USAID | AFGHANISTAN
FROM THE AMERICAN PEOPLE

USAID/CAPACITY BUILDING ACTIVTIY (CBA)

BASIC TRAINING IN DATA ANALYSIS FOR STATISTICS UNIT

MANUAL

FY 2017 (06 June - 21 June 2018)

This Training Manual is made possible by the support of the American People through the United States Agency for International Development (USAID.) The contents of this manual are the sole responsibility of Chemonics International and do not necessarily reflect the views of USAID or the United States Government.

Government of Islamic Republic of Afghanistan

Ministry of Education

Basis Training in Data Analysis for Statistics Unit

Manual

Prepared by: EMIS COMPONENT OF USAID/CAPAPCITY BUILDING ACTIVITY (CBA)

Table of Contents

Training Program- I:

BASIC TRAINING IN DATA ANALYSIS

TRAINING OBJECTIVES:

By the end of this training session, participants will be able to perform independently descriptive analysis of the collected Educational Data. They should be able to validate their understanding by providing examples of each type of data.

WHAT WILL THE PARTICIPANTS LEARN?

- ✓ The participants will learn Meaning, Definition, Nature and Limitations of Statistics
- ✓ The participants will be able that "How Statistics Works" in educational data analysis
- ✓ The participants will be able to understand what is Data (variables) and what are the different ways of data collection? and could differentiate between several types of data. they will learn how to appropriately and objectively present the data in hand.
- ✓ The participants will learn about the basic concept of diagrams and graphs. They will be able to differentiate between various types of diagrams and graphs and will understand the appropriate use of them and can visualize the data in hand by means of suitable diagrammatic representation.

ACTIVITIES:

Activity – 1: (The Participants are introduced to the Meaning, Definition, Nature and Limitations of Statistics, use of statistics in educational data, Types of Data & Data Collection Techniques, Different Scales): In this session, the participants are introduced to the statistical theory and how statistics play role in education data, types of data (variables), data collection techniques through PowerPoint presentation by the trainer.

Activity – 2: (Group Work and Discussions: The Participants undertake the group work and discussions on basic concepts about education statistics and calculation of errors and percentages. The activity is facilitated by the trainer): Participant groups are given the target of calculating different examples from data. The participants take part in the discussion related to each topic and will be able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding.

Activity – 3: (The participants are introducing to diagrams and graphs and they will learn how to visualize the information in hand through diagrams and graphs through power point presentation b. The activity is facilitated by the trainer): the participants introduce different types of diagrams and graphs and their use in representing

the education statistics. The participants take part in the discussion related to each topic and will be able to answer trainer's questions to determine their understanding of the topic. The participants discuss various types of diagrams and graphs with the trainer to further enhance their knowledge in data visualization.

Activity – 4: (Group Work and Discussions: The Participants undertake the group work on identifying different types of diagrams and graphs. they will introduce how to visualize the information in hand through diagrams and graphs. The activity is facilitated by the trainer): The participants discuss various types of diagrams and graphs and their use in every day educational data visualization and share their understanding with each other in the group. Participant groups are given the target of calculating different examples from data.

Activity – 5: (The Participants are introduced to measures of location like, (Mean, Median, Mode, the activity is facilitated by the trainer): The participants are introduced to the various measures of location through PowerPoint presentation by the trainer. The participants take part in the discussion related to each topic and are able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding.

Activity – 6: (Group Work: The Participants undertake the group work on identifying different types of calculating different measures of location by means of examples in education data, the activity is facilitated by the trainer): The participants discuss various measures of location and share their understanding with each other in the group. The trainer will demonstrate the computation formulae on data and how to calculate the required information from these measures. Participant groups are given the target of calculating different examples from data.

Activity – 7: (The Participants are introduced to partition values like, (quartiles, deciles, percentiles, the activity is facilitated by the trainer): The participants are introduced to the partition values through PowerPoint presentation by the trainer. The participants take part in the discussion related to each topic and are able to answer trainer's questions to determine their understanding of the topic. The participants are encouraged to discuss their questions with the trainer in order to help enhance their understanding

Activity – 8: (Group Work and Discussions: The Participants undertake the group work on identifying and calculating different partition values by means of examples in education data, the activity is facilitated by the trainer): The participants discuss various partition values and share their understanding with each other in the group. The trainer demonstrates the computation formulae on data and how to calculate the required information from these measures. Participant groups are given the target of calculating different examples from data.

Activity – 9: (Quiz: The Participants take a quiz related to the purpose and use of various Statistical measures): In this session the trainer will provide a short overview of all the work done during last sessions. The participants provided with list of different

measures, and they are required to express their understanding of utilization of each measure.

TOTAL SESSION TIME: 16 Hours (Two days)

Training Agenda

Duration	Activity	What is needed?
TRAINING – I: BASIC STATISTICS TRAINING FOR DATA ANALYSIS		
Day first		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on Meaning, Definition, Nature and Limitations of Statistics, use of statistics in educational data, Types of Data & Data Collection Techniques, Different Scales	– Projector – Power Point Presentation
Tea Break for 15 Minutes		
1 Hour 15 Minutes	Group work, Discussions and Findings	– Projector – Flip Chart – Notebook and Pen
Lunch & Prayer Break (12:00 – 01:00)		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on visualization of data by means of graph and diagrams	– Projector – Power Point Presentation
1 Hour 30 Minutes	Group work and Discussions	– Projector – Flip Chart – Notebook and Pen
Second Day		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on measures of location.	– Projector – Power Point Presentation
Tea Break for 15 Minutes		
1 Hour 15 Minutes	Group work, and Discussions	– Projector – Flip Chart – Notebook & Pen
Lunch & Prayer Break (12:00 – 01:00)		
1 Hour 30 Minutes	Participatory Lecture and Power point presentation by the trainer on Partitioned Values	– Projector – Power Point Presentation
1 Hour 15 Minutes	Group work, and Discussions	– Projector – Flip Chart

		– Notebook and Pen
15 Minutes	Quiz: The Participants take a quiz related to the purpose and use of various Statistical measures	– Question paper

DEFINITION OF STATISTICS:

"STATISTICS" is a body of theories and methods which have been developed for handling the collection, analysis and description of sample data for drawing useful conclusions. It is a primary function to assist the investigator in making decision relative to the parameter of the population from which the sample data are taken.

THE MODERN DEFINITION OF STATISTICS

"STATISTICS" are the numerical statements of facts capable of analysis and interpretation, and the science of STATISTICS is the study of the principles and the methods applied in collecting, presenting, analyzing, and interpreting data in every field of life".

THE MEANING OF STATISTICS

The word **"STATISTICS"** which come from the Latin word "Status" meaning a political state, is used in three different ways as follows:

- ✓ **Firstly;** the word statistics refer to "numerical facts systematically arranged. In this sense the word statistics is always used in the plural; for example, statistics of prices, statistics of births, in all these examples the word statistics denotes a set of numerical data in the respective fields.
- ✓ **Secondly;** the word "Statistics" is defined as a discipline that includes procedures and techniques, used to collect, process, and analyze numerical data to make inferences and to reach decisions in the face of uncertainty. In this sense the word "Statistics" is used in the singular. Thus, the word "Statistics" used in the plural refers to a set of numerical information and in the singular, denotes the science of basing decision on numerical data.
- ✓ **Thirdly**; the word "Statistics" are numerical quantities calculated from sample data; a single quantity that has been so calculated is called a statistic. For instance, the sample mean is a statistic.

THE NATURE OF THIS DISCIPLINE

Statistics as a subject may be divided in to Descriptive statistics and inferential statistics:

DESCRIPTIVE STATISTICS:

Descriptive Statistics is that branch of statistics which deals with concepts and method concerned with summarization and description of the important aspects of numerical data. This area of study consists of the condensation of data, their graphical displays and the computation of few numerical quantities that provide information about the centre of the data and indicate the spread of the observations.

INFERENTIAL STATISTICS:

Inferential statistics is that branch of statistics which deals with procedures for making inferences about the characteristics that describe the large group of data or the whole, called the population, from the knowledge derived from only a part of the data, known as a sample. This area includes the estimation of population parameters and testing of statistical hypotheses. This phase of statistics is based on probability theory as the inferences which are made on the basis of sample evidence, cannot be absolutely certain.

COMPARISON

DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
✓ A cricket player wants to find his average score for the last 20 – games. ✓ Ahmad wants to describe the variation in his four test scores in statistics. ✓ Gulaly wants to determine the average weekly amount she spent on groceries in the past 6 – months.	✓ A cricket player wants to estimate his chance of scoring based on his current season average. ✓ Based on the first four test scores, Ahmad would like to predict the variation in his final statistics test scores. ✓ Based on last 6-month grocery bills, Gulaly would like to predict the average amount she spends on groceries for the upcoming year.

LIMITATIONS (CHARACTERISTICS) OF STATISTICS)

Statistical methods cannot be applied to all kind of phenomena and cannot answer all the queries, although these methods are used in almost every field of life. Some of the limitations of statistics are:

- ✓ Statistical methods are best applicable to quantitative data.
- ✓ Statistical decisions are subject to certain degree of error.
- ✓ Statistical statements are true on an average, i.e. true for a group of individuals and may not be true for an individual.
- ✓ Statistics deals with those characteristics or aspects of thing which can be described numerically either by counts or by measurements.

- ✓ Statistics deals with variability that obscures underlying patterns. No two objects in this universe are exactly alike. If they were, there would have been no statistical problem.
- ✓ Statistics deals with the behavior of aggregates or large groups of data. It has nothing to do with what is happening to a particular individual or object of the aggregate.
- ✓ Statistical results might be misleading the incorrect if sufficient care in collecting, processing and interpreting the data is not exercised or if the statistical data are handled by a person who is not well versed in the subject matter of statistics.

THE WAY IN WHICH STATISTICS WORKS:

As it is such an important area of knowledge, it is definitely useful to have a fairly good idea about the way in which it works, and this is exactly the purpose of this introductory course. The following points indicate some of the main functions of this science:

- ✓ Statistics assists in summarizing the larger set of data in a form that is easily understandable.
- ✓ Statistics assists in the efficient design of laboratory and field experiments as well as surveys.
- ✓ Statistics assists in a sound and effective planning in any field of inquiry.
- ✓ Statistics assists in drawing general conclusions and in making predictions of how much of a thing will happen under given conditions.

DATA COLLECTION & TYPES OF DATA:

The word **data** refers to the information that has been collected from an experiment, a survey or an historical record, etc. The survey data are broadly classified into two types: **categorical** data and **scale data**. Categorical data are collected using a nominal scale or an ordinal scale. Scale data are collected based on interval and ratio scales. Why it is important to know the type of data? General speaking, statistical techniques are determined by the type of data.

OBSERVATIONS

In statistics, an **observation** often means any sort of numerical recording of information, whether it is a physical measurement such as height or weight; a classification such as heads or tails, or an answer to a question such as "**Yes**" or "**No**".

VARIABLES

A characteristic that varies with an individual or an object is called a **variable**. For example, age is variable as it varies from person to person. A variable can assume a number of values. The given set of all possible values from which the variable takes on a value is called its domain. If for a given problem, the domain of a variable contains only one value, then the variable is referred to as a **"constant".**

TYPES OF VARIABLES

✓ QUANTITATIVE VARIABLES

A variable is called a quantitative variable when a characteristic can be expressed numerically, or the characteristic which is countable, such as age, height, weight, income or number of children in a family.

✓ QUALITATIVE VARIABLES

The variable is referred to as qualitative variable if the characteristic is non-numerical or characteristic which is measurable, such as education, sex, eye-colour, intelligence, poverty, satisfaction etc. A qualitative characteristic is also called an attribute.

✓ DISCRETE VARIABLES

A discrete variable is one that can take only a discrete set of integers or whole numbers. That is the values are taken by jumps or breaks. A discrete variable represents counts data such as number of persons in a family, the number of rooms in a house etc.

✓ CONTINUOUS VARIABLES

A variable is called a continuous variable if it can take on any value fractional or integral within a given interval, i.e. its domain is an interval with all possible values without gapes. A continuous variable represents measurement data such as the age of a person, the height of a plant, the temperature at a place.

NOMINAL SCALE

In nominal scale, the numbers represent labels or tags for identifying objects, properties, or events. For example, gender is a nominal scale variable; in a data set 'males' and 'females' could be coded as 0 = female and 1 = male. The nominal scale does not possess order, distance, or origin. For gender variable, we can reverse the code, i.e. use '1' for female and '0' for male, or use any other two arbitrary chosen numbers. Similarly, rainfall may be classified as heavy, moderate, and light. We may use numbers 1, 2, and 3, to denote the three categories of rainfall. The numbers when they are used only to identify the categories of the given scale carry no numerical significance and there is no particular order for the grouping.

ORDINAL OR RANKING SCALE

It includes the characteristics of nominal scale and in addition has the property of ordering or ranking of measurements. For example, the performance of students is rated as “excellent”, “Good”, “fair”, etc. Numbers 1, 2, 3, 4, etc. are also used to indicate ranks. The only relation that holds between any pair of categories is that of “greater than” (or more preferred). In other words; Ordinal scale provides categorical data but unlike nominal scale, there is a logical ordering of the categories.

INTERVAL SCALE

On an interval scale, measurements are not only categorized and ordered (therefore, having the properties of the two previous scales), but the distances between each interval on the scale are equal right along the scale from the low end to the high end. Two points next to each other on the scale, no matter whether they are high or low, are separated by the same distance.

For example, when temperature is measured in Centigrade scale, the difference between 960C and 980C is the same as between 1000C and 1020C. However, interval scale suffers from one limitation that it does not have an absolute (i.e., fixed) zero value. Thus, a temperature reading of 100C does not mean that the temperature is 10 times hotter than something measuring 100C even though the value on the scale is 10 times as large. There is a popular joke: "If it's twice as cold today as it was yesterday, and it was zero degree yesterday, how cold is it today?" This illustrates the limitation of interval measurements such as Celsius and Fahrenheit temperature: by setting zero at an arbitrary point, they make it impossible to multiply and divide meaningfully.

RATIO SCALE

It is a special kind of an interval scale, where the scale of measurement has a true zero point as its origin. The ratio scale is used to measure weight, volume, length, distance, money, etc. The key to differentiating interval and ratio scale is that the zero point is meaningful for ratio scale.

EXAMPLE OF MEASUREMENT SCALES

NOMINAL - LEVEL DATA	ORDINAL – LEVEL DATA	INTERVAL – LEVEL DATA	RATIO – LEVEL
✓ Gender (Male, Female) ✓ Eye colour ✓ Religion ✓ Specialization ✓ Nationality	✓ Grades (A, B, C, D, F) ✓ Position (1^{st} , 2^{nd}, 3^{rd}) ✓ Ranking of cricket player ✓ Rating (Poor, Good, Excellent) ✓ Socio – economic status (poor, middle class, rich)	✓ Temperature ✓ IQ Score	✓ Age ✓ Weight ✓ Height ✓ Time ✓ Salary ✓ Distance

ERROR OF MEASUREMENT

Experience has shown that a continuous variable can never be measured with perfect fineness because of certain habits and practices, methods of measurements, instruments used, etc. The measurements are thus always recorded correct to the nearest units and hence are of limited accuracy.

The actual or true values are, however, assumed to exist. For example, if a student's weight is recorded as 60.00 kg (correct to the nearest kilogram), his true weight in fact lies between 59.995 kg and 60.005 kg.

Thus, there is a difference, however small it may be, between the measured value and the true value. This sort of departure from the true value is technically known as the error of measurement.

ERROR

If the observed value and the true value of a variable are denoted by "x" and "x+ε" respectively, then the difference (x+ε) – x, i.e. ε is the error.

ABSOLUTE ERROR

This is the actual difference between an estimate or approximation and true value. e.g.: A house-wife may expect to spend Rs10 on her shopping. But actually, spend Rs12.5 therefore the absolute error is:

$$= 12.5 - 10 = 2.5$$

RELATIVE ERROR

The absolute error is divided by the estimate to find out the relative error and it may be expressed in percentage.

$$\text{Relative error} = R.E. = \frac{Absulute error}{Estimate} * 100$$

e.g.:

$$i) \quad R \cdot E = \frac{2.5}{10} \times 100 = 25\%$$

$$ii) \quad R \cdot E = \frac{2.5}{20} \times 100 = 12.5\%$$

Here it is clear that in this example, the 2nd estimate is better than the first. In the sense that in first case the error was 25% of the original estimate, while in the 2nd case error was 12.5% of the original estimate.

SAMPLING ERROR

A sampling error is the difference between the value of statistic obtained from random sample and the value of the corresponding population parameter because of chance variation in the selection of the elementary units. OR

The error which arises due to only a sample being used to estimate the population parameter is termed as sampling error.

NON - SAMPLING ERROR

The errors which occur during the process of gathering data, regardless of whether a sample or a complete census is taken, these errors are called non-sampling errors. The main sources of non-sampling errors are:

- ✓ Failure to measure some of the units in the selected sample
- ✓ Observational errors due to defective measurement techniques
- ✓ Error introduced in editing coding and tabulating the results.

TYPES OF DATA

There are two different forms of data:

- ✓ Primary data;
- ✓ Secondary data;

PRIMARY DATA

Primary data may be collected directly from the original source and is called a primary data. The primary data may be collected by the following ways:

- ✓ Direct interview method;
- ✓ Through mail (correspondence);

DIRECT INTERVIEW METHOD

In this method, the investigator contacts the unit and has personal interview with the unit. The information from the unit is recorded on the questionnaire or schedule.

This information will be more reliable and correct, but more expenditure may be involved, more time will be spent as the person himself will be going from place to place to collect the data. The investigator knows the purpose for which the data is being collected.

THROUGH MAIL

The data may be collected through correspondence. The questionnaires or schedule are sent by mail with the instruction for filling the same, and to return them. It is less costly to get the data by mail. The main drawback of this method is the poor response. Usually the response by mail in surveys has been found to be about 40%.

SECONDARY DATA

Sometimes we find that the data which we need had already been collected by some agencies for their study or the data is available in the published records. We may make use of such collected data which is known as a secondary data. Although for the agency which collected it from the original source, it is a primary data; but for the others it is secondary data. For instance;

- ✓ The University results can be used for analyzing and comparing pass percentages.
- ✓ The census records may be used for getting several types of information.
- ✓ The information and data on most of the aspects are published by central and state government departments and are easily available for use as a secondary data.

TIME SERIES DATA (OR) CHRONOLOGICAL DATA

The data which is collected with respect to time is called time series data or chronological data. For instance, the prices of different commodities in a period of time the temperature of Kabul city in different months of the year.

CROSS-SECTIONAL DATA (OR) GEOGRAPHICAL DATA

The data which is collected with respect to place is called cross-sectional data or geographical data. For instance: the prices of different commodities at different places, the temperature of different cities over a course of time.

COLLECTION OF DATA

For an inquiry or study, we first collect the data, and then arrange it in a proper manner. This arrangement of the data in a proper manner on the basis of similarities and dissimilarities is known as classification and tabulation of the data.

We classify the collected data in such a manner that the information is in a condensed table form and having almost all the characteristics present in the original data. After this we make use of the statistical methods to analyze the condensed table of the data to draw the conclusions.

FOUR DATA COLLECTION TECHNIQUES:

Any researcher is only as good as the data that drives it, so choosing the right technique of data collection can make all the difference. In this section, we will look at four different data collection techniques– **Observation, Questionnaire, Interview** and **Focus Group** Session – and evaluate their suitability under different circumstances.

OBSERVATION:

Seeing is believing, they say. Making direct observations of simplistic phenomena can be a very quick and effective way of collecting data with minimal disturbance. Establishing the right mechanism for making the observation is all you need.

Merits:

- Non-responsive sample subjects are a non-issue when you're simply making direct observation.
- If the observation is simple and doesn't require interpretation (e.g. the number of students in a class), this model doesn't require a very extensive and well-tailored training regime for the survey workforce.
- Infrastructure requirement and preparation time are minimal for simple observations.

Demerits:

- More complex observations that ask observers to interpret something (e.g. how many of the students are excellent in their studies, mediocre or very low) require more complex training and are prone to bias.
- Analysis may rely heavily on experts who must know what to observe and how to interpret the observations once the data collection is done.
- There is the possibility of missing out on the complete picture due to the lack of direct interaction with sample subjects.

QUESTIONNAIRE

For collection of data we may use a printed proforma. This proforma is called a questionnaire or schedule. A simple questionnaire is the form that one has to fill to get ID-card. An objective type test paper is another example of questionnaire. The papers are evaluated, the result is prepared, and the grades are awarded

Questionnaires are stand-alone instruments of data collection that will be administered to the sample subjects either through mail, phone or online. They have long been one of the most popular data collection techniques.

Merits:

- Questionnaires give the researchers an opportunity to carefully structure and formulate the data collection plan with precision.
- Respondents can take these questionnaires at a convenient time and think about the answers at their own pace.
- The reach is theoretically limitless. The questionnaire can reach every corner of the globe if the medium allows for it.

Demerits:

- Questionnaires without human intervention (as we have taken them here) can be quite passive and lose out some big information, leaving the responses open to interpretation. Interviews and Focus Group Sessions are instrumental in overcoming this shortfall of questionnaires.
- Response rates can be quite low. Questionnaires can be designed well by choosing the right questions to optimize response rates, but very little can be done to encourage the respondents without directly conversing with them.

INTERVIEWS:

Conducting interviews can help you overcome most of the shortfalls of the previous two data collection techniques that we have discussed here, by allowing you to build a deeper understanding of the thinking behind the respondents' answers.

Merits:

- Interviews help the researchers uncover rich, deep understanding and learn information that they may have missed otherwise.
- The presence of an interviewer can give the respondents additional comfort while answering the questionnaire and ensure correct interpretation of the questions.
- The physical presence of a persistent, well-trained interviewer can significantly improve the response rate.

Demerits:

- Reaching out to all respondents to conduct interviews is a massive, time-consuming exercise that leads to a major increase in the cost of conducting a survey.
- To ensure the effectiveness of the whole exercise, the interviewers must be well-trained in the necessary soft skills and the relevant subject matter.

FOCUS GROUP SESSIONS:

Group Sessions take the interactive benefits of an interview to the next level by bringing a carefully chosen group together for a moderated discussion on the subject of the survey.

Merits:

- The presence of several relevant people together at the same time can encourage them to engage in a healthy discussion and help researchers uncover information that they may not have envisaged.
- It helps the researchers verify the facts immediately; any inaccurate response will most likely be countered by other members of the focus group.
- It gives the researchers a chance to view both sides of the coin and build a balanced perspective on the matter.

Demerits:

- Finding groups of people who are relevant to the survey and persuading them to come together for the session at the same time can be a difficult task.
- The presence of excessively loud members in the focus group can subdue the opinions of those who are less vocal.
- The members of a focus group can often fall prey to group-think if one of them turns out to be remarkably persuasive and influential. This will bury the diversity of opinion that may have otherwise emerged. The moderator of a Focus Group Session must be on guard to prevent this from happening.

DATA TABULATION

Tabulation is a process of summarizing classified or grouped data in the form of a table, so that it is easily understood, and a researcher is quickly able to locate the desired information, Statistical data arranged in a tabular form facilitates computation of various statistical measures like, mean, standard deviation and correlation etc. Tables can be classified according to their purpose, stage of inquiry, nature of data or number of variables, (i.e. characteristics) used. On the basis of number of variables or characteristics, tables may be classified as follows:

- ✓ One – way or single variable table;
- ✓ Two – way table or two variable table (also known as cross table or contingency table) and;
- ✓ Manifold or multivariate table (i.e. more than two variables) table. Manifold tables enable full information to be incorporated and facilitate analysis of all related facts.

FREQUENCY DISTRIBUTION

Frequency distribution is a summary table in which the data are arranged into numerically ordered class intervals or categories. The number of class intervals depends on the number of observations in the data. Larger number of observations allows for a larger number of class groups. In general, the frequency distribution should have at least 5 class intervals but no more than 20. Each class interval should have the same width. To determine the width of a class interval, the range of the data is divided by the number of class intervals (the resulting figure is rounded off). The smallest data value is the lower-

class limit of the first class; add the class width to find the lower-class limit of the second class; the process is continued until highest data value is covered.

Frequency distribution can be represented graphically through a histogram. A histogram is a chart in which the rectangular bars are considered at the boundaries of each class.

PRESENTATION & SUMMARIZATION OF DATA

After the data has been collected the next step is to present this in an arranged form. The data can be arranged in ascending or descending orders. There are two important ways in which we can arrange the data.

- ✓ Classification according to attribute; and
- ✓ Classification according to class – intervals;

CLASSIFICATION ACCORDING TO ATTRIBUTE

The data in the qualitative form, i.e. possessing one character or possessing other character can be put in the table form. For example; if an investigator observes the people with respect to their behavior of smoking and nonsmoking. Further he may wish to be more specific and form the table in which each observation has two characteristics (attribute), i.e. sex of the person and the state of his /her behavior.

	male	Female	Total	
Smoker	10	5	15	
Non – Smoker	3	7	10	
Total	13	12	25	Grand Total

Similarly; the data with more characteristics can be represented in the table form and this classification is according to attribute. In the above example; the collected data were arranged in the table form which is called contingency table. This table is an array of natural numbers, where these natural numbers represent counts or frequencies. Therefore, in a nut shell we can define as: “An array of frequencies or counts represented in the table form is known as a contingency table”.

CLASSIFICATION ACCORDING TO THE CLASS INTERVAL

Some suitable intervals of the quantitative data are formed and the numbers of observations falling in the intervals are recorded. “This distribution of the frequencies in the different classes or intervals is known as Frequency Distribution.” A systematic way to form a frequency distribution is as follows:

- ✓ Observe whether the character or the variable is measurable (Continuous) or countable (Discrete).

- ✓ Decide the number of classes or groups to be formed such that most of the information contained in the data is retained.
- ✓ Find range = Maximum value – Minimum value.
- ✓ Divide range by the number of groups (classes) which will give the interval of the classes.
- ✓ Form the classes taking into consideration that in the first group lower limit should be lower than the minimum value and in the last group the upper limit of the class should be more than the maximum value. These points we keep in mind due to the fact that usually in the first and in the last group, the observations are less in number and for the analysis of the data we assume that the observations in a group are at the class mark.
- ✓ Against the classes so formed, the data are fed in the table with the use of tally marks. One should start from the very first observations and use the tally marks in lien of it and then for the second observation use the tally marks in the class in which it lies and proceed further till the last observation is fed in the table. In the class when a tally marks reach the number four then the fifth tally mark in a group is the cross of the first four tally marks. Proceeding in this way, we save the time no error in the counting is involved.
- ✓ Count the number of tally marks in a group and it is the frequency of the class.

EXAMPLE:

The marks of 25 – students in the class are given below (out of 50): 25, 26, 27, 14, 15, 21, 12, 25, 26, 41, 5, 12, 0, 29, 33, 7, 20, 32, 14, 38, 27, 11, 43, 40 & 35.

The data can be classified as in the following frequency distribution table:

SOLUTION:

Minimum Value		0	
Maximum Value		43	
Range		43 – 0 = 43	
Number of Classes		5	
Width of Class		9 (rounded)	
Classes (Interval)	Tally Marks	Frequency	Cumulative Frequency
0 - 10	|||	3	3
10 - 20	~~||||~~ |	6	9
20 - 30	~~||||~~ ||||	9	18
30 - 40	||||	4	22
40 - 50	|||	3	2

It may be noted that 40 will be in 40 – 50 intervals rather than in 30 – 40 intervals. This is a normal practice.

This form of arrangements is called the frequency distribution table. Here there is no break between the classes 0 – 10, 10 – 20, 20 – 30… And as such it is called continuous frequency distribution and the classification is exclusive type.

When there are gaps in – between the intervals, the distribution is called discrete, and the classification is inclusive classification. For example, the same data can be put as:

Classes (Interval)	Tally Marks	Frequency
0 - 9	|||	3
10 - 19	𝍸 |	6
20 - 29	𝍸 ||||	9
30 - 39	||||	4
40 - 49	|||	3

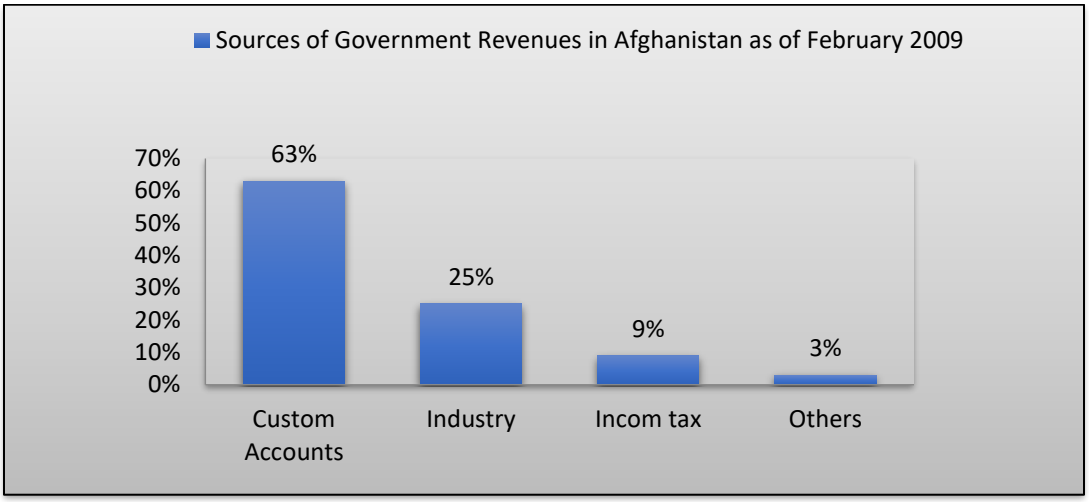
In fact, it depends on the data, which type of classification is to be used. For measurable characters use exclusive type of classification, and for countable characters use inclusive type of classification, for preparing a frequency distribution usually the number of intervals should be preferable between 5 and 20, and if the interval is say 10 – 20, then 10 is called the lower limit and 20 is called the upper limit, and 10+20/2 = 15, is called the mid-value, mid-point or the class-mark. For small data it is possible to form a frequency distribution in the above manner either by writing the data in an array or in the class intervals.

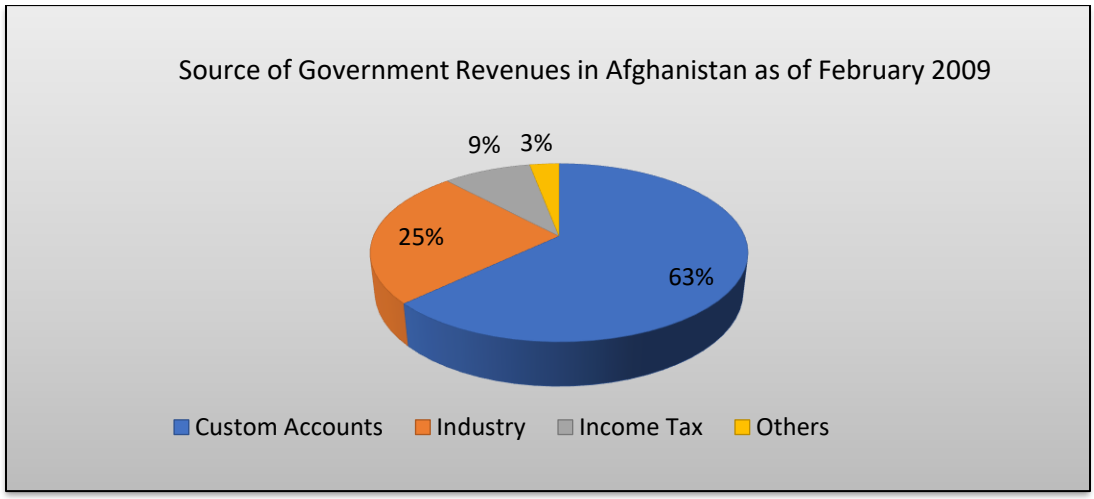
Tables and Charts For Categorical Data

A summary table for categorical data is similar to the frequency distribution table for numerical data. The summary table is often represented graphically using bar chart and pie chart. In a bar chart, each category is depicted by a bar, the length of which represents the frequency or percentage of observations falling into a category. The pie chart divides the pie into slices according to the percentage in each category.

Example:

The table below and the charts present various sources of government revenues in Afghanistan as of February 2009. As may be noted from the data, the largest source of government revenues in Afghanistan is custom duties, followed by industry. Custom accounts for 63 percent, Industry 25 percent, Income taxes 9 percent, and other 3 percent of the total government revenues. All these data and analysis are clearly readable from the charts. In fact, charts are very effective tool for communication.





CROSS TABULATION

In order to simultaneously study the responses to two categorical variables, a two-way table of cross-classification known as cross-table or contingency table is developed. When constructing a cross-tabulation, the independent variable is usually placed in columns and the Gender dependent variable is usually placed in rows. It is common to display cross – table by constructing side-by-side bar chart.

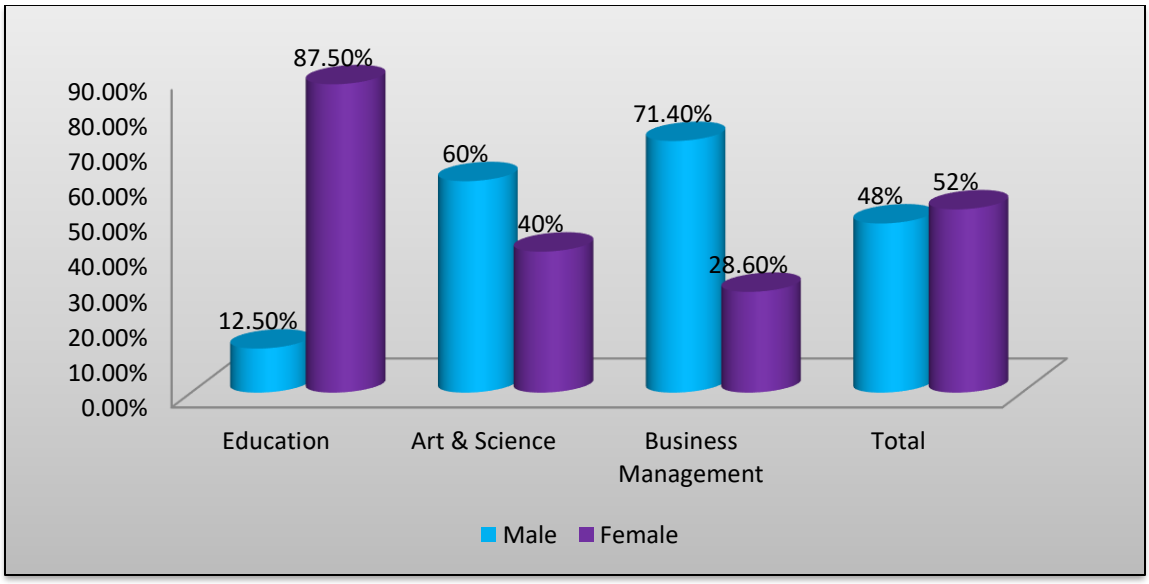
Example:

The adjacent tables present cross tabulation of data of 25 graduates on two categorical variables, namely "gender" shown in columns, and "major subject of study at college" shown in rows. Gender is placed in column because it is treated as an independent variable. The table also shows row, Column, and total percentages.

The row percentage suggests that between male and female, higher percentage (87.5%) of female prefer to study Education. Similarly, higher percentage (71.4%) of male prefers to study business management. Thus, it appears that choice of subject at college level depends on gender.

		Gender		
Major subject of study at college		Male	Female	Total
Education	Count	1	7	8
	Row %	12.5%	87.5%	100%
	Column %	8.3%	53.8%	32%
Art & Science	Count	6	4	10
	Row %	60%	40%	100%
	Column %	50%	30.8%	40%
Business Management	Count	5	2	7
	Row %	71.4%	28.6%	100%
	Column %	41.7%	15.4%	28%
Total	Count	12	13	25
	Row %	48%	52%	100%
	Column %	100%	100%	100%

The gender-subject cross-table is shown by a side – by – side chart shown as bellow.



GRAPHICAL REPRESENTATION OF DATA (OR) VISUAL REPRESENTATION OF DATA

There is another way to present statistical data in visual form. This visual display of statistical data in the form of points, lines, area and other geometrical forms and symbols, is in the most general terms known as a GRAPHICAL REPRESENTATION. Statistical data can be studied with this method without going through figures, presented in the forms of tables.

Such visual representation can be divided into two main groups, Graphs and Diagrams. The basic difference between a graph and diagram is that a graph represents the data by a continuous curve, usually shown on a graph paper, while diagram is any other one, two or three-dimensional form of visual representation.

DIAGRAMMATIC REPRESENTATION OF DATA

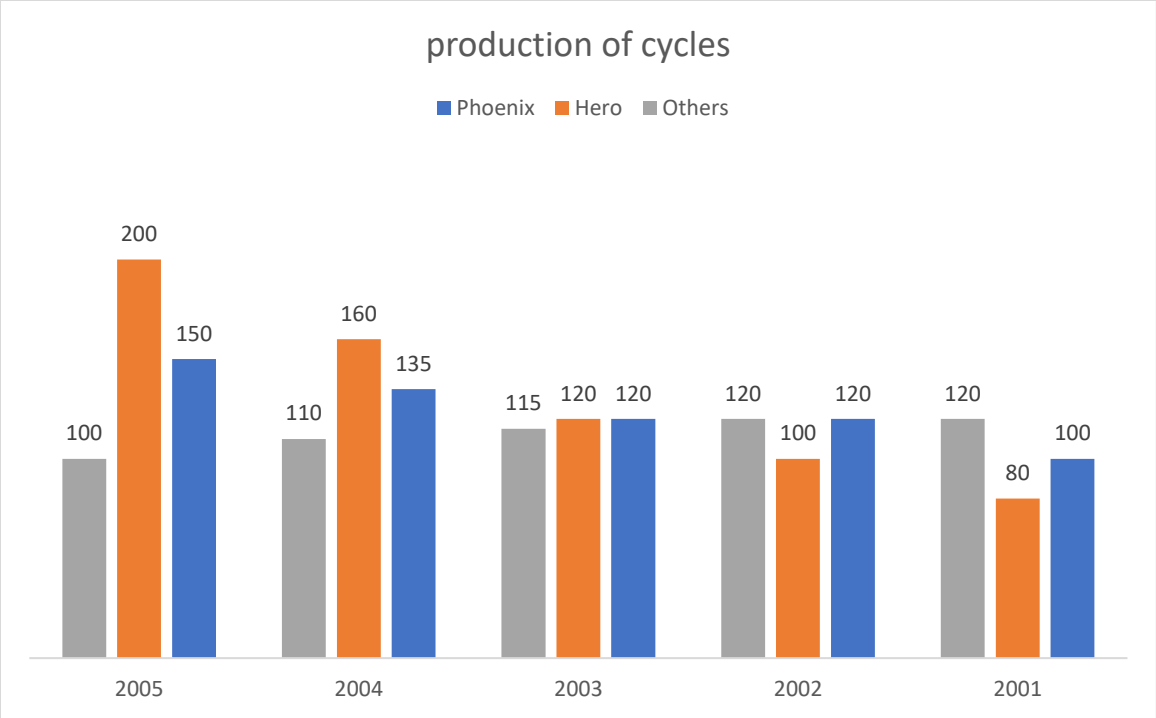
BAR DIAGRAM

A diagram in which the solid bars having length proportional to the numerical value are selected, the BAR DIAGRAME may further be divided in to the following types:

I. Simple Bar Diagram (Simple Bar Chart)
II. Component Bar Diagram (Component Bar Chart)
III. Multiple Bar Diagram (Multiple Bar Chart)

I. SIMPLE BAR DIAGRAM

A simple bar chart consists of horizontal or vertical bars of equal width and length proportional to values they represent. As the basis of comparison is one-dimensional, the widths of these bars have not mathematical significance but are taken in order to make the chart look attractive. Let us consider an example:



II. COMPONENT BAR DIAGRAM

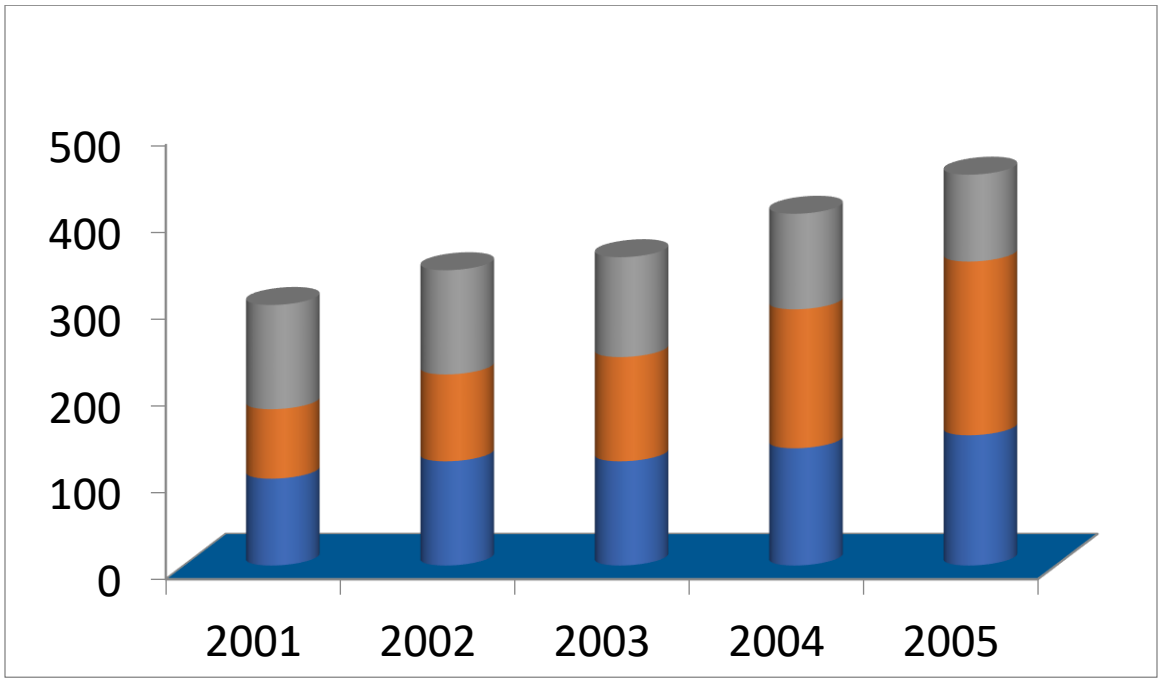
This kind of bar chart consist of solid bars having length proportional to the numerical value are selected and these bars may further be divided into components of which the numerical value is the total.

EXAMPLE: In the table give below are the data on production of cycles of various makes in china in different years (in thousands).

Years	Phoenix	Hero	Others	Total
2001	100	80	120	300
2002	120	100	120	340
2003	120	120	115	355
2004	135	160	110	405
2005	150	200	100	450

SOLUTION:

The bars having the proportional lengths of 300, 340, 355, 405, & 450 are selected and then each bar has been divided into three components in the given diagram.



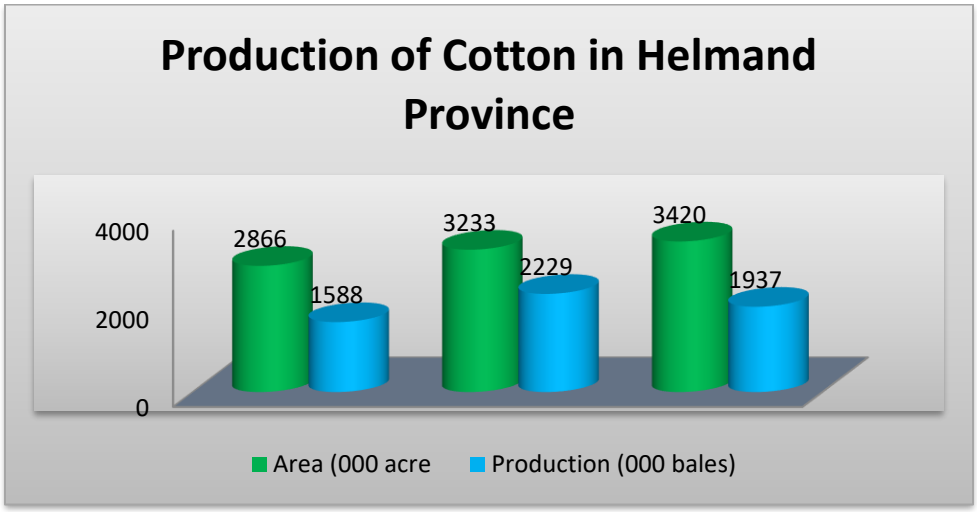
This diagram is known as a component bar diagram.

MULTIPLE BAR DIAGRAM

A multiple bar chart shows two or more characteristics corresponding to the values of a common variable in the form of grouped bars, whose lengths are proportional to the values of the characteristics and each of which is shaded or coloured differently to aid identification. This is a good device for the comparison of two or three kinds of information. For example; imports, exports, and production of a country can be compared from year to year by grouping the three bars together.

EXAMPLE: Draw multiple bar charts to show the area and production of cotton in the Helmand province from the following data:

Years	Area (000 acre)	Production (000 bales)
1965 – 66	2866	1588
1970 – 71	3233	2229
1975 – 76	3420	1937



QUESTION:

The basic difference between component bar chart and multiple bar charts is; the component bar chart should be used when we have information regarding totals and their components.

FOR EXAMPLE:

The total number of male students out of which some are Pashto medium and some are English medium. The number of Pashto medium male students and the number of English medium male students add up to give us the total number of male students. On the contrary, in the example of exports and imports, the imports and exports do not add up to give us the totality of someone thing!

RECTANGLES & SUB-DIVIDED RECTANGLES

The area of rectangle is equal to the product of its length and breadth. To represent a quality by rectangle, both length and breadth of the rectangle are used. Sub-divided rectangles are drawn for the data where the quantities along with their components are to be computed. These diagrams are generally drawn to compare the budgets of various families. In the construction of sub-divided rectangles, we are required to:

Change each component into the percentage of the corresponding total.

Draw one rectangle for each total taking equal lengths (100 units), and breadths proportional to the total.

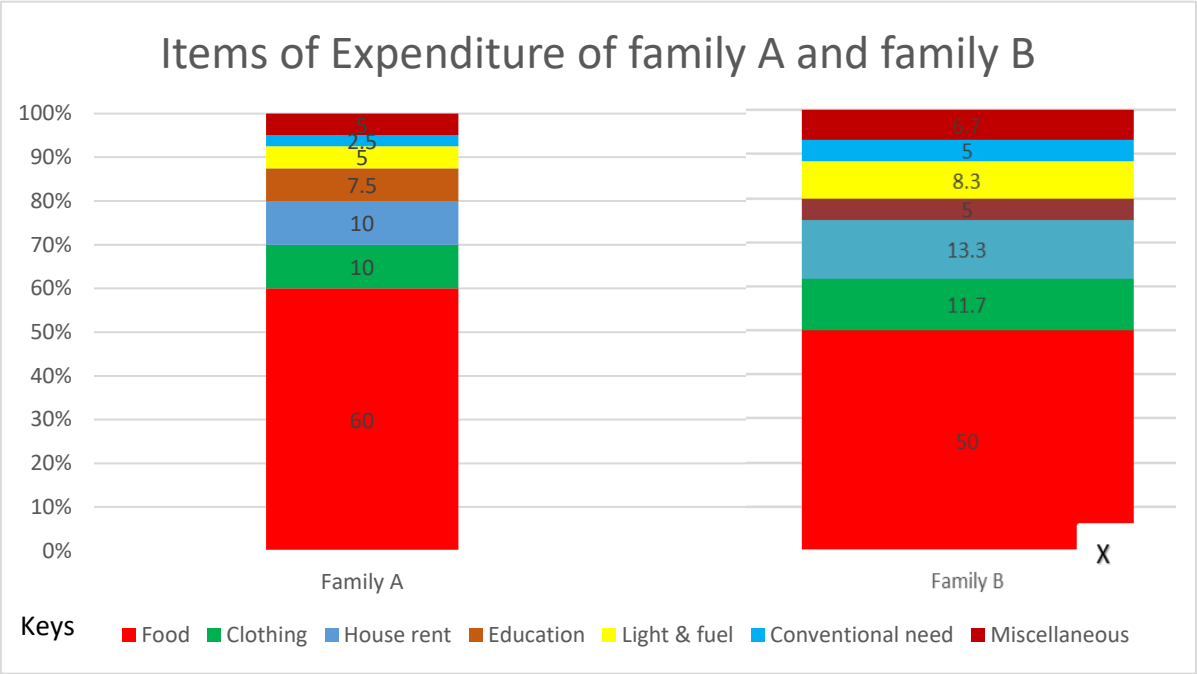
Divide every rectangle so drawn into parts equal in number to the number of components. Each part shaded or coloured will represent percentage size of the component.

EXAMPLE: Compare the budgets of family – A & family – B with suitable diagrams:

Items of Expenditure	Family - A	Family - B
Food	24	60
Clothing	4	14
House Rent	4	16
Education	3	6
Light & Fuel	2	10
Conventional Needs	1	6
Miscellaneous	2	8

The necessary computation required for the drawing of subdivided rectangles, are given below and the diagram is shown below:

Items of Expenditure	Family - A		Family – B	
	Actual Expenditure	Percentage Expenditure	Actual Expenditure	Percentage Expenditure
Food	24	60	60	50
Clothing	4	10	14	11.7
House Rent	4	10	16	13.3
Education	3	7.5	6	5
Light & Fuel	2	5	10	8.3
Conventional Needs	1	2.5	6	5
Miscellaneous	2	5	8	6.7
Total	40		120	



PIE DIAGRAM (PIE – CHART)

A diagram of circles in which the area of the circles is proportional to the numerical value is known as Pie Diagram.

As the area of the circle is equal to πr^2, so we take the square root of the numerical value as the radius of the circle, and thus this circle represents the numerical value. The circle is further divided into angles corresponding to the components, so that the total of the area is 360o.

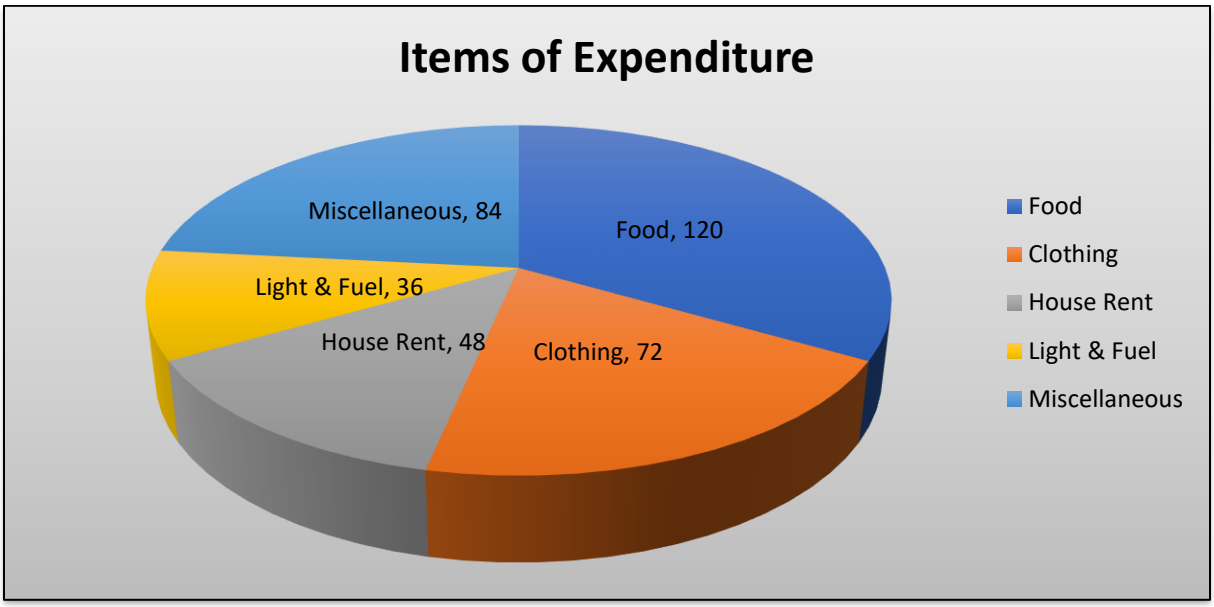
A data which is presented by Pie – Diagram can also be represented by Bar – Diagram. This is only another way of representing the data diagrammatically.

For example; represent the total expenditure and expenditure on various items of a family by Pie – Diagram.

Items:	Food	Clothing	House Rent	Fuel & Light	Miscellaneous	Total
Expenditure:	50	30	20	15	35	150

Angles are calculated by the following formula: Angle = Component Part* 360o / Whole Quantity. The corresponding angles needed to draw the chart are computed below:

Items of Expenditure	Expenditure	Angles of the Sectors (In Degree)
Food	50	120º
Clothing	30	72º
House Rent	20	48º
Light & Fuel	15	36º
Miscellaneous	35	84º
Total	150[whole Quantity]	



CONSTRUCTION OF GRAPHS

In the construction of Graphs, the first step is to make a starting point known as the origin. In the left – hand bottom corner of the graph paper.

Two straight lines perpendicular to each other are drawn through the origin. The horizontal line is called the X – axis or abscissa and the vertical line is called the Y – axis or ordinate. The two lines together are known as co – ordinate axis. Some suitable scales are selected along X – axis and Y – axis. Points are plotted and joined to get the required graph. While constructing a graph, the following points should be kept in mind.

- ✓ A scale and the form of presentation are to be selected in such a way that the true impression of the data is to be represented is giver by the graph.
- ✓ Every graph must have a clear and comprehensive title at top. Where necessary sub – title should be added.
- ✓ The source of the data must be given. A key and footnotes should be provided when necessary.
- ✓ The independent variable should always be placed on the horizontal axis.
- ✓ The vertical scale should always begin with zero; otherwise the graph will give a false impression.
- ✓ The horizontal axis does not have to begin with zero, unless of course, the independent variable or the lower limit of the first class – interval is zero.
- ✓ The axis of the graph should be properly labeled. Labels should clearly state both, the variable and units.e.g. “Distance” and “Kilometers”, “Scale” and “Afghanis” etc.
- ✓ Curves if more than one must be clearly distinguished either by different colours or by differentiated lines (solid, dashed, dot – dashed).
- ✓ The graph should not be loaded with too many curves.

TYPES OF GRAPHS

Graphs can be divided into two main categories namely:

- ✓ Graph of Time – series or Graph of Historical data and;
- ✓ Graph of Frequency Distribution;

GRAPH OF TIME SERIES (HISTORIGRAM)

A curve showing changes in the value of one or more items from one period of time to the next is known as the graph of a time series. This curve is also called a Historigram.

This Histogram display the variations in time series data dealing with prices, production, imports, exports, population etc. To construct Historigram, time is taken along X – axis and the values of variable along Y – axis. Points are plotted and are then connected by means of straight line segments to get the “HISTORIGRAM’.

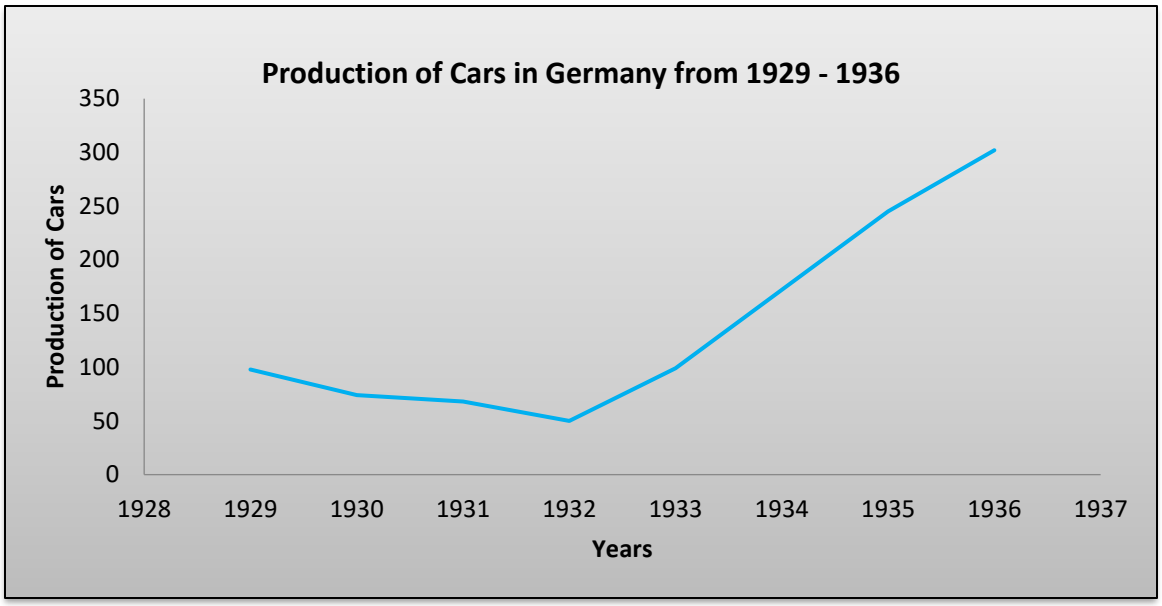
EXAMPLE:

The following table gives the number of cars produced in Germany during the years; 1929 – 1936. Draw a suitable graph, i.e. Historigram of the series:

Years:	1929	1930	1931	1932	1933	1934	1935	1936
No. of Cars:	98	74	68	50	99	172	245	302

SOLUTION:

The Historigram is drawn for the data by taking years on horizontal axis and the number of cars on vertical axis as below:



DESCRIPTIVE STATISTICS

WHAT IS DESCRIPTIVE STATISTICS: Descriptive statistics are numbers that are used to summarize and describe data. There are two different types of descriptive statistics: (a) measures of central tendency (Measures of Location) and (b) measures of dispersion (Measures of Variation).

MEASURE OF CENTRAL TENDENCY

A value which is used in this way to represent the distribution is called an average. Since the averages tend to lie in the centre of the distribution, they are called measure of central tendency. They are also called measure of location because they locate the center of the distribution.

TYPE OF AVERAGE

The most commonly used averages are:

- ✓ THE ARITHMETIC MEAN
- ✓ THE GEOMETRIC MEAN
- ✓ THE MEDIUM (QUARTILES, DECILES & PERCENTILES)
- ✓ THE MODE

THE ARITHMETIC MEAN

The arithmetic mean is a defined as a value obtained by dividing the sum of the values by their number. It is donated by $\bar{x}$ (x – bar) and is given by the formula:

$$\bar{x} = \frac{\sum x}{n} [for\ ungrouped\ data]$$

$$\bar{x} = \frac{\sum fx}{\sum f} [for\ grouped\ data]$$

n= number of values in the data

$\sum$ = Sigma= sign of summation

f = frequency

CALCULATION OF ARITHMETIC MEAN (اوسط حسابی) BASED ON UNGROUPED DATA

EXAMPLE: the values are:

25,26,27,11,15,21,12,25,26,41,5,12,0,29,33,7,20,32,14,38,27,11,43,40,45

$$\bar{x} = \frac{\sum x}{n} = \frac{578}{25} = 23.12$$

CALCULATION OF ARITHMETIC MEAN BASED ON GROUPED DATA

EXAMPLE:

Classes	F	x	f(x)
0 - 9	3	4.5	13.5
10 - 19	6	4.5	87
20 - 19	9	24.5	220.5
30 - 39	4	34.5	138
40 - 49	3	44.5	133.5
	$25 = \sum F$		$592.5 = \sum f(x)$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

COMBINED ARITHMETIC MEAN

If $\bar{x}_1, \bar{x}_2, \bar{x}_3$ ..... $\bar{x}_n$ be the arithmetic mean of k-subgroups of data with respective frequencies $n_1, n_2, n_3 \ldots\ldots n_k$. Then the combined mean $\bar{x}_c$ is defined by:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + \ldots\ldots + n_k\bar{x}_k}{n_1 + n_2 + n_3 + \ldots\ldots + n_k}$$

$$\bar{x}_c = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

THE WEIGHTED ARITHMETIC MEAN:

The multipliers of a set of numbers, which express more or less adequately the relative importance of various values in a set of data are technically called the weight.

We assign weights w_1, w_2, w_3 ..... w_n to the values in a set of data according to their relative importance, when the values are not of equal importance. The weighted mean denoted by "$\bar{x}_w$" of set of n-values x_1, x_2, x_3 ..... x_n with corresponding weights w_1, w_2, w_3 ..... w_n is then defined as:

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

THE GEOMETRIC MEAN

The G.M of n-positive values is defined as the nth root of their product. In other words, it's obtained by multiplying together all the n-values and then taking the nth root of their product. Thus

G.M= $\sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$

This method of calculating G.M satisfactory only if there are two or three values, but if the number of values (n) is large the problem of compiling the nth-root of the product of values by the above method is tedious work. To facilitate the computation of Geometric Mean, we make use of logarithms. The above formula when reduced to its logarithms form can be written as:

$$\text{G.M=Antilog}\left[\frac{\sum \log x}{n}\right] \left[for\ ungrouped\ data\right]$$

$$\text{G.M=Antilog}\left[\frac{\sum f \log x}{\sum f}\right]\left[for\ grouped\ data\right]$$

[The geometric mean is appropriate to average ratio and rates of change]

THE MEDIAN

This is the positional average of a set of data. Median is a value which divides an arrayed (ترتیب شوی) set of data into two equal halves. The number of values greater than the median is equal to the number of values smaller than the median. It's given by formula:

$$\text{Median} = \frac{(n+1)}{2} \text{ th value [for ungrouped data]}$$

$$\text{Median} = l + \frac{h}{f}\left(\frac{n}{2} - C\right) \text{ [for grouped data]}$$

Note:

Median=50%

Quartiles=25%

Deciles=10%

Percentiles=100%

Where: L= lower class-boundary of the median class

h= class- interval of the median class

f= frequency of the median class

C= cumulative (جمع شونده) frequency of the class preceding the median class

n= total frequency $= \sum f$

OTHER PARTITIONED VALUES

QUARTILES

Quartiles are the values which divide an arrayed (ترتیب شوی) set of data into four equal parts. The first, 2nd and 3rd quartiles are donated by Q1, Q2 and Q3 respectively (بترتیب) they are given by the formula:

In case of ungrouped data;

$$Q_1 = \frac{(n+1)}{4} \text{ Th value [first or lower quartile]}$$

$$Q_2 = \frac{2(n+1)}{4} \text{ Th value [second quartile]}$$

$Q_1 = \frac{3(n+1)}{4}$ Th value [3rd quartile or upper quartile]

Where “n” denotes the number of values

In case of grouped data:

$$Q_1 = l + \frac{h}{f}(\frac{n}{4} - C)$$

$$Q_2 = l + \frac{h}{f}(\frac{2n}{4} - C)$$

$$Q_3 = l + \frac{h}{f}(\frac{3n}{4} - C)$$

Where;

$l =$ Lower class boundary of the quartile class

$h =$ Class interval of the quartile class

$f =$ Frequency of the quartiles class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the quartile class

DECILES

Deciles are the values which divide an arrayed set of data into ten equal parts. The 1st, 2nd, 3rd, and 9th deciles are donated by $D_1, D_2, D_3 \ldots D_9$ respectively. They are given by formula:

In case of ungroup data:

$D_1 = \frac{(n+1)}{10}$ th value

Where “n” denotes the number of values in a set of data

$D_2 = \frac{2(n+1)}{10}$ th value

In case of grouped data;

$D_3 = \frac{3(n+1)}{10}$ th value

:

$D_9 = \frac{9(n+1)}{10}$ th value

$$D_1 = l + \frac{h}{f}(\frac{n}{10} - C)$$

$$D_2 = l + \frac{h}{f}(\frac{2n}{10} - C)$$

$$D_3 = l + \frac{h}{f}(\frac{3n}{10} - C)$$

:

$$D_9 = l + \frac{h}{f}(\frac{9n}{10} - C)$$

Where is:

$l =$ Lower class boundary of the docile class

$h =$ Class interval of the docile class

$f =$ Frequency of the docile class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the deciles class

PERCENTILE

Percentiles are the values which divide an arrayed set of data into hundred equal parts. The first, 2nd, 3rd … and 99th percentiles are donated by P1, P2, P2 … P99 respectively they are given by the formula:

IN CASE OF UNGROUPED DATA	IN CASE OF GROUPED DATA
$P_1 = \frac{(n+1)}{100}$ th value	$P_1 = l + \frac{h}{f}(\frac{1n}{100} - C)$
$P_2 = \frac{2(n+1)}{100}$ th value	$P_2 = l + \frac{h}{f}(\frac{2n}{100} - C)$
$P_3 = \frac{3(n+1)}{100}$ th value	$P_3 = l + \frac{h}{f}(\frac{3n}{100} - C)$
:	:
$P_{99} = \frac{99(n+1)}{100}$ th value	$P_{99} = l + \frac{h}{f}(\frac{99n}{100} - C)$

Where:

$l =$ Lower class boundary of the Percentile class

$h =$ Class interval of the Percentile class

$f =$ Frequency of the Percentile class

$n =$ Total frequency= $\sum f$

$c =$ Cumulative frequency of the class preceding the Percentiles

THE MODE (OR) MODEL VALUE

The French word “mode” meaning fashion has been adopted (اختیارول) to convey the idea of “most frequent”. The mode is defined as a value which accurse most frequently in a set of data. A set of data may have more than one mode or no mode at all.

CALCULATION OF MODE BASED ON UNGROUPED DATA

$$Mode = l + \frac{(f_m - f_1) \times h}{(f_m - f_1) + (f_m - f_2)}$$

Where:

$l =$ Lower class boundary of the modal class

$f_m =$ Maximum frequency (model class)

$f_1 =$ Frequency of the class preceding the model class

$f_2 =$ Frequency of the class following the model class

$h =$ Class interval of the model

EMPIRICAL RELATION BETWEEN MEAN, MEDIAN AND MODE

In a symmetrical distribution the values of the mean, median and mode are coin side (same). But if these values differ the frequency distribution is said to be skewed or asymmetrical.

For a moderately skewed distribution there exist an empirical relationship among the mean, median and mode.

$$\text{Mean} = \frac{1}{2}(3 \text{ Median} - \text{Mode})$$

$$\text{Median} = \frac{1}{3}(2 \text{ Mean} + \text{ Mode})$$

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$