad (5.21)$$

Thus, the coefficient of mean deviation of example 7.2 is computed as

$$Coefficient\ of\ MD = \frac{M\ D}{\bar{x}} x100 = \frac{9.3}{92.8} x100 = 10.02\%$$

### Points to Note about Mode

1. Its computation is simple as compared to standard deviation
2. It is less affected by extreme values as compared to standard deviation
3. It is better than range or quartile deviation in terms of data utilization, since it is based on all values in the distribution.
4. It is not suitable for further statistical computations.

## 5.3.5 Standard Deviation and Variance

Standard deviation is defined as the square root of the mean of the squares of the deviations of individual observations from their arithmetic mean. The population standard deviation is denoted as $\sigma$ (sigma) and the sample deviation is denoted by $s$. Thus, population standard deviation is expressed as

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}} \quad (5.22)$$

Equation (5.22) can be simplified as

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \bar{x}^2} \quad (5.23)$$

For grouped data, population and sample standard deviation are expressed as

$$\sigma = \sqrt{\frac{\sum f(x_i - \bar{x})^2}{\sum f}} \quad (5.24)$$

$$s = \sqrt{\frac{\sum f(x_i - \bar{x})^2}{\sum f - 1}} \quad (5.25)$$

Equation (5.24) and (5.25) can further be simplified as

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} \quad (5.26)$$

$$s = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx_i)^2}{\sum f}}{\sum f - 1}} \quad (5.27)$$

Let $x_1,\ x_2, \cdots, x_n$ be a set of $n$ observations with sample mean $\bar{x}$, the sample standard deviation is expressed as

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{5.28}$$

Equation (5.28) can further be simplified as

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} \tag{5.29}$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x_i)^2}{n}}{n-1}} \tag{5.30}$$

If $N_1$ and $N_2$ represent the number of items in two different data sets with their means and standard deviations given as $\bar{x}_1, \bar{x}_2$ and $\sigma_1, \sigma_2$ respectively, then the standard deviation of the combined distribution is given as

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{X}_c - \bar{x}_1)^2 + N_2(\bar{X}_c - \bar{x}_2)^2}{N_1 + N_2}} \tag{5.31}$$

Where $\bar{X}_c$ is as defined in equation (5.31)

## 5.3.6 Variance

Variance is the square of standard deviation and is denoted as $\sigma^2$. Variance is more often specified than the standard deviation and it can be said to have the same properties with the standard deviation. Thus all the equations expressed as (7.7) to (7.15) without the square root, also hold for the computations of variance denoted as $\sigma^2$.

When several variances with different number of observations are given for $k$ items, a pooled variance for $k$ number of items can be obtained using the expression:

$$\sigma_{pooled}^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2 + \cdots + n_k\sigma_k^2}{n_1 + n_2 + \cdots + n_k} \tag{5.32}$$

**Example 5.18**

Find the standard deviation, variance and sample variance of the data in example 5.17.

| $x_i$ | 9 | 14 | 15 | 10 | 11 | 12 | 15 | 16 | 17 | 18 | $\sum$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i - \bar{x}$ | -4.7 | 0.3 | 1.3 | -3.7 | -2.7 | -1.7 | 1.3 | 2.3 | 3.3 | 4.3 | 0 |
| $(x_i - \bar{x})^2$ | 22.09 | 0.09 | 1.69 | 13.69 | 7.29 | 2.89 | 1.69 | 5.29 | 10.89 | 18.49 | 84.1 |
| $x_i^2$ | 81 | 196 | 225 | 100 | 121 | 144 | 225 | 256 | 289 | 324 | 1961 |

**Standard Deviation**

$$\sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}} = \sqrt{\frac{84.1}{10}} = \sqrt{8.41} = 2.9$$

Alternatively,

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \bar{x}^2} = \sqrt{\frac{1961}{10} - 13.7^2} = \sqrt{8.41} = 2.9$$

**Variance**

$$\sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{84.1}{10} = 8.41$$

Alternatively,

$$\sigma^2 = \frac{\sum x^2}{N} - \bar{x}^2 = \frac{1961}{10} - 13.7^2 = 8.41$$

**Sample Variance**

$$s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{84.1}{9} = 9.3$$

Alternatively,

$$s^2 = \frac{\sum x^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{1961-\frac{137^2}{10}}{9} = \frac{84.1}{9} = 9.3$$

## Example 5.19

Find the standard deviation, variance and sample variance of the data solved in example 5.2.

**Solution**

**Table 5.2 (Recalled)** $\bar{X}$ = 92.77142857

| Class | $f$ | $x_i$ | $x_i^2$ | $x_i f_i$ | $f x_i^2$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $f(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|---|---|
| 65 – 74 | 4 | 69.5 | 4830.25 | 278 | 19321 | -23.3 | 542.89 | 2171.56 |
| 74 – 83 | 11 | 78.5 | 6162.25 | 863.5 | 67784.75 | -14.3 | 204.49 | 2249.39 |
| 83 – 92 | 14 | 87.5 | 7656.25 | 1225 | 107187.5 | -5.3 | 28.09 | 393.26 |
| 92 – 101 | 31 | 96.5 | 9312.25 | 2991.5 | 288679.75 | 3.7 | 13.69 | 424.39 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 101 – 110 | 4 | 105.5 | 11130.25 | 422 | 44521 | 12.7 | 161.29 | 645.16 |
| 110 – 119 | 3 | 114.5 | 13110.25 | 343.5 | 39330.75 | 21.7 | 470.89 | 1412.67 |
| 119 – 128 | 3 | 123.5 | 15252.25 | 370.5 | 45756.75 | 30.7 | 942.49 | 2827.47 |
| **TOTAL** | **70** | | | **6494** | **612581.5** | | 2363.83 | 10123.90 |

**Standard Deviation**

$$\sigma = \sqrt{\frac{\sum f(x_i-\bar{x})^2}{\sum f}} = \sqrt{\frac{10123.90}{70}} = \sqrt{144.6271429} \quad = 12.03$$

Alternatively,

$$\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} \quad = \sqrt{\frac{612581.5}{70} - 92.77142857^2}$$

$$= \sqrt{8751.164286 - 8606537959} \quad = \sqrt{144.6263268} = 12.03$$

**Variance**

$$\sigma^2 = \frac{\sum f(x_i - \bar{x})^2}{\sum f} = \frac{10123.90}{70} = 144.6$$

Alternatively,

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{612581.5}{70} - 92.77142857^2 =$$

$$8751.164286 - 8606537959 = 144.6$$

**Sample Variance**

$$s^2 = \frac{\sum f(x_i-\bar{x})^2}{\sum f-1} = \frac{10123.90}{69} = 146.7231884 \approx 146.7$$ Alternatively,

$$s^2 = \frac{\sum fx^2 - \frac{(\sum fx_i)^2}{\sum f}}{\sum f-1} = \frac{\mathbf{612581.5} - \frac{\mathbf{(6494)^2}}{\mathbf{70}}}{69} = \frac{10123.84286}{69} = 146.7$$

## 5.3.7 Coefficient of Variation

When comparing the variability of two sets of data, it is necessary to take into account the units of measurements for the two data sets. If the unit of measurements differ, comparison of variations become unreasonable due to the influence of unit on the magnitude of such results. For the purpose of comparison, it therefore seems reasonable to use a measure that does not depend on any unit of

measurement. One such measure is obtained by dividing the standard deviation by the mean. This measure is called the coefficient of variation, and it is usually expressed as a percentage of the mean. Thus,

$$Coefficient\ of\ Variation\ (COV) = \frac{\sigma}{\bar{x}} x100 \qquad (5.33)$$

## Example 5.20

Find the coefficient of variation of the data solved in examples 7.3 and 7.4

$$Coefficient\ of\ Variation\ (COV) = \frac{\sigma}{\bar{x}} x100 = \frac{2.9}{13.7} x100 = 21.12\%$$

$$Coefficient\ of\ Variation\ (COV) = \frac{\sigma}{\bar{x}} x100 = \frac{12.03}{92.77} x100 = 12.97\%$$

## Exercises

1. The arithmetic mean of a certain set of observations is 50. If two items with values 60 and 68 are added to this data, the mean increases to 52. Find the number of items in the original data.
2. Below are the ages of buildings obtained in a building census:

| Age | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 4 | 9 | 17 | 23 | 20 | 14 | 10 | 3 |

Calculate:

(a) The mean, median and the modal age of the buildings

(b) The first and third quartiles division of the buildings ages

(c) The $20^{th}$ and $80^{th}$ Percentiles division of the buildings ages

(d) The $3^{rd}$ and $7^{th}$ Deciles division of the buildings ages

(e) Calculate (i) Variance (ii) Standard deviation (iii) Mean deviation (iv) Coefficient of variation (v) Coefficient of Mean deviation for the sales figure.

3. Draw the histogram of question 2, and from your chart, estimate the median, $80^{th}$ percentile, $7^{th}$ Deciles and Quartile deviation.
4. A manufacturing company has the profit distributions for the year ended 2019 as follows:

| Profit(N'000) | 50-60 | 60-70 | 70-80 | 80-90 | 90 - 100 | 100-110 | 110-120 | 120-130 | 130-140 | 140-150 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 8 | 12 | 14 | 7 | 10 | 9 | 11 | 9 | 15 | 5 |

Using the assumed mean of ₦95,000, calculate the average profit made by the company.

5. Arithmetic Mean of 98 items is 50. Two items 60 and 70 were left out at the time of calculations. What is the correct mean of all items?
6. Given the test scores of five (5) randomly selected students as follows: 14, 18, 16, 15 and 10. Compute the arithmetic mean, Geometric Mean and Harmonic mean respectively.
7. The price of a particular brand of flour has been increasing consistently from year 2016 to 2019 as follows: Year 2016 - #1500, Year 2017 - #1600, Year 2018 - #1800 and Year 2019 - #2000. What is the average yearly percentage increase in price?
8. A student's mean scores in part 1 and part 2 of his diploma examinations are 65% and 70% respectively. If the part 2 result is given twice as much weight as that of part 1, calculate the student's overall mean score.
9. Arithmetic mean of a group of 100 items is 50 and of another group of 150 items is 100. What will be the mean of all the items?
10. Arithmetic mean of 50 items were 100 and its median is 95. At the time of calculations, two items 180 and 90 were wrongly taken as 100 and 10. What is the correct mean and median?
11. Compute the mean and standard deviation of $N$ natural numbers $1, 2, 3, \cdots, N-1, N$.

$$HINT: \left(\sum_{1}^{N} x_i = \frac{N(N+1)}{2}\right) and \sum_{1}^{N} x_i^2 = \frac{N\{N+1)(2N+1)}{6}$$

12. Given the mean and standard distributions of two distributions of 80 and 120 items as 40, 2.5 and 30, 4 respectively. Find the standard deviation of all the items taken together.
13. If the number of distributions in exercise 12 is increased to 3 with the number of items given as 150, mean and standard deviation given as 40 and 6 respectively. Find the standard deviation of the combined distributions.
14. The mean and standard deviation of 100 items were calculated as 50 and 4 respectively by a computer analyst who has mistakenly taken the value 30 in place of 50 for one item. Having noticed this anomaly, you are required to calculate the correct mean and standard deviation for the whole items.
15. The mean and standard deviation of 100 items are 50 and 4 respectively. Find the sum of squares of all the items.
16. Crate of soft drinks sold in a supermarket for a 30 day sales period are given in the frequency distribution table below:

# Chapter Six

# 6.0 MEASURES OF SYMMETRY AND PEAKEDNESS

The two previous chapters have discussed extensively the features of mean and standard deviation, as two of the most important parameters which describe the frequency distribution of data set in terms of locations and spread of items about the mean. However, other important features such as symmetry and Peakedness of the distribution have not been discussed. How symmetrical is the distribution about the mean and how peaked is the distribution of a given data will be exemplified Moments is a whole series of measures which when properly interpreted, give a wealth of information about the shape of distribution. It is pertinent to state that both the mean ($\bar{x}$) and standard deviation ($\sigma$) are the first two members of the moments series.

## 6.1 Normality Test

This test is mostly carried out in descriptive statistics using **Jarque-Bera (JB)** statistic. This is a goodness-of-fit test of whether data have the skewness and kurtosis matching a normal distribution. The test is named after Carlos Jarque and Anil K. Bera. The test statistic JB is defined as:

$$JB = \frac{n-k+1}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right) \quad (6.1)$$

The decision criteria in using JB statistic is such that when its computed P-value is greater than 0.05 at 5% level of significance, we assume normality of the given observations.

## 6.2 Moments

Let $x_1, x_2, \cdots, x_n$ be a sample of size $n$, the $k^{th}$ sample moment about the origin (point zero) i.e. at a distance $x$ from origin zero is estimated as

$$M_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (6.2)$$

If a force equal to the frequency $f$ associated with $x$, then the $k^{th}$ moment about origin is defined as

$$M_k = \frac{1}{\sum f}\sum_{i=1}^{n} f x_i^k \quad (6.3)$$

Thus, the $1^{st}$ and $2^{nd}$ moments are derivable from equations (6.1) and (6.2) by setting $k = 1$ and $k = 2$ respectively. The resulting moment's equations are given as

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (6.4)$$

$$M_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad (6.5)$$

$$M_1 = \frac{1}{\sum f}\sum_{i=1}^{n} f x_i \quad (6.6)$$

$$M_2 = \sum_{i=1}^{n} f x_i^2 \quad (6.7)$$

When the moment is centered on the mean, the $k^{th}$ sample moment is defined as

$$M_k = \frac{1}{n}\sum (x_i - \bar{x})^k \quad (6.8)$$

$$M_k = \frac{1}{N}\sum f(x_i - \bar{x})^k \quad (6.9)$$

Thus, the $1^{st}$ and $2^{nd}$ moments about the mean are derivable from equations (6.7) and (6.8) by setting $k = 1$ and $k = 2$ respectively. The resulting moment's equations are given as

$$M_1 = \frac{1}{n}\sum (x_i - \bar{x}) \quad (6.10)$$

$$M_1 = \frac{1}{N}\sum f(x_i - \bar{x}) \quad (6.11)$$

$$M_2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \sigma^2 \quad (6.12)$$

$$M_2 = \frac{1}{N}\sum f(x_i - \bar{x})^2 = \sigma^2 \quad (6.13)$$

## 6.3 Measures of Skewness

Skewness is a procedure of measuring the extent to which a given set of data departs from symmetry. Any set of data that is not symmetric can either be skewed to the right or left (asymmetrical). Symmetry is the ability of data set to exhibit a normal curve. However, some data set show a tendency to have a prolonged tail either to the right or left. When a distribution is symmetric, it would have a mean that is equal to its median and mode, otherwise the mean tends to lie on the same side of the median and mode in the part of the longer tail. The shape of histogram is used to determine whether a set of data is skewed to the right or left. If the histogram has a longer tail to the right than to the left, it is said to be positively skewed, and if the histogram has a longer tail to the left than to the right, it is said to be negatively skewed as illustrated in the diagrams below:

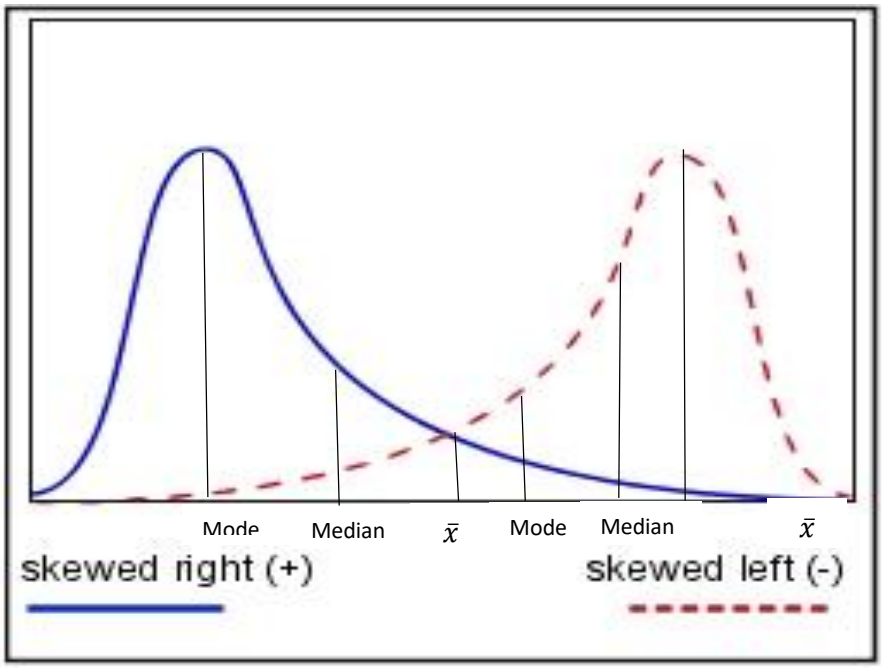


**Fig. 6.1 Asymmetric curves**

According to Figure 8.1 above, the Mean > Median > Mode for a positively skewed distribution while for a negatively skewed distribution, the Mean < Median < Mode. For symmetric curve, the Mean = Median = Mode as shown in the curve below

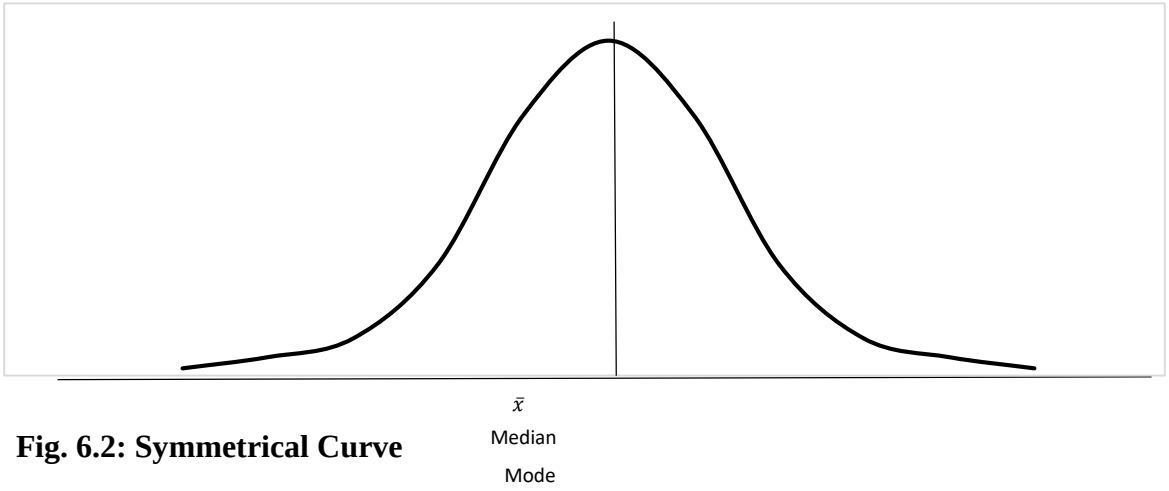


**Fig. 6.2: Symmetrical Curve**

It is pertinent to note that in a symmetrical distribution, all the odd moments about the mean apart from $M_1$ (i.e $M_3, M_5, M_7, \cdots$) are equal to zero, if otherwise the distribution is skewed. However, the computation of odd moments alone is not sufficient for measuring Skewness. The knowledge of the observations spread (i.e. standard deviation) is equally of importance to exhibit the degree of symmetry.

Thus, the various moments divided by their equivalent power of standard deviation give adequate measures of the coefficient of Skewness. Using the Greek letter $\psi$ to denote the coefficient of Skewness, symbolically we have the following expressions for the measures

$$\psi_3 = \frac{M_3}{\sigma^3} = \frac{\sum(x_i-\bar{x})^3}{n\sigma^3}$$

$$\psi_5 = \frac{M_5}{\sigma^5} = \frac{\sum(x_i-\bar{x})^5}{n\sigma^5} \quad (6.14)$$

$$\psi_7 = \frac{M_7}{\sigma^7} = \frac{\sum(x_i-\bar{x})^7}{n\sigma^7}$$

$\vdots$

The most prominently used among the moments equation (8.14) is the third moment ($\psi_3$) because it is relatively easier to compute and also because the higher the moment, the more will it vary from sample to sample. Thus, $\psi_3$ can further be expressed as

$$\psi_3 = \frac{M_3}{M_2^{\frac{3}{2}}} \quad (6.15)$$

Since standard deviation ($\sigma$) is the square root of second moment about the mean, $M_2$ ($\sigma^2$), where $M_3 = \frac{1}{n}\sum(x_i - \bar{x})^3$ and $M_2 = \frac{1}{n}\sum(x_i - \bar{x})^2$.

For frequency distribution data, the third and the second moments about the mean are expressed as $M_3 = \frac{1}{N}\sum f(x_i - \bar{x})^3$ and $M_2 = \frac{1}{N}\sum f(x_i - \bar{x})^2$.

Another measure of Skewness apart from moments approach is through the study of relationship between the 3 M's of location (mean, median, and mode), defined by Karl Pearson as

$$Coefficient\ of\ Skewness\ (C.S) = \frac{\bar{x}-Mode}{Standard\ Deviation} \quad (6.16)$$

*Since in a moderately skewed distribution,* $Mode = \bar{x} - 3(\bar{x} - Median)$

The coefficient of Skewness becomes

$$C.S = \frac{3(\bar{x}-Median)}{Standard\ Deviation} \quad (6.17)$$

We have a symmetrical distribution $if\ C.S = 0$, positively skewed $if\ C.S > 0$ and negatively skewed $If\ C.S < 0$.

Other measures of Skewness that are available for use, but which are not considered in this text are Quartile measure and Percentile measure of Skewness commonly known as Bowley's and Kelly's measures respectively.

**Example 6.1**

In a moderately skewed distribution, the mean and the median are 25.6 and 26.1 inches respectively.

(a) What is the mode of the distribution?
(b) Determine the coefficient of Skewness if the standard deviation is 1.2 and comment on its distribution.

**Solution**

$(a)$ $Mode = \bar{x} - 3(\bar{x} - Median) = 25.6 - 3(25.6 - 26.1) = 27.1$

(b) $Coefficient\ of\ Skewness\ (C.S) = \frac{\bar{x} - Mode}{Standard\ Deviation}$

$$= \frac{25{,}6 - 27.1}{1.2} = \frac{-1.5}{1.2} = -1.25$$

Since $C.S < 0$, the distribution is negatively skewed.

**Example 6.2**

Find the coefficient of Skewness and comment on the distribution of the frequency table given below

| Class | 0-2 | 2-4 | 4-6 | 6-8 | 8-10 | 10-12 | 12-14 |
|---|---|---|---|---|---|---|---|
| F | 4 | 6 | 5 | 12 | 10 | 8 | 5 |

**Solution**

| $x_i$ | $f$ | $fx_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $f(x_i - \bar{x})^2$ | $(x_i - \bar{x})^3$ | $f(x_i - \bar{x})^3$ |
|---|---|---|---|---|---|---|---|
| 1 | 4 | 4 | -6.48 | 41.9904 | 167.9616 | -272.0978 | -1088.3912 |
| 3 | 6 | 18 | -4.48 | 20.0704 | 120.4224 | -89.9154 | -539.4924 |
| 5 | 5 | 25 | -2.48 | 6.1504 | 30.7520 | -15.2530 | -76.2650 |
| 7 | 12 | 84 | -0.48 | 0.2304 | 2.7648 | -0.1106 | -1.3271 |
| 9 | 10 | 90 | 1.52 | 2.3104 | 23.1040 | 3.5118 | 35.1181 |
| 11 | 8 | 88 | 3.52 | 12.3904 | 99.1232 | 43.6142 | 348.9137 |
| 13 | 5 | 65 | 5.52 | 30.4704 | 152.3520 | 168.1966 | 840.9830 |
| | **50** | **374** | | | **596.4800** | | **-480.4609** |

$$\bar{x} = \frac{\sum f x_i}{\sum f} = \frac{374}{50} = 7.48$$

$$\psi_3 = \frac{M_3}{M_2^{\frac{3}{2}}} = \frac{\sum f(x_i - \bar{x})^3}{[\sum f(x_i - \bar{x})^2]^{\frac{3}{2}}} = \frac{-480.4609}{(596.4800)^{\frac{3}{2}}} = -0.03298103$$

Since $\psi_3 < 0, the\ distribution\ is\ negatively\ skewed.$

## 6.4 Measures of Kurtosis

Kurtosis shows the degree of Peakedness of a distribution relative to a normal distribution. Kurtosis is a measure of the extent to which the curve is more flat topped or more peaked than the normal curve. A normal curve must have its features similar to that of Figure 6.2.

When a normal curve is not very peaked or very flat topped, it is said to be **Mesokurtic.** It has the same degree of convexity as the normal curve. If the curve has a relatively high peak, it is said to be **Leptokurtic** and if the curve is unusually flat topped, it is said to be **Platykurtic** as shown in Figure 8.3 below.

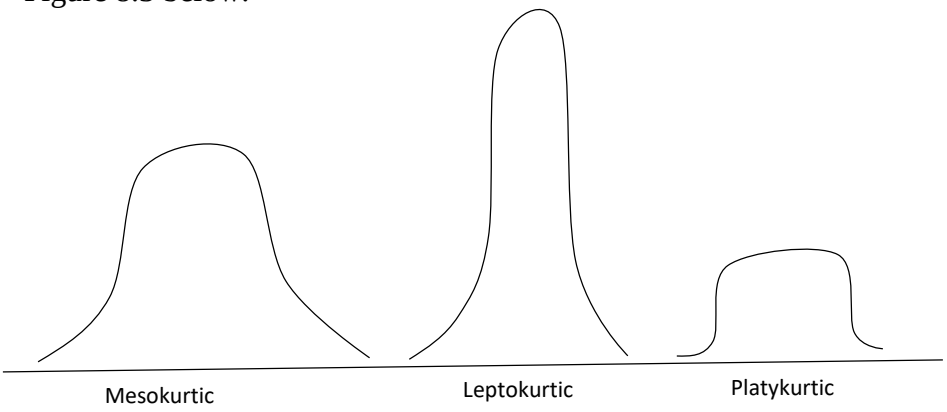


**Fig. 6.3: Degree of Peakedness**

The Kurtosis is measured by

$$\psi_4 = \frac{M_4}{\sigma^4} = \frac{M_4}{M_2{}^2} = \frac{\sum(x_i - \bar{x})^4}{[\sum(x_i - \bar{x})^2]^2} = \frac{\sum f(x_i - \bar{x})^4}{[\sum f(x_i - \bar{x})^2]^2} \qquad (6.18)$$

In a Mesokurtic curve, $\psi_4$ will be equal to 3 which implies that the curve is symmetrical. If $\psi_4$ is greater than 3, the curve is leptokurtic and it will be Platykurtic if $\psi_4$ is less than 3.

Thus, equation (6.17) may be written as

$$Kurtosis = \psi_4 - 3 = \frac{M_4}{\sigma^4} - 3 \qquad (6.19)$$

If Kurtosis is positive, it means that the curve is relatively peaked than the normal curve. If Kurtosis is negative, the curve is more flat-topped than the corresponding normal curve.

**Example 6.3**
Find the kurtosis for the distribution given in example 6.2.
Solution

| $x_i$ | $f$ | $fx_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $f(x_i - \bar{x})^2$ | $(x_i - \bar{x})^4$ | $f(x_i - \bar{x})^4$ |
|---|---|---|---|---|---|---|---|
| 1 | 4 | 4 | -6.48 | 41.9904 | 167.9616 | 1763.1937 | 7052.7748 |
| 3 | 6 | 18 | -4.48 | 20.0704 | 120.4224 | 402.8210 | 2416.9258 |
| 5 | 5 | 25 | -2.48 | 6.1504 | 30.7520 | 37.8274 | 189.1371 |
| 7 | 12 | 84 | -0.48 | 0.2304 | 2.7648 | 0.0531 | 0.6370 |
| 9 | 10 | 90 | 1.52 | 2.3104 | 23.1040 | 5.3380 | 53.3795 |
| 11 | 8 | 88 | 3.52 | 12.3904 | 99.1232 | 153.5220 | 1228.1761 |
| 13 | 5 | 65 | 5.52 | 30.4704 | 152.3520 | 928.4453 | 4642.2264 |
| | **50** | **374** | | | **596.4800** | | **15583.2567** |

$$\psi_4 = \frac{M_4}{M_2{}^2} = \frac{15583.2567}{(596.4800)^2} = 0.043799227$$

Since $\psi_4 = 0.04$ which is less than 3, it shows that the distribution is far from symmetry and is Platykurtic.

## Exercises

1. In a moderately skewed distribution, the mean and the median are 25.6 and 26.1 inches respectively. (a) What is the mode of the distribution? (b) Compute the coefficient of Skewness.
2. If the mode and median of a moderately asymmetrical series are 16" and 17" respectively. What would be its most probable mean?
3. If the standard deviation of a symmetrical distribution is 3, what will be the value of the 4$^{th}$ moment about the mean in order that the distribution be (a) Mesokurtic (b) Platykurtic?
4. The text scores of 10 undergraduates' students in a college are given as:
   9, 14, 15, 10, 11, 12, 15, 16, 17, 18
   Find the coefficient of Skewness and kurtosis using the method of moments
5. Calculate the coefficient of Skewness using the Pearson and Moment methods for the data below.

| Class | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| frequencies | 1 | 3 | 4 | 2 | 6 | 4 |

Comment on the distribution of the data.

6. Calculate the kurtosis of data in example 5 and comment on the symmetry of the data.