**http://mail.bracuniversity.net/openwebmail/images/openwebmail.gifWhat is “Statistics”?**

It is difficult to define statistics in a few words; since its dimension, scope, function, use and importance are constantly changing over time. No formal definition thus has emerged so far and no definition is perhaps beyond controversy.

According to Fisher (1947)[[1]](#footnote-1), the science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics, applied to observational data.

Croston and Cowden (1948) defined statistics as the subject of collection, presentation and analysis of numerical data.

As Yule and Kendal (1950) opined, Statistics means quantitative data, which are affected to a marked extent by multiplicity of causes.

American Heritage Dictionary defines statistics as: “The mathematics of the collection, organization and interpretation of numerical data especially the analysis of population characteristics by inference form sampling.”

**Types of Statistical Applications:**

The field of statistics consists of two branches –

**Descriptive statistics** focuses on collection, summarization, presentation and analysis of the data using suitable numerical and graphical methods to look for patterns in a data set.

**Inferential statistics** utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data (population).

**Uses and importance of Statistics and Statisticians:**

The scope and uses of statistics are so wide and universal that they can’t be enumerated instantly in a few words. Statistics has now been recognized as a separate discipline of human knowledge in its own right.

Statistics has it extensive application in the following fields:

1. **Surveys:**

* Determine which political candidate is more popular.
* Discover what foods teenagers prefer for breakfast
* Estimate the number of children living in a given school district
  1. **Government Operation:**

Government often conducts experiments to aid in the development of public policy and social programs. Such experiments include:

* + - Consumer price
    - Fluctuations in the economy
    - Employment patterns
    - Population trends
    - Opinion polls.
  1. **Scientific research:**

Statistical sciences are used to enhance the validity of inference in all the fields of science, medical science etc. Such as:

* + Radio carbon dating to estimate the risk of earthquakes.
  + Clinical trials to investigate the effectiveness of new treatments.
  + Field experiments to evaluate the irrigation methods.
  + Measurements of water quality
  1. **Business and Industry:**

Statisticians using statistical tools quantify the unknowns in order to optimize resources. They:

* + Predict the demand for product and services.
  + Check the quality of items manufactured in a facility
  + Manage investment portfolios. And so on.

**Statistics in the Business World:**

In the business world, statistics has four important applications:

* To summarize business data
* To draw conclusions from that data
* To make reliable forecasts about business activities
* To improve business process.

**Some Basic Vocabulary of Statistics:**

**Population:**

A set of all values or elements defined on some common characteristics is called a population.

Thus population means an aggregate of elements possessing certain characteristics of interest in any particular investigation or enquiry. A population consists of all the items or individual about which researcher want to draw a conclusion.

‘N’ denotes the size of population.

Example: If we want to study the average weight of the student of 1st semester BBA then the set that consists of all the weights of the student of 1st semester BBA will be the population in this case.

**Parameter**

A parameter is a numerical measure that describes a characteristic of a population.

**Sample:**

A small and representative (desirably) part of population is known as sample.

In many particular situations it is impossible or even impractical to study the whole population, in such case only a small and representative part of population is taken under consideration to draw inferences about the population by analyzing that part of population. Such a part of population is known as sample.

Sample size is denoted by ‘n’.

**Statistic:**

A statistic is a numerical measure that describes a characteristic of sample.

**Variable:**

The measurement of elements of a population having certain characteristics may vary from element to element either in magnitude or in quality. These measurable characteristics are called variables.

Thus a measurable characteristic, which can vary from element to element with in its domain called a variable. Usually we denote the variables by capital letters and their values by small letters.

Example: Height, weight, age, SSC and HSC marks, family size, sex, etc. are some variables of 1st semester BBA students of BRAC University.

**Types of Variables**

There are two basic types of variables -

1. **Qualitative variable (also known as categorical variable or attribute)**

A qualitative variable is one for which numerical measurement is not possible. In other word when the characteristic being studied is nonnumeric, it is called a qualitative variable or an attribute.

For example: Hair color (brown, black, white etc.), religion (Muslim, Hindu, etc.), sex (male, female), home district (Dhaka, Rajshahi, Bogra etc.), occupational status (employed, unemployed, self-employed, others) etc.

An individual is simply assigned to any one of the several mutually exclusive categories on the basis of observation on the individual. The qualitative observations can neither meaningfully ordered nor physically measured, these can only be classified and then enumerated.

In dealing with the qualitative data, researchers are usually interested in how many or what proportion fall in each category.

For Example:

* + - What percent of students of BRAC Universities of English medium background?
    - What proportion of people opted in favor of construction of the new Airport?
    - How many Muslims and how many Hindus are there in Bangladesh?

1. **Quantitative variable (also known as numerical variable)**

A quantitative variable is one for which the resulting observations have numeric value and thus possesses a natural ordering. Quantitative (numeric) variable are further subdivided as discrete and continuous variables.

* **Discrete variable:**

When a variable can assume only isolated values within a given range is called discrete variable.

Example: Number of children in a family, number of road accident in a year, number of phone call received in a phone booth, height of nails etc.

* **Continuous variable:**

A variable is said to be a continuous variable if it can theoretically assume any value within a given range or ranges.

Example: height of a person –since it can take any value between 5.6 feet and 5.8 feet.

Flow Chart of Types of Variables

**Exercise:**

1. Classify each variable as qualitative or quantitative:
2. Marital status of nurses in a hospital
3. Time it takes to run a marathon
4. Weights of lobsters in a tank in a restaurant
5. Colors of automobiles in a shopping centre parking lot
6. Ages of people living in a personal care home
7. Classify each variable as discrete or continuous:
8. Number of pizzas sold by Pizza Express each day
9. Lifetimes (in hours) of 15 iPod batteries
10. Weights of the backpacks of first graders on a school bus
11. Number of students each day who make appointments with a mathematics tutor at a local college
12. Blood pressures of runners in a marathon

**Data:**

Numerical facts gathered from a statistical investigation are called a data.

In a statistical analysis the first work is to collect data the raw materials of statistics after identifying a specific problem and field of enquiry.

*Data* is in fact the *plural form* of *‘datum’*. Single information of a phenomenon on any subject of interest is called a datum. So data is called the collection of datum.

Example: If we are interested about the height of the students of 1st semester in BBA of BU, then a single value (that is the height of a student) is called a datum, and the set of all values of height will be data.

**Sources and Types of data:**

Based on the sources data can be of two types.

**Primary data:**

A data is said to be primary data if it is obtained from an investigation conducted for the first time. Thus the data collected for the first time by the investigator as original data are known as primary data.

**Secondary data:**

When a statistical analysis is conducted on a data set available from a prior investigation is called a secondary data.

Example: National income data collected by the government are primary data but they become secondary data for those who use them.

**Raw data:**

In any statistical investigation, when data first collected usually appear in raw form where, information has been recorded merely in arbitrary order in which they happened to occur. This is known as the raw data set.

Raw data, collected for any statistical investigation, is unable to represent the summaryinformation, which are although preliminary but necessary for analyses with advanced statistical method. So it is necessary to represent the raw data in such a way, which will enable us to extract the preliminary ideas about the variable(s) under study, to get some summary measures and also to perform further statistical analysis.

**Dealing with Raw Data: How to prepare data for further Statistical operation**

In the next few subsequent segments we are going to discuss on some techniques of statistics that we usually used to condense raw data, to make the data prepared for further statistical application.

The most frequently used methods for data condensation or/and representation are

1. Classification
2. Tabulation
3. Graphical representation

**Classification:**

Classification is the process of arranging data values of a variable in groups or classes according to their affinities or of our interest. It is the first step towards further processing of a heterogeneous mass of data in to a number of homogeneous groups and subgroups by their respective characteristics.

**Purpose of classification:**

Classification is necessary to serve the following purpose:

1. To eliminate unnecessary details.
2. To bring out clearly point of similarity and dissimilarity.
3. To enable one to form mental picture of the object.
4. To enable one to make comparisons.
5. To pin point the most significant features of the data at glance.
6. To enable a statistical treatment of the collected data.

**Principles of determining the number of classes / determination of number of classes:**

Usually we determine the number of classes in the light of the following conjoined considerations

1. The number of observations of a variable.
2. The lowest and highest value of a variable.
3. Even distribution of the values with in classes.
4. A regular sequence of frequencies.
5. Avoidance of extremely large or small number of classes.

**Do your Self**

* 1. Write down the Limitations of Statistics.
  2. Discuss the Importance / Scope / Uses of Statistics.

**Tabulation:**

A statistical method of data condensation by which we can represent summary information of one or more variables, is defined as tabulation.

A statistical table is the logical listing of collected data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, terms and statements in the form of titles, headings and notes which make clear the full meaning of data and their origin.

**Principles of the constructions of a table:**

Some of the most basic principles that one should consider in constructing table are as follows:

1. The table should be self-explanatory. The title describing the contents of the table should be clear, concise and to the point.
2. The table should be as simple as possible. Two or three tables are often preferable to a large table containing too many details and variables.
3. The specified units of measurements for the data should be given.
4. Necessary code or symbols used in table should be explained in a footnote.
5. Sources of data should be mentioned.

**Frequency distribution:**

The number of times a particular value of a certain variable occurs in a set of observations is called the frequency of that value and the manner in which the frequencies are distributed in the different classes is known as the frequency distribution of the values of that variable.

That is frequency distribution can be defined as a summary presentation of a number of observations of an attributes or values of a variable arrange according to their occurrence either individually (in case of discrete data) or in a range (in case of both discrete or continuous data)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 01: Frequency distribution of number of children per family**   |  |  | | --- | --- | | Number of children | Number of families | | 0 | 10 | | 1 | 27 | | 2 | 15 | | 3 | 18 | | 4 | 9 | | **Table 02: Frequency distribution of height of trees in Sundarban**   |  |  | | --- | --- | | Height of the tree  (In Feet) | Number of trees | | 0-50 | 1000 | | 50 – 100 | 2735 | | 100 – 150 | 1589 | | 150 – 200 | 1518 | | More than 200 | 719 | |

**Class limit:**

Class limits are the highest and the lowest values that can be included in the class.

For example if we consider the class 50 – 100, here 50 is the lower limit and 100 is the upper limit. In such case no values greater than 100 shall fall into that class. Similarly no values less than 50 shall fall into that class either.

**Class interval:**

The difference between the upper limit and the lower limit of a class is called the class interval.

Class interval is usually denoted by c, , h or w.

For example the class interval of the class ‘50 – 100’ is 50.

**Class frequency:**

The number of observation falling with in a particular class is called its frequency or class frequency.

**Class midpoint or class mark:**

The value of the variable that lies in the middle of the upper and lower limits is called mid value or midpoint of the class.

It can be obtained as follows:

Class midpoint; Where = Upper limit of the class, = Lower limit of the class

**Relative frequency (also known as proportion):**

Instead of presenting the frequencies in absolute terms, it is sometimes convenient to express the frequencies in percentages. The relative frequency (also known as proportion) corresponding to a class is simply the ratio of the total number of items in that class to the total number of elements in the total set.***Multiplying relative frequency by 100 one can obtain the percentage of observation that belongs to any particular class***.

**Cumulative frequency:**

The cumulative frequency corresponding to a class is the total of all frequency up to and including that class.

Example: let us consider the following table showing the distribution of mark of 20 students

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class limit  *Fill up the blank shaded area* | Class mid value | Frequency | Relative frequency | Cumulative frequency | Cumulative relative frequency |
| 0 – 10 | 5 | 4 | 0.148 | 4 | 0.148 |
| 10 – 20 | 15 | 8 | 0.296 | 4+8 | 0.444 |
| 20 – 30 |  | 5 |  | 4+8+5 |  |
| 30 - 40 |  | 4 |  |  |  |
| 40 – 50 |  | 3 |  |  |  |
| 50 – 60 |  | 2 |  |  |  |
| 60 – 70 | 65 | 1 |  |  |  |
| Total | |  |  |  |  |

**Exercise:**

The following information, extracted from a survey of a Microfinance institution (MFI) represents the amount of loan request of 50 potential borrowers from any particular branch of that MFI.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1850 | 9250 | 6100 | 4500 | 5100 | 1800 | 6100 | 6500 | 6999 | 6780 |
| 3100 | 7475 | 6400 | 4950 | 8789 | 6100 | 6480 | 7050 | 9900 | 4790 |
| 4400 | 7900 | 6900 | 3865 | 5556 | 4859 | 6999 | 6780 | 8050 | 9900 |
| 5600 | 6600 | 9980 | 4800 | 8855 | 5550 | 1200 | 4790 | 6500 | 8050 |
| 3858 | 7300 | 8050 | 6200 | 7155 | 4980 | 8050 | 6480 | 7050 | 1500 |

For the given data construct a suitable frequency distribution table featuring the following components

|  |  |  |
| --- | --- | --- |
| 1. Class mid value | 1. Tally Bars | 1. Frequency |
| 1. Relative frequency | 1. Cumulative frequency | 1. Cumulative relative frequency |

Using the aforementioned information also answer the following

1. Determine the number of loan request between tk 4000-6000
2. Determine the proportion of loan request between 4000 – 6000.
3. Determine the number of loan request below tk. 7000.
4. Determine the proportion of loan request below tk. 7000.

**Tables and Charts for categorical data**

When you have categorical data, the investigators need to tally responses into categories and then present the frequency or percentage in each category in tables and charts.

**The summary Table**

A summary table indicates the frequency, amount or percentage of items in a set of categories so that you can see differences between the categories. A summary table lists the categories in one column and the frequency, amount or percentage in a different column.

The following table illustrates a summary table that asked people where they prefer to do their banking.

Table 1: Table of percentage distribution of banking preference of the customer of BANK XYZ

|  |  |
| --- | --- |
| Banking Preference | Percentage (%) |
| ATM | 16 |
| Automated or live telephone | 2 |
| Drive-through service at branch | 17 |
| In person at branch | 41 |
| Internet | 24 |

**Example 1:**

***Summary table of levels of Risk of Mutual Funds***.

A sample of 868 mutual funds has been selected and questions were asked to assess and categories the risk associated with the customer’s investments in mutual funds. Of the 868 mutual funds 202 funds are classified as the low risk funds, 311 funds are classified as average-risk fund and the rest of 355 funds are categorized as high- risk. Hence the summary table of levels of risk of mutual funds is given below.

Table 2: Frequency and Percentage Summary Table Pertaining to Risk Level for 868 Mutual Funds

|  |  |  |
| --- | --- | --- |
| Fund Risk Level | Number of funds | Percentage of funds (%) |
| Low | 202 | 23.37 |
| Average | 311 | 35.83 |
| High | 355 | 40.89 |
| Total | 864 | 100.00 |

**The Bar Chart**

In a bar chart, a bar shows each category. The length of the bar represents the amount, frequency or percentage of values falling into a category.

Figure 1 displays the bar chart for the people’s preference to do their banking as depicted in table 1. Bar chart allows researchers to compare the percentages in different categories. In figure 1: respondents are most likely to bank in person at a branch and on the internet, followed by drive through service at a branch and ATM. Very few respondents mentioned automated or live telephone.



**Example 2:**

**Bar Chart of levels of risk of Mutual Funds.**

Construct a bar chart for the levels of risk of mutual funds (based on data shown in table 2) and interpret the result.



**The Pie Chart**

The Pie Chart is a Circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

In table 1 of this lecture 16% of the respondents stated that they prefer to bank using ATM. Thus in constructing the pie chart, the 360 degrees that makes up a circle is multiplied by 0.16, resulting in a slice of the pie that takes up 57.6 degrees of the 360 degrees of the circle. In this figure, bank in person at the branch takes 41% of the pie and automated or live telephone takes only 2%.

In case of pie diagram –

Angle of the slice of pie for a particular categoryFrequency (or Percentage) of that particular item.

If the frequency / value / percentage of any component is f from the whole N then the angle of pie for that particular component is



**Which chart should one use – a bar chart or a pie chart?**

The selection of a particular chart often depends on the intention of the researcher. If a comparison of categories is most important, one should use a bar chart. If observing the portion of the whole that is in a particular category is most important, one should use a pie chart.

**Exercise:**

Using data given in table 2 construct a pie chart for the levels of risk of mutual funds and interpret the results.

1. Complete the following table.

|  |  |  |
| --- | --- | --- |
| Grades on Statistics examination | Frequency | Relative Frequency |
| A: 90 – 100 |  | 0.08 |
| B: 80 – 89 | 36 |  |
| C: 65 – 79 | 90 |  |
| D: 50 – 64 | 30 |  |
| F: Below 50 | 28 |  |
| Total | 200 | 1.00 |

1. A qualitative variable with three classes (X, Y and Z) is measured for each 20 randomly sampled from a target population. The data (observed class for each unit) are listed below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Y | X | X | Z | X | Y | Y | Y | X | X |
| Z | X | Y | Y | X | Z | Y | Y | Y | X |

1. Compute the frequency for each of the three classes.
2. Compute the relative frequency for each of the three classes.
3. Display the results, part a, in a frequency bar graph.
4. Display the results, part b, in a pie chart.
5. Assume telecommunication companies in Bangladesh spent about BDT 300 million in advertising. The spending is as follows:

|  |  |  |
| --- | --- | --- |
| **Media** | **Amount ($ millions)** | **Percentage (%)** |
| **Radio** | 20 | 6.67 |
| **Internet** | 30 | 10.00 |
| **Cinema** | 5 | 1.67 |
| **Direct mail** | 15 | 5.00 |
| **Magazines** | 35 | 11.67 |
| **Newspapers** | 65 | 21.67 |
| **Outdoor** | 45 | 15.00 |
| **TV** | 35 | 11.67 |
| **Other** | 50 | 16.67 |
|  | 300 | 100 |

1. Construct a bar chart and a pie chart.
2. Which graphical method do you think is best to portray these data?
3. The international Rhino Federation estimates that there are 25280 rhinoceroses living in the wild in Africa and Asia. A breakdown of the number of rhinos of each species is reported in the accompanying table.

|  |  |
| --- | --- |
| Rhino Species | Population Estimate |
| White rhino | 18000 |
| Black rhino | 4240 |
| Greater One-horned rhino | 2800 |
| Sumatran Rhino | 200 |
| Javan Rhino | 40 |
| Total | 25280 |

1. Construct a relative frequency table for the data.
2. Display the frequencies in a bar graph.
3. Display the frequencies in a pie chart.
4. What proportion of the 25280 rhinos are White rhinos? Black?
5. The following data set represents the scores on intelligence quotient (IQ) examinations of 40 sixth-grade students at a particular school:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 114 | 122 | 103 | 118 | 99 | 105 | 134 | 125 | 117 | 106 |
| 109 | 104 | 111 | 127 | 133 | 111 | 117 | 103 | 120 | 98 |
| 100 | 130 | 141 | 119 | 128 | 106 | 109 | 115 | 113 | 121 |
| 100 | 130 | 125 | 117 | 119 | 113 | 104 | 108 | 110 | 102 |

1. Organize the data in classes such as 90 – 100, 100 – 110 and so on.
2. Present the data set in a frequency histogram.
3. Determine the mean scores on intelligence quotient (IQ) examinations.
4. Determine the proportion of scores above the average scores.

**Organizing Numerical Data**

When the number of data values is large, one can organize data into an ordered array or a stem and leaf display to help understand the information the researcher has.

**The Ordered Array**

An ordered array is a sequence of data, in rank order, from the smallest value to the largest value.

**The Stem and Leaf Display**

To construct a Stem and Leaf plot each numerical value is divided into two parts. The leading digit(s) becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis and the leaf values are staked against each other along the horizontal axis

Stem and leaf plot is a graphical technique of representing quantitative data that can be used to examine the shape of a frequency distribution, the range of the values and point of concentration of the values. This is, in essence a display technique taken from the area of statistics called exploratory data analysis (EDA).

Tukey (1977) first proposed the technique. It allows us to use the information contained in a frequency distribution to show

* The range of score
* Concentration of scores
* The shape of the distribution
* Presence of any specific values or scores not represented in the entire data set
* Whether there are any stray or extreme values in the distribution.

**Example:**

1. The following data represented the marks obtained by 20 students in a statistics test.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 84 | 17 | 78 | 45 | 47 | 53 | 76 | 54 | 75 | 22 |
| 66 | 65 | 55 | 54 | 51 | 33 | 39 | 19 | 54 | 72 |

Use the stem leaf plot to display the data.

|  |  |
| --- | --- |
| The stem leaf plot for the given data | After arranging the stem leaf plot we get for the given data |

|  |  |
| --- | --- |
| Stem | Leaf |
| 1 | 7,9 |
| 2 | 2 |
| 3 | 3,9 |
| 4 | 5,7 |
| 5 | 1,3,4,4,4,5 |
| 6 | 5,6 |
| 7 | 2,5,6,8 |
| 8 | 4 |

|  |  |
| --- | --- |
| Stem | Leaf |
| 1 | 7,9 |
| 2 | 2 |
| 3 | 3,9 |
| 4 | 5,7 |
| 5 | 3,4,5,4,1,4 |
| 6 | 6,5 |
| 7 | 8,6,5,2 |
| 8 | 4 |

1. Form an ordered array, given the following data from a sample of n=8 midterm exam scores in math:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 63 | 99 | 68 | 72 | 79 | 83 | 71 | 62 |

1. Form an stem and leaf display, given the following data from a sample of n=7 midterm exam scores in physics:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 70 | 44 | 79 | 88 | 83 | 73 | 84 |

**Tables and Charts for categorical data**

**The Histogram**

A histogram is a bar chart for grouped numerical data in which the frequencies or percentages of each group of numerical data are presented as individual vertical bars. In a histogram there are no gaps between adjacent bars as there is in a bar chart of categorical data.

***Usually in histogram***

* The variable of interest is displayed or plotted along the horizontal (X) axis.
* Frequency or the percentage of the values per class is displayed or plotted along the vertical (Y) axis.

Example::

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1.3: frequency distribution of male and female by age group | | | | | | |
| Age group | Male | Female | Age group | | Male | Female |
| 0-4 | 3243 | 1621 | 35-39 | 936 | | 461 |
| 5-9 | 2842 | 1413 | 40-44 | 773 | | 378 |
| 10-14 | 2398 | 1192 | 45-49 | 633 | | 306 |
| 15-19 | 2125 | 1056 | 50-54 | 503 | | 240 |
| 20-24 | 1776 | 880 | 55-59 | 391 | | 184 |
| 25-29 | 1450 | 716 | 60-64 | 278 | | 130 |
| 30-34 | 1173 | 580 | 65+ | 749 | | 208 |

**The Polygon-**

**The Frequency Polygon**

In constructing frequency polygon the mid values of the class intervals of the frequency distribution are placed on the horizontal (X) axis and the corresponding frequencies are represented on the vertical (Y) axis. The co-ordinates points thus obtained joined by straight line. The left most point is to be joined with the mid value of the immediate previous interval and the right most co- ordinate point is to be joined with the mid value of the immediate next interval. Thus we obtain a polygon known as frequency polygon.

|  |  |  |
| --- | --- | --- |
| Table 1.4: Frequency distribution of male and female by age group | | |
| Marks | Midvalue | Frequency |
| 40-50 | 45 | 2 |
| 50-60 | 55 | 6 |
| 60-70 | 65 | 8 |
| 70-80 | 75 | 3 |
| 80-90 | 85 | 2 |
| 90-100 | 95 | 1 |

**The Percentage Polygon**

Constructing multiple histograms on the same graph to compare two or more data sets often gets confusing. Superimposing the vertical bars of one histogram on another histogram makes interpretation difficult. When there are two or more groups, one should use a percentage polygon.

A percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoint at their respective class percentages. The following table 1.5 and figure 1.5 illustrates the construction of the percentage polygon.

Table 1.5: Frequency distribution of Marks obtained by students

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mark Group | Mid value | Frequency of students | | | Percentage of Students | | |
| Section 2 | Section 3 | Section 5 | Section 2 | Section 3 | Section 5 |
| 40-50 | 45 | 2 | 1 | 3 | 9.1 | 4.0 | 8.8 |
| 50-60 | 55 | 6 | 8 | 5 | 27.3 | 32.0 | 14.7 |
| 60-70 | 65 | 8 | 6 | 5 | 36.4 | 24.0 | 14.7 |
| 70-80 | 75 | 3 | 5 | 11 | 13.6 | 20.0 | 32.4 |
| 80-90 | 85 | 2 | 3 | 6 | 9.1 | 12.0 | 17.6 |
| 90-100 | 95 | 1 | 2 | 4 | 4.5 | 8.0 | 11.8 |
| Total | | 22 | 25 | 34 | 100 | 100 | 100 |

**Cross Tabulations:**

The study of patterns that may exist between two or more categorical variables is common in practice. Often by cross-tabulating the data, these patterns can be explained. One can present cross tabulations in tabular form (contingency tables) or graphical from (side by side charts).

**The Contingency table**

A contingency table presents the results of two categorical variables. The joint responses are classified so that the categories of one variable are located in the rows and the categories of the other variable are located in the columns. The values located at the intersections of the rows and columns are called ***cells***. Depending on the type of contingency table constructed, the cells for each row-column combination contain the frequency, the percentage of the overall total, the percentage of the row total, or the percentage of the column total.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1.6: Frequency distribution of students by religion and sex | | | |
| Religion | Sex | | Total |
| Male | Female |
| Muslim | 25 | 20 | 45 |
| Hindu | 12 | 12 | 24 |
| Christian | 8 | 6 | 14 |
| Buddha | 5 | 3 | 8 |
| Others | 2 | 2 | 4 |
| Total | 52 | 43 | 95 |

**The Side –by – side bar chart**

|  |  |
| --- | --- |
|  |  |

A useful way to visually display the results of cross-classification data is by constructing a side by sidebar chart. Figure 1.6.1 and figure 1.6.2 uses the data from table 1.6.

**Example& Exercise:**

A sample of 500 shoppers was selected in a large metropolitan area to determine various information concerning consumer behavior. Among the questions asked was “do you enjoy shopping for clothing?” the results are summarized in the following cross classified table:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1.7: Frequency distribution of preference of shopping for clothing of the consumer | | | |
| Enjoy shopping for clothing | Sex | | Total |
| Male | Female |
| Yes | 136 | 224 | 360 |
| No | 104 | 36 | 140 |
| Total | 240 | 260 | 500 |

1. Construct contingency tables based on total percentages, row percentages and column percentages.
2. Construct a side –by side bar chart of enjoying shopping for clothing based on gender.
3. What conclusion can you draw from these analyses?

***Exercise:***

The following table represents the information of 50 individuals collected in a socio-economic survey. Using the information given in table 1 answer question A - D

Table 1: Summary information of 50 individuals

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sl. # | Sex | Religion | Previous month’s Income | Division | Marital Status |
| 1 | M | Islam | 1500 | Dhaka | Married |
| 2 | F | Hindu | 3100 | Rajshahi | Married |
| 3 | M | Buddha | 4400 | Sylhet | Married |
| 4 | M | Christian | 5600 | Khulna | Unmarried |
| 5 | F | Hindu | 3858 | Dhaka | Divorced |
| 6 | M | Islam | 9250 | Rajshahi | Married |
| 7 | M | Islam | 7475 | Chittagong | Married |
| 8 | M | Hindu | 7900 | Khulna | Unmarried |
| 9 | F | Buddha | 6600 | Rangpur | Divorced |
| 10 | F | Islam | 7300 | Dhaka | Unmarried |
| 11 | M | Islam | 6100 | Barishal | Married |
| 12 | M | Buddha | 6400 | Rajshahi | Married |
| 13 | M | Christian | 6900 | Sylhet | Married |
| 14 | F | Islam | 9980 | Khulna | Unmarried |
| 15 | M | Islam | 8050 | Dhaka | Divorced |
| 16 | M | Christian | 4500 | Rajshahi | Married |
| 17 | M | Islam | 4950 | Chittagong | Married |
| 18 | M | Hindu | 3865 | Dhaka | Unmarried |
| 19 | F | Hindu | 4800 | Rajshahi | Divorced |
| 20 | M | Buddha | 6200 | Sylhet | Unmarried |
| 21 | F | Islam | 5100 | Barishal | Married |
| 22 | M | Islam | 8789 | Rajshahi | Married |
| 23 | M | Christian | 5556 | Sylhet | Married |
| 24 | F | Islam | 8855 | Khulna | Unmarried |
| 25 | M | Buddha | 7155 | Dhaka | Divorced |
| 26 | M | Islam | 1800 | Rajshahi | Married |
| 27 | F | Islam | 6100 | Chittagong | Married |
| 28 | M | Christian | 4859 | Khulna | Married |
| 29 | M | Islam | 5550 | Rangpur | Married |
| 30 | F | Christian | 4980 | Dhaka | Unmarried |
| 31 | M | Hindu | 6100 | Barishal | Divorced |
| 32 | F | Islam | 6480 | Rajshahi | Married |
| 33 | M | Christian | 6999 | Sylhet | Married |
| 34 | M | Islam | 1200 | Khulna | Unmarried |
| 35 | F | Christian | 8050 | Dhaka | Divorced |
| 36 | F | Hindu | 6500 | Rajshahi | Unmarried |
| 37 | M | Christian | 7050 | Chittagong | Married |
| 38 | F | Islam | 6780 | Khulna | Married |
| 39 | M | Hindu | 4790 | Rangpur | Married |
| 40 | M | Buddha | 6480 | Barishal | Married |

***Question A:***

|  |
| --- |
| 1. How many variables are listed in table I? |
| 1. Mention the variable name listed in Table I. |

***Question B:***

Construct a frequency distribution table to represent the summary information of the variable “Division” and determine proportion of respondent from Dhaka.

***Question C:***

Complete the following table # 3 and answer (a) & (b)

***Table 3: Frequency distribution of sex by Religion***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sex | Religion | | | | Total |
| Islam | Hindu | Christian | Buddha |
| Male |  |  |  |  |  |
| Female |  |  |  |  |  |
| Total |  |  |  |  |  |

|  |
| --- |
| 1. What is the modal response for the variable “Sex” |
| 1. What proportion of respondents are “Buddha” |

***Question D:***

Complete the following table # 4 and answer a), b) & c)

***Table 4: Frequency distribution of previous month’s income***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Income Group | Tally | Frequency | Relative frequency | Cumulative relative frequency |
| Below – 2000 |  |  |  |  |
| 2000 – 4000 |  |  |  |  |
| 4000 – 6000 |  |  |  |  |
| 6000 – 8000 |  |  |  |  |
| 8000 - 10000 |  |  |  |  |
| 1. What proportion (Percentage) of people had previous month’s income between 2000 - 6000 | | | | |
| 1. What proportion (Percentage) of people had previous month’s income less than 4000 | | | | |
| 1. Construct Histogram to display the data represented in table 4 | | | | |

**Components of numerical measures of data**

When we speak of a data set, we refer to either a sample or to a population. If statistical inference is one’s goal, s/he will wish ultimately to use to use sample numerical descriptive measures to make inferences about the corresponding measures for population. Although a large number of numerical methods are available to describe quantitative data sets. Most of these methods measures one of the two data characteristics:

***Central tendency*** - This measures the extent to which all the values grouped around a typical or central value.

***Variation or Dispersion***–This measures the amount of dispersion or scattering of values away from a central value.

**Measures of Central Tendency**

Most sets of data show a distinct tendency to group around a central point. That is in a data set (population or sample) the values have a tendency to cluster around a certain point. This tendency of clustering the values around the center of the series is usually called central tendency. The numerical measure of this tendency of concentration is variously known as the measure of central tendency or measure of location or the measure of average.

**Necessity of measuring the central tendency:**

The necessities of measuring central tendency or average are as follows –

1. They give us an idea about the concentration of the values in the central part of the distribution.
2. It is the value of the variable, which is typical of the whole se.
3. It represents all relevant information contained in the data in as few numbers as possible.
4. They give precise information, not information of a vogue general type.

**Characteristics of a good measure of central tendency:**

The following are the characteristics of an ideal measure of central tendency

1. It should be easy to understand.
2. It should be easy to calculate.
3. It should be based upon all observations.
4. It should be rigidly defined.
5. It should be unduly affected by extreme values.
6. It should be suitable for further algebraic treatment.
7. It should be less affected by sampling fluctuation.

**Different measure of central tendency:**

The following are the different measure of central tendency

|  |  |  |
| --- | --- | --- |
| 1. Arithmetic mean | 1. Median | 1. Mode |
| 1. Geometric mean | 1. Harmonic mean | 1. Weighted mean |

**Arithmetic mean (AM):**

Adding the values of the observations and then dividing the sum by the number of observations obtain the arithmetic mean of a series of observations.

Arithmetic mean (AM) for

* sample observation is denoted by 
* Population mean is denoted by.

***Formula for Ungrouped Data***

Suppose there are n values  for a variable X, then the AM denoted by  is defined as

**Example:**

Banglatel is studying the number of minutes used by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 90 | 77 | 94 | 89 | 119 | 112 |
| 91 | 110 | 92 | 100 | 113 | 83 |

What is the mean (arithmetic mean) number of minutes used?

**Answer:**

Average use of the rate plan

Thus the arithmetic mean number of minutes used last month by the sample of cell phone users is 97.5 minutes.

**Exercise:**

1. ***“Dolphine Autos”*** employed 12 sales people. The number of new cars sold last month by the respective sales people were as given in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 15 | 23 | 10 | 4 | 18 | 8 |
| 10 | 28 | 13 | 19 | 14 | 12 |

Determine the average number of car sold by the sales people. Also determine the proportion of sales people performing below average.

1. During the last month Shameem Refrigeration and Air Conditioning Company completed 129 different assignments for their clients and earned mean revenue of 13449 tk per assignment. If the managing director wants to know the total revenue for the month can you compute the total revenue? What it is?
2. Following data represents the battery life (in shots) for a sample of 12 three-pixel digital cameras:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 300 | 180 | 380 | 260 | 35 | 380 |
| 85 | 170 | 460 | 120 | 110 | 240 |

Determine the average number of shots taken for each battery. Also determine the proportion of batteries performing above average.

Again for a group data as given in the following table

***Formula for***

***Grouped***

***Data***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| : |  |  | … | … |  |
|  |  |  | … | … |  |

Such that  then the AM is denoted by is defined as

**Example & Exercise:**

Calculate the mean for the following frequency distribution for n=100:

|  |  |
| --- | --- |
| Class interval | Frequency |
| 0-10 | 10 |
| 10-20 | 20 |
| 20-30 | 40 |
| 30-40 | 20 |
| 40-50 | 10 |

**Answer:**

Calculation:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class interval | Frequency | Mid values |  | Arithmetic mean  ?? |
| 0-10 | 10 | 5 | 50 |
| 10-20 | 20 | 15 |  |
| 20-30 | 40 |  |  |
| 30-40 | 20 |  |  |
| 40-50 | 10 | 45 | 450 |
| Total |  |  |  |

**Exercise:**

1. The following data represent the distribution of the age of employees within two different divisions of publishing company. Determine which company have relatively aged group of employees.

|  |  |  |
| --- | --- | --- |
| Age of employees | Number of employees of division | |
| X | Y |
| 20 – 30 | 6 | 13 |
| 30 – 40 | 19 | 30 |
| 40 – 50 | 9 | 24 |
| 50 – 60 | 10 | 0 |
| 60 – 70 | 2 | 4 |

**When to use Arithmetic Mean:**

In the following cases arithmetic mean should not be used:

1. In highly - skewed distributions.
2. In distributions with open
3. When the distribution is unevenly spread. Concentration being small or large at irregular points.
4. When an average rate of growth or change over a period of time is required.
5. When the observation are from geometric progression.
6. When averaging rates (that is speed, fluctuations in the prices of articles, etc.)
7. When there are very large and very small values of observations.

**Median (Me):**

If the values of a series are arranged in an ascending or descending order of magnitude then the middle most value in this arrangement is called the median of the series.

Median is usually denoted by Me.

**Determination of Median:**

Let n be the number of observations.

**For ungrouped data:**

***Formula for***

***Ungrouped***

***Data***

1. When ***n is odd*** the value of the th observation will be the median.
2. When ***n is even*** the median will be the AM of the values of th and th observation in the series.

**Example:**

The ages of a family of seven members are given as 12, 7, 2, 34, 17, 21 and 19. Find the median age.

|  |  |
| --- | --- |
| Step 1bd14565_ | Count the total number of elements, n=?bd14565_ Here n= 7 bd14565_ 7 is a odd number |
| Step 2bd14565_ | Arrange the values in ascending order bd14565_ 2, 7, 12, 17, 19, 21, 34 |
| Step 3bd14565_ | Median: Me = Value of th observation  = Value of th observation  = Value of 4th observation = 17 |
| Step 4bd14565_ | Median age of the family is 17 years |

**Example:**

The ages of a family of eight members are given as 12, 7, 2, 34, 17, 40, 21 and 19. Find the median age.

|  |  |
| --- | --- |
| Step 1bd14565_ | Count the total number of elements, n=?bd14565_ Here n= 8 bd14565_ 8 is a even number |
| Step 2bd14565_ | Arrange the values in ascending order bd14565_ 2, 7, 12, 17, 19, 21, 34, 40 |
| Step 3bd14565_ | Median: Me = AM of the values of th and observation  = AM of the values of \_ \_ \_ and \_ \_ \_ observation  = =? |
| Step 4bd14565_ | Median age of the family is ? ? ? years |

**For ungrouped data:**

Me is given by the formula, ****

***Formula for***

***Grouped***

***Data***

|  |  |  |  |
| --- | --- | --- | --- |
|  | = Median |  | = Frequency of the median class |
|  | = Lower Limit of the median class |  | = Width of the median class |
|  | = Cumulative frequency of the pre median class |  | = Total number of observation |

Where

***MEDIAN CLASS*** *is the class that contains th observation of the given data.*

**Example:** Table 1.6 displays summary information of the parent of 50 students. Compute the median age of woman.

|  |  |
| --- | --- |
| Table 1.6: Income distribution of the student’s of ECO 202 | |
| Income of parent  (in thousand taka) | Frequency |
| Below 20 | 3 |
| 20 – 40 | 4 |
| 40 – 60 | 6 |
| 60 – 80 | 8 |
| 80 – 100 | 12 |
| 100 – 120 | 10 |
| 120 and over | 7 |
| Total | 50 |

Hints:

***Step 1:*** Compute the cumulative frequencies.

***Step2:*** Determine**, one half of the total number of cases.

***Step 3:*** Locate the median class.

***Step 4:***Determine the lower limit () of the median class.

***Step5:***Sum the frequencies of all the classes prior to the median class. This is.

***Step 6:***Determine the frequency of the median class.

***Step 7:***Determine the width of the median class.

**You got all the quantities to compute median. So compute the median. …**

**Exercise:**

1. The following table gives the data pertaining to kilowatt hours of electricity consumed by 100 randomly selected flat owners of Japan garden city.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Consumption  (in K-watt hours) | 0-100 | 100-200 | 200-300 | 300-400 | 400-500 |
| No. of users | 6 | 25 | 36 | 20 | 13 |

Calculate

|  |  |
| --- | --- |
| 1. Mean consumption of electricity | 1. Median use of electricity |
| 1. Standard deviation of electricity consumption | 1. Skewness of electric consumption. |

1. The following data represents the amount (in thousands taka) of loan requirements of the people of two different upazilla. Using median comment on which upazilla has the greater average demand of loans.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Upazilla 1 | 42 | 12 | 26 | 18 | 9 | 35 | 28 | 39 | 8 |
| Upazilla 2 | 8 | 15 | 10 | 18 | 22 | 20 | 26 | 42 | 35 |

**When to use Median:**

The median is generally the best average in open – end grouped distribution, especially where if plotted as a frequency curve one gets a J or reverse J shaped curve.

**Mode:**

The mode is the value of the variable that occurs most frequently; that is for which the frequency is a maximum.

Mo denotes mode.

**Determination of mode:**

**For ungroup data / categorical variable** mode is the value of the variable for which the frequency is highest.

For the data sets:

1. 7, 8, 6, 7, 9, 7, and 4: Here ‘7’ appears highest 3 times, hence mode is ‘7’and the data is unimodal.

***Mode for***

***Ungrouped***

***Data***

1. 6, 4, 8, 5, 8, 1, 2, 5, 4, 7, 5, 2, 4, and 3: here ‘5’ and ‘4’ both occur highest 3 times hence the mode ‘5’ and ‘4’ and the data is bimodal.
2. 1, 5, 7, 2, 6, 9, and 4: there is no mode.
3. Consider the following table representing the frequency distribution of religion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Religion | Muslim | Hindu | Buddhist | Christian | Others |
| Frequency | 18 | 75 | 12 | 4 | 2 |

Here the highest frequency ‘75’ occurs for the category ‘Hindu’. Hence mode for the given data is \_ \_ \_ \_ \_ \_ \_.

**For grouped data** mode is obtained by using the following formula

|  |  |
| --- | --- |
|  | = Mode |
|  | = Lower Limit of the Modal class |
|  | = Frequency of the modal class |
|  | = Frequency of the pre modal class |
|  | = Frequency of post modal class |
|  | = Width of the modal class |

Where,

**When to use Mode:**

Generally speaking mode can be used to describe qualitative data. Mode is particularly useful average for discrete data.

**Exercise:**

1. The frequency distribution below represents the weights in pounds of a sample of packages carried last month by a small airfreight company.

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Frequency | Class | Frequency |
| 10.0 – 11.0 | 1 | 15.0 – 16.0 | 11 |
| 11.0 – 12.0 | 4 | 16.0 – 17.0 | 8 |
| 12.0 – 13.0 | 6 | 17.0 – 18.0 | 7 |
| 13.0 – 14.0 | 8 | 18.0 – 19.0 | 6 |
| 14.0 – 15.0 | 12 | 19.0 – 20.0 | 2 |

Find the mean, median and mode.

1. Suppose that 100 students are enrolled in a statistics class and the following are the test scores received by them:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 77 | 44 | 49 | 33 | 38 | 76 | 68 | 68 | 39 | 44 |
| 29 | 41 | 32 | 45 | 83 | 58 | 73 | 47 | 40 | 26 |
| 34 | 47 | 66 | 53 | 55 | 58 | 49 | 45 | 61 | 41 |
| 54 | 50 | 51 | 66 | 80 | 73 | 57 | 61 | 56 | 50 |
| 38 | 45 | 51 | 44 | 41 | 68 | 45 | 92 | 43 | 12 |
| 59 | 36 | 55 | 47 | 61 | 53 | 32 | 65 | 51 | 33 |
| 59 | 55 | 43 | 66 | 44 | 41 | 25 | 39 | 72 | 37 |
| 55 | 92 | 83 | 77 | 45 | 62 | 45 | 36 | 78 | 48 |
| 45 | 82 | 71 | 48 | 46 | 69 | 38 | 72 | 56 | 64 |
| 37 | 16 | 44 | 57 | 63 | 71 | 40 | 64 | 57 | 51 |

1. Organize the data in classes such as 10 – 20, 20 – 30 and so on
2. Using the above data draw histogram, frequency polygon, ogive and stem leaf plot.
3. Find the mean median and mode for the given data.
4. The following data set represents the record high temperatures in degree Fahrenheit () for each of the 50 US states:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 112 | 100 | 117 | 106 | 114 | 118 | 105 | 110 | 109 | 112 |
| 110 | 118 | 117 | 116 | 118 | 112 | 114 | 114 | 105 | 109 |
| 116 | 112 | 114 | 115 | 118 | 117 | 118 | 92 | 106 | 110 |
| 88 | 108 | 110 | 121 | 113 | 120 | 119 | 111 | 104 | 111 |
| 107 | 113 | 98 | 117 | 105 | 110 | 118 | 112 | 114 | 114 |

1. Construct a suitable frequency distribution table using interval 85 – 95, 95 – 105 and so on.
2. Determine the modal temperature.
3. Determine the proportion of states having temperature that is more than modal temperature.
4. The data given represent the ages of patients admitted to a small hospital on February 28, 2004.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 85 | 75 | 66 | 43 | 40 | 41 | 88 | 80 |
| 56 | 56 | 67 | 69 | 89 | 83 | 65 | 53 |
| 75 | 74 | 87 | 83 | 52 | 44 | 48 | 49 |

1. Construct a frequency distribution table.
2. Compute the sample mean median and mode from the frequency distribution table.
3. Compute the sample mean, median and mode from the raw data.
4. The rate of return for 30 stocks is:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8.3 | 9.6 | 9.5 | 9.1 | 8.8 | 11.2 | 7.7 | 10.1 | 9.9 | 10.8 |
| 10.2 | 8.0 | 8.4 | 8.1 | 11.6 | 9.6 | 8.8 | 8.0 | 10.4 | 9.8 |
| 9.2 | 6.5 | 8.9 | 7.4 | 12.5 | 13.8 | 8.6 | 11.2 | 10.5 | 11.2 |

Organize this information into a stem-leaf display. Hence answer the following

1. How many rates are less than 9.0?
2. Determine the mode.
3. Determine median.
4. What are the maximum and the minimum rates of return?
5. 168 handloom factories have the following distribution of average number of workers in various income groups:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Income Groups: | 800 - 1000 | 1000 - 1200 | 1200 – 1400 | 1400 – 1600 | 1600 – 1800 |
| Number of firms: | 40 | 32 | 26 | 28 | 42 |
| Average Number of Workers: | 8 | 12 | 8 | 8 | 4 |

Find the mean salary paid to the workers.

***Answer: 1228.84***

***J K Sharma, 91***

1. A class of 50 students sits for a class test. The following table gives result of the students who passed the examination:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Marks: | 40 | 50 | 60 | 70 | 80 | 90 |
| Number of Students: | 8 | 10 | 9 | 6 | 4 | 3 |

If the mean marks for all the students were 51.6, find out the mean marks of the students who failed.

***Answer: 21Marks***

***J K Sharma, 93***

1. The average declared by a group of 10 chemical companies was 18 percent. Later on it was discovered that one correct figure, 12 was misread as 22. Find the correct average dividend.

***Answer: 17 percent***

***J K Sharma, 93***

1. A company wants to pay bonus to members of the staff. The following “Table 1” demonstrates the amount to be paid as bonus and” table 2” represents the actual amount of salary drawn by the employees of that company:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 1: Monthly Bonus Policy*   |  |  | | --- | --- | | Monthly salary (in tk.) | Bonus | | 3000 – 4000 | 1000 | | 4000 – 5000 | 1200 | | 5000 – 6000 | 1400 | | 6000 – 7000 | 1600 | | 7000 – 8000 | 1800 | | 8000 – 9000 | 2200 | | 9000 – 10000 | 2300 | | 10000 - 11000 | 2400 | | *Table 2: Monthly Salary*   |  |  |  |  |  | | --- | --- | --- | --- | --- | | *3250* | *3780* | *4200* | *4550* | *6600* | | *6200* | *6800* | *7250* | *3630* | *8320* | | *9420* | *9520* | *8000* | *10020* | *10280* | | *11000* | *6100* | *6250* | *7630* | *3820* | | *5400* | *4630* | *5780* | *7230* | *6900* | |

For the given information determine –

1. How much would the company need to pay by way of bonus?
2. What shall be the average bonus paid per member of the staff?

***Answer: tk. 42000 and tk. 1680***

***J K Sharma, 99***

1. The mean of 200 observations was 50. Later on, it was found that two observations were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

***Answer: 50.9***

***J K Sharma, 93***

1. There are two units of a garment in two different cities employing 760 and 800 persons, respectively. The arithmetic means of monthly salaries paid to persons in these two units are tk 18750 and tk. 16950 respectively. Find the combined arithmetic mean of salaries of the employees in both the units.

***Answer: tk. 17827 (appx.)***

***J K Sharma, 96***

1. An investor buys Tk. 12000 worth of shares of a company each month. During the first 5 months he bought the shares at a price of tk. 100, tk. 120, tk. 150, Tk. 200 and tk. 240 per share respectively. After 5 months what is the average price paid for the shares by the investor.

***Answer: tk. 146.34 (appx.)***

***J K Sharma, 99***

1. The mean yearly salary paid to all employees in a company is tk. 2400000. The mean yearly salaries paid to male and female employees are tk. 2500000 and tk. 1900000 respectively. Determine the percentage of male to female employees in the company.

***Answer: Male 83.33% and Female 16.67%***

***J K Sharma, 97***

1. The mean monthly salaries paid to 100 employees of a company were tk. 5000. The mean monthly salaries paid to male and female employees were tk. 5200 and tk. 4200 respectively. Determine the percentage of males and females employed by the company.

***Answer: Male 80% and Female 20%***

***J K Sharma,127***

1. A charitable organization decided to give Old-age pension to people over sixty years of age. The scales of pension were fixed as follows (*see* Table 1) and the ages of persons who secured the pension are given in table 2:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Pension policy**   |  |  | | --- | --- | | Age Group | Pension /Month | | 60 – 65 | 200 | | 65 – 70 | 250 | | 70 – 75 | 300 | | 75 – 80 | 350 | | 80 - 85 | 400 | | **Table 2: Actual salary drawn by employees**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | 74 | 76 | 60 | 83 | 67 | | 71 | 84 | 68 | 74 | 81 | | 75 | 61 | 61 | 66 | 79 | | 62 | 69 | 67 | 72 | 64 | | 63 | 72 | 78 | 64 | 73 | |

Determine –

1. How much money would the organization need to pay by way of pension?
2. What shall be the average pension payable person and the standard deviation?

***Answer:* and**

***J K Sharma, 160***

1. In 2014, a person spends tk. 1800 monthly on an average for the first four months and tk. 2000 monthly for the next eight months and saves tk. 5600 in that a year. Determine the person’s average monthly income.
2. The average of 11 results is 60. If the average of first 6 results is 58 and that of the last six is 63, find the sixth result.

**The Weighted Mean:**

The weighted mean is a special case of the arithmetic mean. It occurs when there are several observations of the same value.

***To explain:*** Suppose the ***Shumi’s Hot Cake*** offers three different kinds of burger packages small, medium and large for Tk. 100, Tk. 125 and Tk. 150. Of the last 10 burgers sold 3 were small, 4 were medium and 3 were large. To find the mean price of the last 10 burger packages sold we can calculate using the usual formula of the arithmetic mean as follows –



The mean selling price of the last 10 burger packages sold is Tk. 125.

An easier ways to find the mean selling price is to determine the weighted mean. In this method we multiply each observation by the number of times it happens as described below –



In this case the weights are frequency counts. However, any measure of importance could be used as a weight. In general the weighted mean of a set of numbers designated with the corresponding weights  is computed by:



Example:

Madina Construction Company pays its part time employees hourly basis. For different level of employee the hourly rate are Tk. 50, Tk. 75 and Tk. 90. There are 260 hourly employees, 140 of which are paid at Tk. 50 rate, 100 at Tk. 75 and 20 at the Tk. 90 rate. What is the mean hourly rate paid to the employees?

Answer:

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate as follows -

.

The weighted mean hourly wage is Tk. 62.69 or Tk. 63.00 (approximately).

**Example & Exercise:**

The US postal service handles seven basic types of letters and cards: 3rd class, 2nd class, 1st class, airmail, special delivery, registered and certified. The mail volume during 2004 is given in the following table

|  |  |  |
| --- | --- | --- |
| Types of mailing | gm delivered  (in millions) | Price per gm |
| 1st class | 77600 | 0.13 |
| AIR mail | 19000 | 0.17 |
| Special delivery | 1300 | 0.35 |
| Registered mail | 750 | 0.40 |
| Certified mail | 800 | 0.45 |

What was the average revenue per gm for these services during the year?

WESTECS sold 95 Executive Men’s Suits for the regular price of TK. 4,900. For the summer sale the suits were reduced to Tk. 3,500 and 126 were sold. At the final year end clearance, the price was further reduced to Tk. 2,500 and the remaining 79 suits were sold.

1. What was the weighted mean price of a WESTECS suit?
2. WESTECS paid Tk. 2000 a suit for the 300 suits. Comment on the store’s profit per suit if a salesperson received a Tk. 150 for each one sold.

**Quartile:**

If the items in a series are arranged in ascending order of their magnitudes then those values of the variable that divide the total frequency in to four equal parts are called quartiles.

There are three quartiles denoted by. The second quartile  coincides with the median. The lower quartile  is the point such that one fourth of the total frequency is less than  and three forth is greater than.

**Problem:**

For the following data compute the three quartiles.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 99 | 75 | 84 | 33 | 45 | 66 | 97 | 69 | 55 | 61 |
| 72 | 91 | 74 | 93 | 54 | 76 | 62 | 91 | 77 | 68 |

**Answer:**

Arrange the data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 33 | 45 | 54 | 55 | 61 | 62 | 66 | 68 | 69 | 72 |
| 74 | 75 | 76 | 77 | 84 | 91 | 91 | 93 | 97 | 99 |

### Hints:

### First find the median

Median ****= AM of the values of

= AM of the values of 10th and 11th observation

= 

1st quartile = median of the 1st half of observations =? ? ?

3rd quartile = median of the 2nd half of observations =? ? ?

* **Merits and demerits of different measure of central tendency -**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Merits | **Demerits** |  |
| Arithmetic mean | 1. Rigidly defined. 2. Easy to understand and calculate. 3. Based upon all observation. 4. Most amenable to algebraic treatment. 5. Not based on position in the series. | 1. Cannot be defined graphically. 2. Cannot be used in case of qualitative data. 3. Affected very much by extreme values. 4. May not occur in the series. 5. Difficult to calculate in the case of the data with open-end class. | Arithmetic mean |
| Median | 1. Rigidly defined. 2. Easy to understand and calculate. 3. Not affected very much by extreme values. 4. Can be calculated in the case of the data with open-end class. 5. Can be defined graphically. | 1. In case of even number of observations it is not defined exactly. 2. Not based on all observations. 3. Not easy for algebraic treatment. 4. For calculating median it is necessary to arrange the data either ascending or descending order. | Median |
| Mode | 1. Most typical and representative value of a distribution. 2. Not at all affected by extreme values. 3. Can be calculated in the case of the data with open-end class. 4. Easy to understand and calculate. 5. Can be defined graphically. | 1. Not clearly defined in case of bimodal or multi modal distribution. 2. Not based on all observation. 3. Not suitable for further algebraic treatment. 4. Affected by sampling fluctuations. | Mode |

Measure of Dispersion or Variation

Dispersion is the spread or scatter of item values from a measure of central tendency. Dispersion is usually measured as an average of deviations about some central value. Dispersion thus is a type of average and is sometimes called a second order average.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group 1 | 49 | 50 | 50 | 51 |
| Group 2 | 0 | 0 | 100 | 100 |

Example: let us consider two groups of students with score in a particular examination as shown in the table. The AM for each group is 50. It is clear from the data that the first group consists of near average intelligent student and the 2nd group is made up of very bright and very dull students. It is evident that the distributions of both groups have the same AM. But they differ in variation from ; such variation is usually measured by the measure of dispersion.

**Characteristics of a good measure of variation or dispersion:**

The following are the characteristics of an ideal measure of variation or dispersion

1. It should be easy to understand.
2. It should be easy to calculate.
3. It should be based upon all observations.
4. It should be rigidly defined.
5. It should be unduly affected by extreme values.
6. It should be suitable for further algebraic treatment.
7. It should be less affected by sampling fluctuation.

**Purpose of measure of dispersion or variation:**

Measure of dispersion is important for the following purpose.

1. To determine the reliability of an average.
2. To compare the variability.
3. To compare two or more series with regard to their variability.
4. To facilitate the use of other statistical measures.
5. It is one of the most important quantities used to characterize a frequency distribution.

**Types of measure of dispersion:**

Measure of dispersion or variation may be either absolute or relative. ***Absolute measure*** of variation are expressed in the same statistical unit in which the original data are given such as takas, kilograms, tones, etc. and may be used to compare the variation in two distributions, provided the variables are expressed in the same units and of same average size.

On the other hand often it is necessary to compare the distribution in two or more different frequency distributions having variables expressed in different units. In such a case dispersion is calculated by dividing the absolute measure of dispersion by a measure of central tendency. The resultant numerical value is a ***relative measure*** of dispersion.

**Different types of Absolute and Relative measure of dispersion are listed below:**

|  |  |
| --- | --- |
| Absolute measure of dispersion | Relative measure of dispersion |
| 1. Range 2. Quartile deviation 3. Mean deviation 4. Variance and Standard deviation | 1. Coefficient of range 2. Coefficient of quartile deviation 3. Coefficient of mean deviation 4. Coefficient of variation and standard deviation |

These measures are discussed below:

**Range and Coefficient of Range:**

The range of a set of data values is the difference between the highest and the lowest values in the set. If  the smallest and the largest values respectively in a set then the range “R” is defined as ****.

For group data the range is taken either as the difference between the lower boundary of the first class and the upper boundary of the last class or as the difference between the highest and the lowest mid-values.

The coefficient of dispersion corresponding to range called coefficient of range and it is obtained by

Coefficient of range = ; Where Largest value and  Smallest value

**Quartile Deviation and Coefficient of Quartile Deviation:**

Quartiles divide the observations in to four equal parts, when observations are arranged in order of magnitudes median, denoted by , is the middle most observation and & are the middle most observations of the lower and upper half respectively.

Therefore and  gives us some measure of dispersion. The AM of these two measures give us the quartile deviation and is denoted by  and is defined as



The coefficient of variation corresponding to quartile deviation is called the coefficient of quartile deviation and is defined as

Coefficient of 

**Mean Deviation or Average Deviation:**

Mean deviation is the arithmetic average of the variation of the value of individual items in the series from their central tendency.

If  denote the value of N observations then the mean deviation about an average (or measure of central tendency) A is defined as

=

In case of frequency distribution

=

The coefficient of dispersion corresponding to mean deviation is known as coefficient of mean deviation and is obtained by dividing mean deviation by the particular average used in computing mean deviation.

That is coefficient of mean deviation, Co. MD =

Thus if mean deviation has been computed from AM, then the Co. MD = .

**Variance or Standard Deviation and Coefficient of variance:**

The standard deviation may be defined as the root of the mean of squares of the deviation of individual items from the AM.

**Population Variance:**

The formula for computing variance of a set of sample observations is given below :

**Case 1:**

If  are N values of a population of size N, then the population variance commonly designated as, is defined as

, Where of the distribution

**Problem:**

Let a ***population*** of 10 students got the marks in the examination as given in the table below. Find the variance of the given data.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13 | 15 | 14 | 16 | 2 | 8 | 9 | 23 | 28 | 12 |

**Answer:**

For the required solution please complete the following steps and table:

***Step 1:*** First find the AM of the given value. Population AM, =? ?

***Step 2:*** Then complete the following table:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 13 |  |  |
| 15 |  |  |
| 14 |  |  |
| 16 |  |  |
| 2 |  |  |
| 8 |  |  |
| 9 |  |  |
| 23 |  |  |
| 28 |  |  |
| 12 |  |  |
|  |  |  |

***Step 3:*** Here N= total number of observations= 10.

***Step 4:*** compute =??

**Case 2:**

For grouped data if the values  occur with frequencies  respectively then the variance of the distribution will be



**Problem:**

Let a ***population*** of 40 students got the marks in the examination as given in the table below. Find the variance of the given data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 15 | 20 | 25 | 30 | 35 |
|  | 6 | 8 | 15 | 7 | 4 |

**Answer:**

For the required solution please complete the following steps and table:

***Step 1:*** First find the AM of the given value. Population AM, =? ?

***Step 2:*** Then complete the attached table:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 15 | 6 |  |  |
| 20 | 8 |  |  |
| … | … | … | … |
| 35 | 4 |  |  |
|  |  |  |  |

***Step 3:*** Here N = total number of observations= 40.

***Step 4:*** compute =?

**Sample Variance:**

The formula for computing variance of a set of sample observations is given below:

**Case 1:**

If  are n values of a sample of size n, then the sample variance commonly designated as , is defined as

, Where Sample mean of the distribution

**Problem:**

Let a ***sample*** of 10 students got the marks in the examination as given in the table below. Find the variance of the given data.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13 | 15 | 14 | 16 | 2 | 8 | 9 | 23 | 28 | 12 |

**Answer:**

For the required solution please complete the following steps and table:

***Step 1:*** First find the AM of the given value. Sample Mean AM, ? ?

***Step 2:*** Then complete the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 13 |  |  | 8 |  |  |
| 15 |  |  | 9 |  |  |
| 14 |  |  | 23 |  |  |
| 16 |  |  | 28 |  |  |
| 2 |  |  | 12 |  |  |

***Step 3:*** Here N= total number of observations= 10.

***Step 4:*** compute =???

**Case 2:**

For grouped data if the values  occur with frequencies  respectively then the variance of the distribution will be



**Problem:**

Let a ***sample*** of 40 students got the marks in the examination as given in the table below. Find the variance of the given data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 15 | 20 | 25 | 30 | 35 |
|  | 6 | 8 | 15 | 7 | 4 |

**Answer:** For the required solution please complete the following steps and table:

***Step 1:*** First find the AM of the given value. Population AM, =? ?

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 15 | 6 |  |  |
| 20 | 8 |  |  |
| 25 | 15 |  |  |
| 30 | 7 |  |  |
| 35 | 4 |  |  |
|  | 40 |  |  |

***Step 2:*** Then complete the attached table:

***Step 3:*** Here N = total number of observations= 40.

***Step 4:*** compute =?

**Standard deviation:**

The standard deviation of a given data is obtained by taking the square root of the corresponding variance value.

That is standard deviation of the variable X, 

The coefficient of dispersion corresponding to variance is known as coefficient of variation  and is obtained by dividing standard deviation by the AM.

That is coefficient of variation =.

### NOTE: For computational convenience we will use the following formulae

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ungroup data | Group data |  |
| Population |  |  | Population |
| Sample |  |  | Sample |
|  | Ungroup data | Group data |  |

**Calculation of combined Standard deviation:**

The combined standard deviation of two sets of data containing and observaitons with means and and standard deviations and respectively is given by

Where,

combined standard deviation

And

This formula combined standard deviation of two sets of data can be extended to compute the standard deviation of more than two sets of data on the same lines

**Question:**

From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

|  |  |  |
| --- | --- | --- |
|  | Organization X | Organization Y |
| Number of wage-earners | 550 | 650 |
| Average monthly wages | 5000 | 4500 |
| Variance of the distribution of wages | 900 | 1600 |

1. Which organization pays a larger amount as monthly wages?
2. Determine the combined variance of all the employees taken together?

**Question:**

For a group of 50 male workers, the mean and standard deviation of their monthly wages are tk. 6300 and tk. 600 respectively. For a group of 40 female workers, these are tk. 5400 and tk. 600 respectively. Find the standard deviation of monthly wages for the combined group of workers.

***Answer: tk. 900***

***J K Sharma, 151***

**Emperical Rule of Standard Deviation**

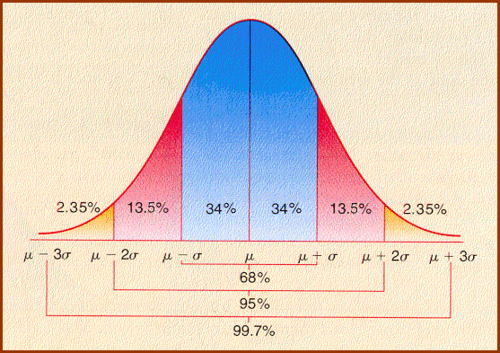
For symmetrical, bell shaped frequency distribution (also called normal Curve), the range with in which a given percentage of values of the distribution are likely to fall within a specified number of standard deviation of the mean is determined as follows:

covers approximately 68.27% of values in the data set

covers approximately 95.45% of values in the data set

covers approximately 99.73% of values in the data set

These ranges are illustrtated in the following figure



**Problem:**

The following data give the number of passengers travelling by airplane from one city to another in one week.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 115 | 122 | 129 | 113 | 119 | 124 | 132 | 120 | 110 | 116 |

Calculate themean and standard deviation and determine the percentage of class that lie between , and . What percentage pf cases lie outside these limits?

**Solution:**

The calculation for mean and standard deviation are given in the following table

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 115 |  |  |
| 122 |  |  |
| 129 |  |  |
| … |  |  |
| … |  |  |
| 110 |  |  |
| 116 |  |  |

and

Therefore,

The percentage of cases that lie between a given limit are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Values within Interval | Percentage of population | Percentage falling Outside |
| and 126.6 | 113, 115, 116, 119, 120, 122, 124 | 70% | 30% |
| =  = 106.80 and 133.20 | 110, 113, 115, 116, 119, 120, 122, 124, 129, 132 | 100% | Nil |

Exercise on Measure of Dispersion:

1. An Advertising company is looking for a group of extras to shoot a sequence for a movie. The ages of the first 20 candidates to be interviewed are

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 50 | 56 | 44 | 49 | 52 | 57 | 56 | 57 | 56 | 59 |
| 54 | 55 | 61 | 60 | 51 | 59 | 62 | 52 | 54 | 49 |

The director of the movie wants men whose ages are tightly grouped around 55 years. Being a statistics buff of sorts, the director suggests that a standard deviation of 3 years would be acceptable. Does this group of extras qualify?

1. The normal daily high temperatures (in degrees Fahrenheit) in January for 10 selected cities are as follows.

50, 37, 29, 54, 30, 61, 47, 38, 34, 61

The normal monthly precipitation (in inches) for these same 10 cities is listed below:

4.8, 2.6, 1.5, 1.8, 1.8, 3.3, 5.1, 1.1, 1.8, 2.5

Which variable represents greater relative variability?

1. A collar manufacturer is considering the production of new collars to attract young men. Thus following statistics of neck circumference are available based on measurements of a typical group of students of a particular university:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mid values (in inches): | 13.0 | 13.5 | 14.0 | 14.5 | 15.0 | 15.5 | 16.0 | 16.5 | 17.0 |
| Number of students: | 2 | 16 | 36 | 60 | 76 | 37 | 18 | 3 | 2 |

Compute the standard deviation and use the criterion , where is the standard deviation and is the arithmatic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collars are worn average half an inch longer than the neck size.

***Answer: 12.2 and 16.4 inches***

***J K Sharma, 155***

1. *ANIK* Electronics is considering employing one of two training programs. Two groups were trained for the same task. Group 1 was trained by program A, group 2 by program B. for the first group, the times required to train the employees had an average of 32.11 hours and a variance of 68.09. In the second group, the average was 19.75 and the variance was 71.14. Which training program has less relative variability in its performance?
2. The administrator of a Georgia hospital surveyed the number of days 200 randomly chosen patients stayed in the hospital following an operation. The data are:

|  |  |
| --- | --- |
| Hospital Stay in days | Number of patients |
| 1 – 3 | 18 |
| 4 – 6 | 90 |
| 7 – 9 | 44 |
| 10 – 12 | 21 |
| 13 – 15 | 9 |
| 16 – 18 | 9 |
| 19 – 21 | 4 |
| 22 – 24 | 5 |

Calculate the following:

* 1. Coefficient of variation (CV).
  2. Comments on the Skew ness of the distribution using the Pearson’s methods.
  3. Calculate the mean median and Mode. And hence comment.
  4. Calculate the mean deviation about mean, about median, and about mode. And also determine the corresponding relative measure of dispersion.

1. The manager of ***Nando’s Chicken*** has just received two dozen tomatoes form her supplier, but she is not ready to accept them. She knows from the invoice that thew average weight is 7.5 ounces, but she insists that all be of uniform weight. She will accept them only if the average weight is 7.5 ounces and the standard deviation is less than 0.5 ounce. Here are the weights of the tomatoes.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 6.3 | 7.2 | 7.3 | 8.1 | 7.8 | 6.8 | 7.5 | 7.8 |
| 7.2 | 7.5 | 8.1 | 8.2 | 8.0 | 7.4 | 7.6 | 7.7 |
| 7.6 | 7.4 | 7.5 | 8.4 | 7.4 | 7.6 | 6.2 | 7.4 |

What would be the manager’s decision and why?

1. Student’s ages in the regular daytime MBA program and the evening program of BRAC University are described by these two samples:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Regular MBA | 23 | 29 | 27 | 22 | 24 | 21 | 25 | 27 | 24 | 26 |
| Evening MBA | 27 | 34 | 30 | 29 | 28 | 30 | 34 | 35 | 28 | 29 |

If homogeneity of the class is a positive factor in learning, use a measure of relative variability to suggest which of the two groups will be easier to teach?

1. In two factories A and B engaged in the same industry, the average monthly wages and standard deviations are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Factory | Average monthly Wages (Tk.) | S.D. of Wages (Tk. ) | No. of Wage Earners |
| A | 4600 | 500 | 100 |
| B | 4900 | 400 | 80 |

Determine

1. Which factory A or B pays larger amount as monthly wages?
2. Which factory shows greater variability in the distribution of wages?
3. What is the mean wage of all workers in two factories taken together?

**Shape characteristics of a distribution:**

The study of shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions. By shape characteristic of a distribution we refer to the extent of its asymmetry and peakedness relative to an agreed upon standard and the study of these two characteristics (that is asymmetry and peakedness) is accomplished through what is known as the measures of skewness and kurtosis.

We study these two characteristics in the following section:

**Skewness:**

The term skewness means the lack of symmetry. The skewness may be either positive or negative. When the skewness is positive the associated distribution is called positively skewed. When the skewness is negative the associated distribution is negatively skewed.

Now some very simple measures of skewness is shown here:

|  |  |
| --- | --- |
| **Method 1** | If for a distribution  The distribution is positively skewed  The distribution is negatively skewed  The distribution is symmetric |
| **Method 2** | Pearson’s coefficient of skew ness  Then if  The distribution is positively skewed  The distribution is negatively skewed  The distribution is symmetric. |

**Kurtosis:**

There is considerable variation among symmetrical distributions. For instance, they can differ markedly in terms of peaked ness. This is what we call kurtosis. Kurtosis, as defined by Spiegel (Spiegel: Theory and Problems of Statistics) is the degree of peaked ness of a distribution, usually taken in relation to a normal distribution.

* A curve having relatively **higher peak** than the normal curve, is known as **leptokurtic**.
* A curve, which is **neither too peaked nor too flat topped**, is known as **mesokurtic**.
* A curve that is **more flat topped** than the normal curve is called **platykurtic**.

**Question:**

If for a distribution Mean=18, Median=32 and Mode=36 the distribution is \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ skewed.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Positively | 1. Symmetrically | 1. None | 1. Negatively |

**Merits and Demerits of different Measures of Dispersion:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Merits | **Demerits** |  |
| Range | * Easy to understand and calculate. * It is based only on extreme observations and no detail in formations is required. * It gives us a quick idea of the variability of a set of data. | * It is not based on all observation. * Range does not give any indication of the character of the distribution with in the two extreme observations. * Range is subject of fluctuations from sample to sample. * Cannot be computed in case of open-end class. | Range |
| **Quartile deviation** | * It is superior to range as a measure of dispersion. * It is applicable in Open-end class. * Easy to understand and compute. * Not affected by extreme values. | * It ignores 50% of items that is the first 25% and last 25% of observations. * Very much affected by sampling fluctuations. * Not suited for further algebraic treatment. | **Quartile deviation** |
| **Mean Deviation** | * It is easy to calculate. * It considers all observations. * Less affected by extreme values. | * Not amenable for further algebraic treatment. * The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of items. * In valid for open – end class. | **Mean Deviation** |
| **Variance** | * Rigidly defined. * Based upon all observation. * Easy to understand * Less affected by sampling fluctuations. * Suitable for further algebraic treatment. | * Difficult to calculate. * Affected by extreme values. * Difficult to calculate for open-end class. | **Variance** |
|  | Merits | **Demerits** |  |

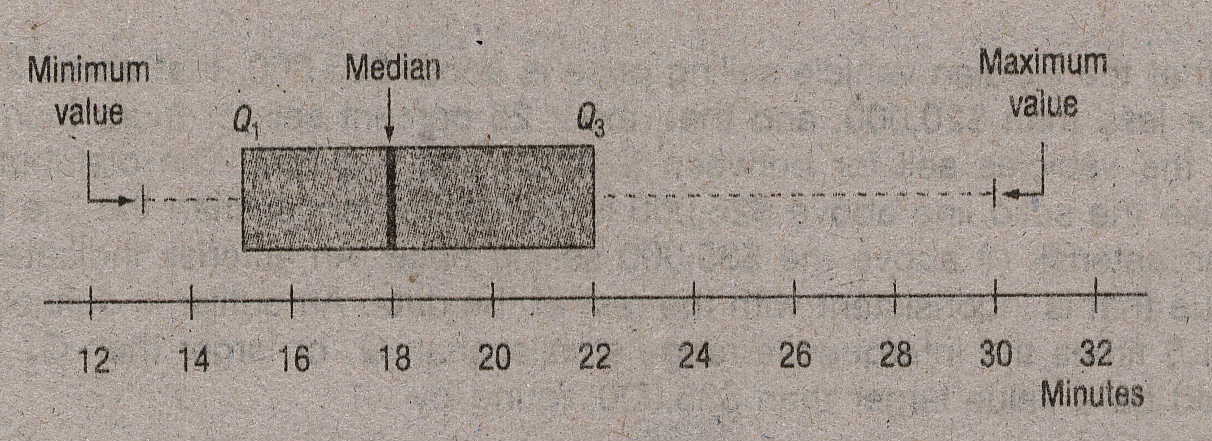
**Box Plot:**

A box plot is a graphic display that shows the general shape of a variable’s distribution. It is based on five descriptive statistics: the minimum value, the first quartile, Median, third quartile and the maximum value.

**Example:**

Pizza Hut offers free delivery of its pizza within 15 miles. Mr. Rahman the owner wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

|  |  |
| --- | --- |
| Minimum value = | 13 minutes |
|  | 15 minutes |
| Median = | 18 minutes |
|  | 22 minutes |
| Maximum value = | 30 minutes |

Develop a boxes plot for the delivery times. What conclusions can you make about make about the delivery times?

**Solution:**

In order to draw box plot follow the steps mentioned below:

***Step 1:***Create an appropriate scale along the horizontal axis.

***Step 2:***Draw a box that starts at (15 minutes) and ends at (22 minutes)

***Step 3:***Place a vertical line to represent the median (18 minutes)

***Step 4:***Extend the horizontal lines[[2]](#footnote-2) from the box out to the minimum value (13 minutes) and the maximum value (30 minutes)

**Interpretation of the Box Plot:**

* The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the inter quartile range[[3]](#footnote-3). That shows the spread or dispersion of the majority of deliveries.
* The box plot also reveals that the distribution of the delivery times is positively skewed. The guiding principle for such conclusion are
* The dashed line to the right of the box from 22 minutes  to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes to the minimum value of 13 minutes.
* The median is not in the middle in the center of the box. The distance from the first quartile to the median is smaller than the distances from the median to the third quartile.

**Question:**

Construct a box plot for the data given below and hence comment on the skewness of the distribution:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 99 | 75 | 84 | 33 | 45 | 66 | 97 | 69 | 55 | 61 |
| 72 | 91 | 74 | 93 | 54 | 76 | 62 | 91 | 77 | 68 |

**Miscellaneous Exercise**

**Question 1:**

Average mark obtained by 15 students was 10 and the average mark obtained by 10 students was 15. What was the average mark obtained by all students?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 10 | 8 | 12 | 15 | 11 |

Answer 11: c

**Question 2:**

Study the following histogram and hence determine the modal class and what proportion of students get marks below 80.

|  |  |
| --- | --- |
| 1. Modal class 90+ and 80.76% | 1. Modal class 50 – 60 and 76.9% |
| 1. Modal class 60 - 70 and 76.9% | 1. Modal class below 50 and 86.9% |

**Question 3:**

A school had 100 students aged 20 years on an average. At the end of the year, 20 students aged 22 years on an average left and 25 students of 18 years on an average joined the school. What is the average age of the present students of the school?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 20.14 | 19.14 | 22.14 | 22 | None |

Answer: b

**Question 4:**

A group of students has hired a bus for Tk. 3000 for going to a picnic. They had an understanding that each participant would share the charge in equal amounts. But because of 10 students not turning up, the charge per student increased by taka 10 over the initial estimates. What is the number of students who originally registered for the picnic?

Answer: 60

**Question 5:**

Salman bought 500 shares of company “X” at tk. 600 and 2 months later bought another 250 shares of the same company at tk. 560. At what price should he purchase additional 250 shares in order to have an average price of tk. 580 per share?

**Answer: Tk. 560**

1. R. A. Fisher (1890- 1962) is known as the father of STATISTICS. [↑](#footnote-ref-1)
2. These horizontal lines outside of the box are sometimes called “whiskers” because the looks a bit like a cat’s whiskers. [↑](#footnote-ref-2)
3. The inter quartile range is the distance between the first and the third quartile. [↑](#footnote-ref-3)