MAT 3103: Computational Statistics and Probability

Chapter 1: Data Visualization

Statistics: The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, where data are collected according to some pre-determined objective. Statistics is especially useful in drawing general conclusions about the population characteristics based on sample observations or population observations.

Terms related to Statistics:

Image source: https://www.sigmamagic.com/blogs/online-sample-size-calculators/

Population: Population consists of all individuals, items or units under investigation in a statistical study. The size of the population is denoted by **N**.

Sample: Sample is a representative part of the population units from which information are to be collected. The size of the sample is denoted by $\boldsymbol{n}$ ($\leq$ N)

Example 1.1: We are interested to study the average number of signals sent from a station in different days of a year. There are two ways to measure the average: one way to collect the information from the station for every day of the year. This process of collection of data is known as **census** and using the **census data** we can calculate the average signal sent per day. Alternatively, instead of recording the information for every day we can record the information for some randomly selected days. This process of collection of data is known as **sample survey** and using the **sample data** we can calculate the average signal sent per day.

Variable: The characteristic which varies from one unit to another is called a variable.

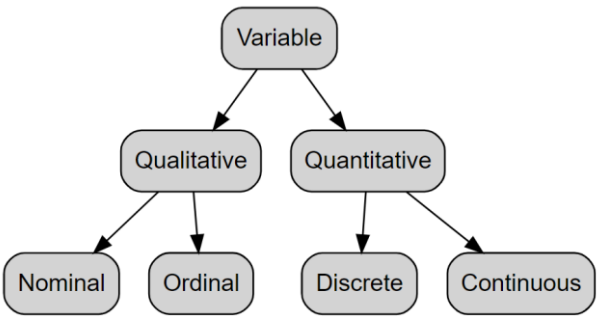
Types of variables:

The reason why we often class variables into different types is because not all statistical analyses can be performed on all variable types.

- **Qualitative Variable:** The variable which cannot be measured by numerical figure is called a qualitative variable or categorical variable. e.g., **gender, religion, name, letter grade, blood group etc.** Qualitative variables are divided into two types:
    - **Nominal Variable:** A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels. e.g., **eye color, gender, wisdom etc**.
    - **Ordinal Variable:** a **qualitative ordinal** variable is a qualitative variable with an order implied in the levels. e.g., **health condition, grade etc.**

- **Quantitative Variable:** The variable which is measured by numerical value is called a quantitative variable. e.g., **age, weight, time, price, speed etc**. Quantitative Variables are further classified as:
    - **Discrete Variable:** A quantitative variable which takes only integer values is called a discrete variable. They take only integer values. e.g., **number of computers in laboratories, number of students in each section of Statistics etc**.
    - **Continuous Variable:** The variable which takes integer as well as fractional values is called a continuous variable. e.g., **height, temperature, income, marks etc.**

We can summarize types of variables in the following diagram:



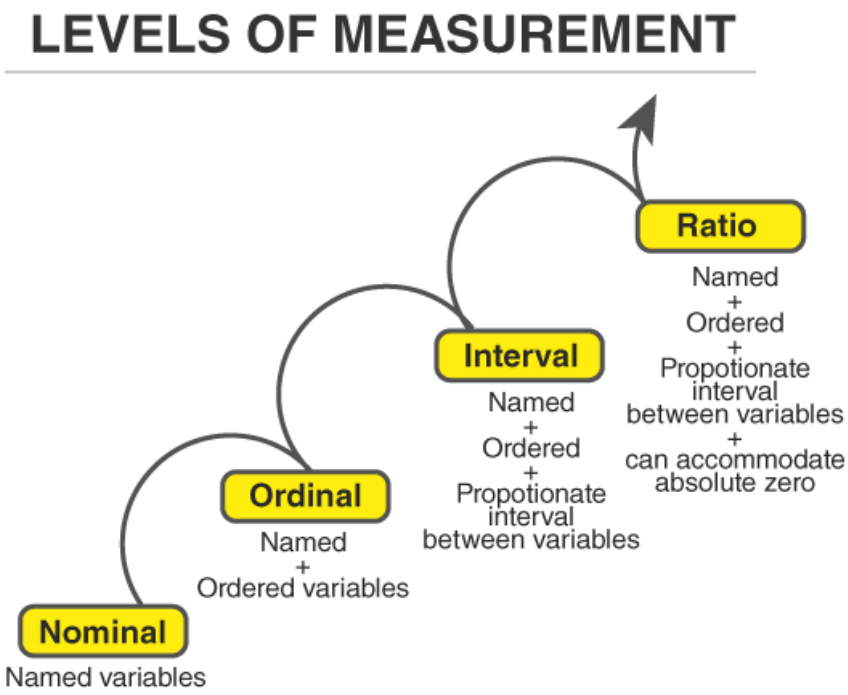
Example 1.2:

Variable	Quantitative	Qualitative	Discrete	Continuous
Number of tresses in a garden	√		√	
A person's marital status		√		
Distance of University from home	√			√
Color of flowers		√		
Number of errors on a math test	√		√	

Scale of Measurement

There are four different scales of measurement. The data can be defined as being one of the four scales. The four types of scales are:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale



1.Nominal Scale

The scale of measurement by which we can classify and identify a qualitative variable according to different categories is called nominal scale.

Example:

- Gender of a garment worker
- Religion of a person
- Marital Status of a worker

2. Ordinal Scale

The scale of measurement by which we can classify and identify and rank a qualitative variable according to different categories is called ordinal scale.

Example

- Economic status of a citizen (Rich, middle class, poor)
- Food Quality (Excellent, good, bad, very bad).

3. Interval Scale

The scale of measurement by which we can measure a quantitative variable on experimental unit with arbitrary zero as origin is called interval scale.

Example

- Temperature

4. Ratio Scale

The scale of measurement by which we can measure a quantitative variable on experimental unit with absolute zero as origin is called ratio scale.

Example

- Number of Children per family
- Number of defects of a product

Example 1.3: A medical researcher wants to estimate the survival time in years of a patient after the beginning of a particular type of cancer and after a particular regime of radio therapy. A sample of 50 patients having cancer and radio therapy who are not alive have been selected randomly from a cancer hospital.

a. What is the population?
b. What is the sample?
c. What is the variable to be measured?
d. Is the variable qualitative, or discrete or continuous?

Solution:

a. The population is the set of all patients listed in the registrar of cancer hospital having that particular type of cancer who died after undergoing the particular type of radiotherapy.

b. The 50 patients selected at random from the cancer hospital is the sample.

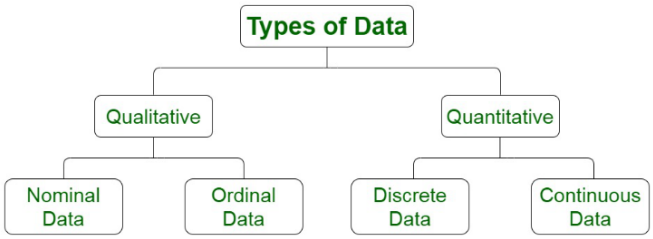
c. Survival times in years is the variable to be measured.

d. The variable is quantitative and continuous variable.

Data: The information collected from population or sample units are known as data.

Types of data:

In statistics, there are four main types of data: Nominal, Ordinal, Discrete, and Continuous. These types of data are used to describe the nature of the data being collected or analyzed, and they help determine the appropriate statistical tests to use.



There are also another 2 types of data:

- **Primary Data:** The data which are collected by investigating population units or sample units are known as primary data. e.g., census data.
- **Secondary Data:** The data which are collected from official records or from published works are known as secondary data. e.g., census report.

Example 1.4: Information of the students recorded by AIUB IT department is primary data for AIUB. If someone uses the data for specific research purposes (with the permission from AIUB), then it will be secondary data for that person.

Sources of data:

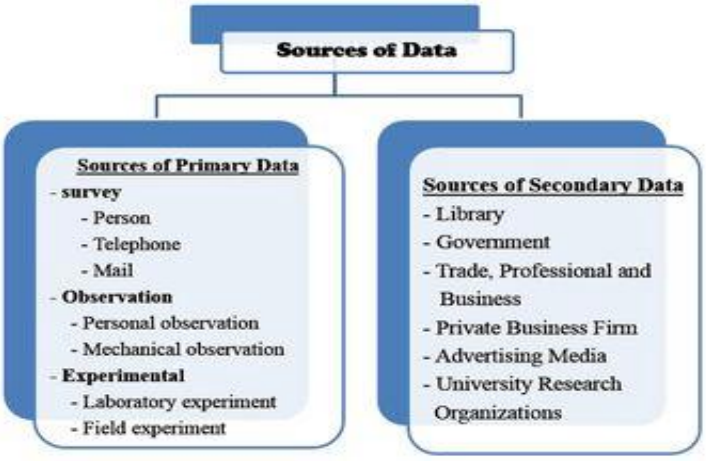


Image source: https://computing2k16.fandom.com/wiki/Primary_vs_Secondary_Sources

Difference Between Primary Data and Secondary Data:

Basis For Difference	Primary Data	Secondary Data
Introduction	Fresh data gathered by conducting original research and investigation	Data already gathered by others
Originality	Yes	No
Information Type	Qualitative	Quantitative
Coverage/Scope	Limited	Wider
Validity/Reliability	More	Less
Required Time	More	Less
Cost And Effort	More	Less
Sources Of Data	Investigation, survey, research, interviews etc.	Blogs, magazines, articles, journals etc.

Ethics in Research:

- Ethics are the set of rules that govern our expectations of our own and others' behavior.
- Research ethics are the set of ethical guidelines that guides us on how scientific research should be conducted and disseminated.
- Research ethics govern the standards of conduct for scientific researchers. It is the guideline for responsibly conducting the research.

- Research that implicates human subjects or contributors rears distinctive and multifaceted ethical, legitimate, communal and administrative concerns.
- Research ethics is unambiguously concerned in the examination of ethical issues that are upraised when individuals are involved as participants in the study.
- Research ethics committee/Institutional Review Board (IRB) reviews whether the research is ethical enough or not to protect the rights, dignity and welfare of the respondents.

The following process helps to ensure ethics at different steps of research:

- Collect the facts and talk over intellectual belongings openly
- Outline the ethical matters
- Detect the affected parties (stakeholders)
- Ascertain the forfeits
- Recognize the responsibilities (principles, rights, justice)
- Contemplate your personality and truthfulness
- Deliberate innovatively about possible actions
- Respect privacy and confidentiality
- Resolve on the appropriate ethical action and be willing to deal with divergent point of view.

Advantages:

- Research ethics promote the aims of research.
- It increases trust among the researcher and the respondent.
- It is important to adhere to ethical principles in order to protect the dignity, rights and welfare of research participants.
- Researchers can be held accountable and answerable for their actions.
- Ethics promote social and moral values.
- Promotes the ambitions of research, such as understanding, veracity, and dodging of error.
- Ethical standards uphold the values that are vital to cooperative work, such as belief, answerability, mutual respect, and impartiality.
- Ethical norms in research also aid to construct public upkeep for research. People are more likely to trust a research project if they can trust the worth and reliability of research.

Objectives:

- The first and comprehensive objective – to guard/protect human participants, their dignity, rights and welfare.
- The second objective – to make sure that research is directed in a manner that assists welfares of persons, groups and/or civilization as a whole.
- The third objective – to inspect particular research events and schemes for their ethical reliability, considering issues such as the controlling risk, protection of privacy and the progression of informed consent.

Array: It is an arrangement of observations either in ascending or descending order.

Example 1.5: Let us consider the observations (x): 12, 19, 16, 10, and 20.

In ascending order: 10, 12, 16, 19, 20.

In descending order: 20, 19, 16, 12, 10.

Data Representation: By suitably organizing data, we can often make a large and complicated set of data more compact and easier to understand. Statistical data can be presented by

1. Tabulation method
2. Graphs and diagrams method

Tabulation method:

Frequency Distribution: A tabular arrangement of data by classes together with the corresponding number of items in each class is called a frequency distribution or frequency table. It is used to represent the value of different levels of quantitative variables.

Example 1.6

The heights of 50 students, measured to the nearest centimeters, have been found to be as follows:

161, 150, 154, 165, 168, 161, 154, 162, 150, 151, 162, 164, 171, 165, 158, 154, 156, 172, 160, 170, 153, 159, 161, 170, 162, 165, 166, 168, 165, 164, 154, 152, 153, 156, 158, 162, 160, 161, 173, 166, 161, 159, 162, 167, 168, 159, 158, 153, 154, 159

Construct a frequency distribution table for the above data.

The resulting table for grouping the height of 50 students:

Class interval of weight	**Tally**	**Frequency**
150-155	~~𝍸 𝍸~~ II	12
155-160	~~𝍸~~ IIII	9
160-165	~~𝍸 𝍸~~ IIII	14
165-170	~~𝍸 𝍸~~	10
170-175	~~𝍸~~	5
Total		50

Terms Associated with Frequency Distributions:

- Class size or width - the differences between lower- and upper-class limits.
- Cumulative frequencies are the cumulative totals of successive frequencies of a frequency distribution.
- Class mark or midpoint - the average of class limits.

$$\textbf{Mid-Point} = \frac{Upper\ Limit + Lower\ limit}{2}$$

- Relative frequency-the relative frequency of a class is obtained by dividing the frequency of that class by the sum of all frequency. Thus, the relative frequency shows a fractional part of the total frequency belongs to the corresponding class.

$$\textbf{Relative Frequency} = \frac{Frequncy\ of\ the\ class}{Sum\ of\ all\ frequncies}$$

- Percentage-The percentage for a class is obtained by multiplying the relative frequency of that class by 100.

$$\textbf{Percentage} = \textbf{(Relative frequency)} \times \textbf{100}$$

Example 1.7:

Class interval of weight	**Frequency**	**Relative Frequency**	**Percentage (RFx100)%**	**Cumulative Frequency**
150-155	12	0.24	24	12
155-160	9	0.18	18	21
160-165	14	0.28	28	35
165-170	10	0.20	20	45
170-175	5	0.10	10	50
Total	50	1.00	100	

- Find the number of students whose length is less than 160cm.
  Ans: There are 12+9= 21 students whose length are less than 160cm.
- Find the percentage of students whose length is above and 165cm.
  Ans: From the above table, we can say that there are $(20 + 10)\% = 30\%$ students whose lengths are above and 165cm.

Graphs and diagrams method: Different graphs and diagrams are-

i) Bar diagram ii) Pie diagram iii) Histogram
iv) Frequency curve v) Scatter diagram

Diagrammatic representation of data:

Bar diagram and pie diagram are generally used to represent the value of qualitative variable diagrammatically.

Bar diagram: Bar diagrams are simple diagrams that are made up of a number of rectangular bars of equal widths whose heights are proportional to the quantities or frequencies they represent.

Pie diagram: Pie diagrams can be defined as a circle drawn to represent the totality of a given data. The circle is also divided into sectors with each sector proportional to the components of the variable it represents. Pie diagram is very useful in drawing comparison among the various components or between a part and the whole. Both diagrams are used to represent value of different levels of qualitative variable.

Example 1.8

A sample of 30 employees from large companies was selected, and these employees were asked how stressful their jobs were. The responses of these employees are recorded below, where very represents very stressful, somewhat mean somewhat stressful and none means no stressful at all.

Draw Bar diagram and Pie Chart of the following table:

Stress on jobs	Number of employees	Angels = $\frac{x \times 360°}{Total}$
Very	10	119.88
Somewhat	14	168.12
None	6	72.00
Total	30	360

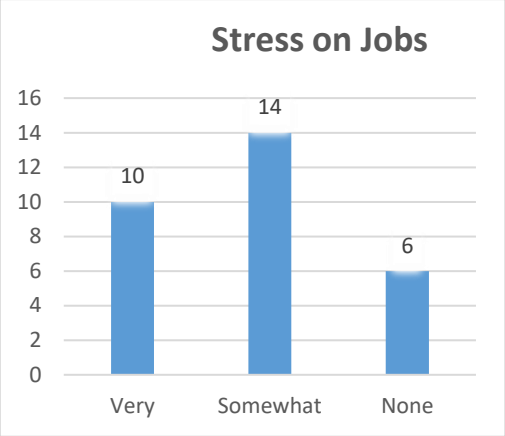


Figure: Bar Diagram

Figure: Pie diagram

Histogram: Consists of set of rectangles having: (a) bases on a horizontal axis with centers at the class marks and length equal to the class interval sizes, and (b) areas proportional to the class frequencies. It is the graphical representation of continuous classes of frequency distribution.

Example 1.9: Uncle Bruno owns a garden with 30 black cherry trees. Each tree is of a different height. We can group the height of the trees as follows in a frequency distribution table:

Class Interval	**Frequency**
60-65	3
65-70	3
70-75	8
75-80	10
80-85	5
85-90	1
Total	30

Draw a histogram using the given data set.

Solution: Histogram of the changing the size of the bin:

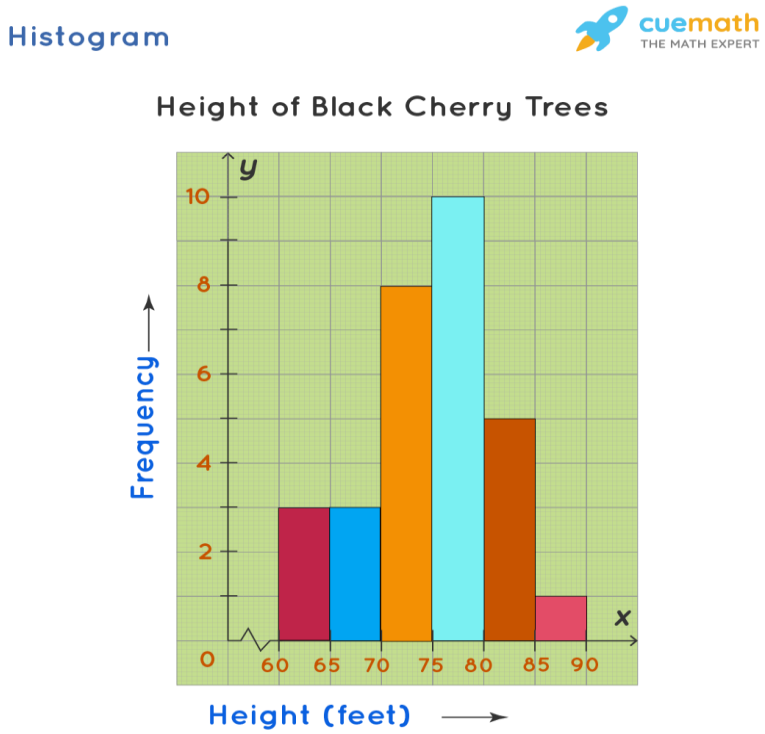


Image source: https://www.cuemath.com/data/histograms

Difference between Bar diagram and histogram:

- Bar diagram is used to represent the qualitative variable and histogram is used to represent quantitative variable.
- Bar diagram is one dimensional and histogram is two dimensional.

Frequency curve:

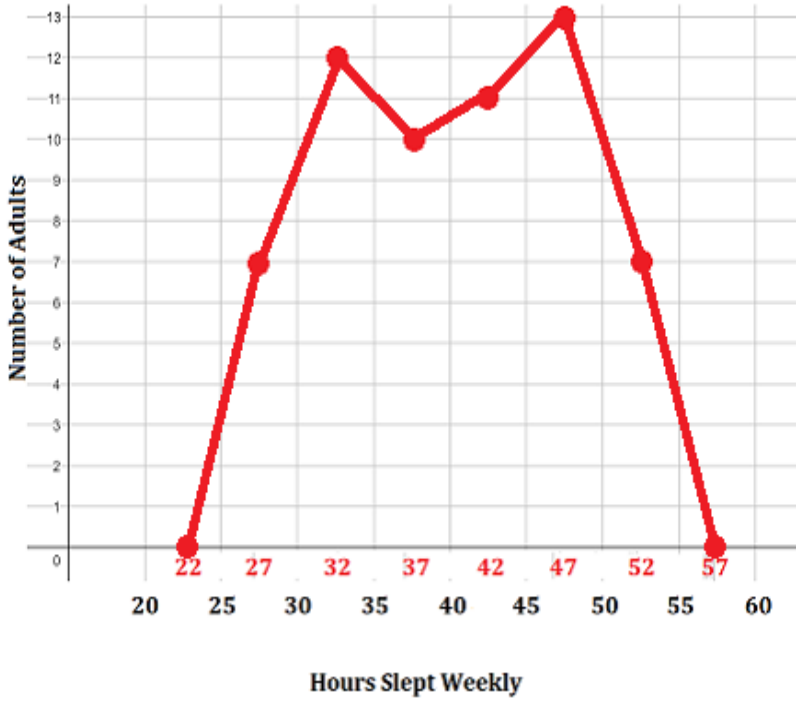
It is a smooth graph of the class frequency plotted against the mid value. It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram by free hand/ smooth hand.

Example 1.10

Adults were surveyed on the number of hours of sleep they get on average in a week. The results of the survey are in the table below. Construct a frequency polygon for this information.

Hours of sleep	No. of Hours	Mid value
25-30	7	27.5
30-35	12	32.5
35-40	10	37.5
40-45	11	42.5
45-50	13	47.5
50-55	7	52.5
Total	60	

Solution: Frequency curve of hours of sleeping of selected adults:



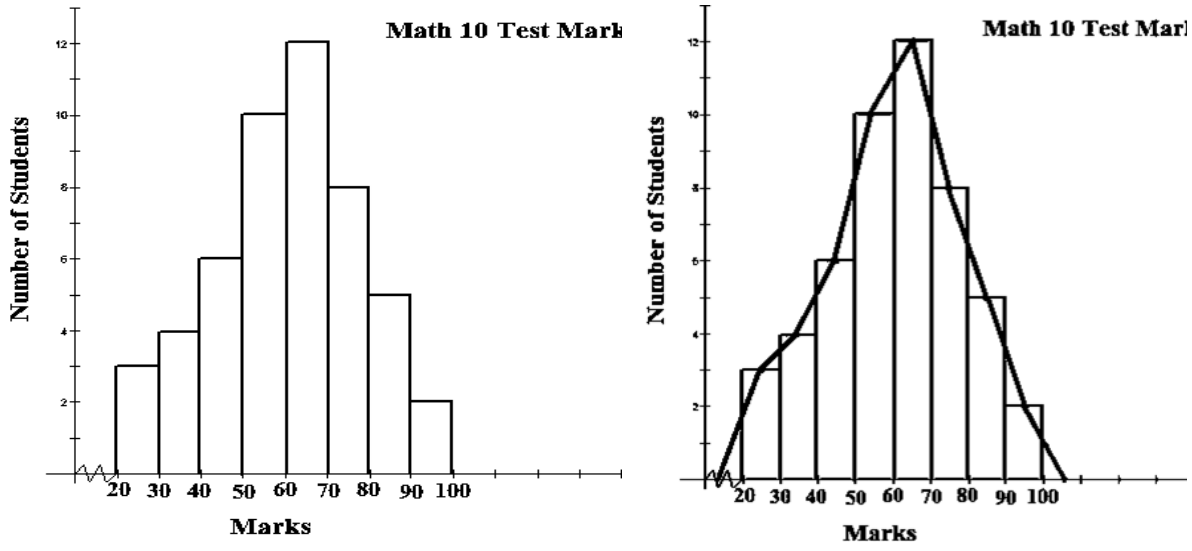
Example 1.11

The following histogram represents the marks made by 40 students on a math test.

Marks	No. of students	Mid value
20-30	3	25
30-40	4	35
40-50	6	45
50-60	10	55
60-70	12	65
70-80	8	75
80-90	5	85
90-100	2	95
Total	40	

1. Represent the data by histogram.
2. Draw a frequency curve of the math score data.

Solution*:* Draw histogram and then draw a frequency curve to represent the data.



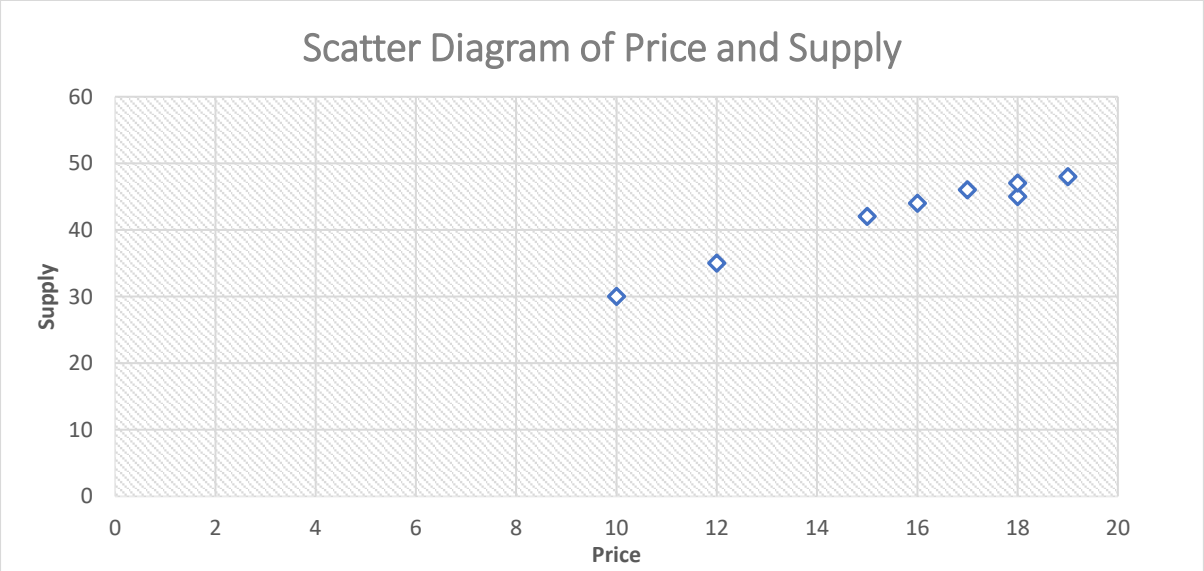
Scatter diagram: A scatter (XY) Plot has points that show the relationship between two sets of data. To construct a **Scatter plot**, Label the x- and y- axis. Choose a range that includes the maximums and minimums from the given data.

Example 1.12

The following data relate to the prices and supplies of a commodity during a period of eight years. Draw a scatter plot.

Price vs Supply	
Price (Taka/kg) (X)	**Supply(kg) (Y)**
10	30
12	35
18	45
16	44
15	42
19	48
18	47
17	46

Solution: Here is the same data as a Scatter Plot. In this example, each dot shows price versus supply.



MATLAB code

Diagrams:

bar produces vertical bar chart, and can be used as bar(y) or bar(x,y) – the first form uses 1:length(y) as

the values for x, which are the bar locations.

```
>> x = [1 2 4 5 9];
>> y = 20-(5-x).^2;
>> bar(x,y)
```

>> title(’Bar Chart’)

Other bar charts can be produced using an optional style argument (bar(x,y,’style’)), where style is one of:

• ’grouped’ - Produces a bar chart where values in each column of y are grouped together, but appear in different colors.

• ’hist’ - produces a bar chart with no space between bars.

>> bar(1:3,[1 2 3;2 3 4;3 4 5],’grouped’)

>> bar(1:3,[1 2 3;2 3 4;3 4 5],’hist’)

pie can be used to produce a pie chart

>> pie([.7 .2 .1],[.1 0 0],{’Stocks’,’Bonds’,’Cash’})

>> title(’Asset allocation’)

Exercise 1

1.1 Identify each of the following underlined variables as qualitative or quantitative.

a.	The jersey number of a player.	
b.	The apartment numbers/ house numbers of a building/street.	
c.	Educational qualification of a student.	
d.	Number of eye blinks a minute.	
e.	Outcome of a game.	

1.2 Identify which of the following variables are qualitative and which are quantitative.

a.	Performance rating of a student.	
b.	Blood group of a person.	
c.	Marital status of people	
d.	Blood pressure of a person.	
e.	Number of feet on a mammal	
f.	Monthly TV cable bills	
g.	Salary of an employee.	

1.3 a. Write down the differences between histogram and bar diagram.

b. Mention the name of the variables the values of which are presented by bar diagram and by pie diagram. Also mention some examples of the variable the value of which are presented by histogram and by frequency curve.

c. Mention important names of graphs and diagrams used to represent statistical data. Which of the graphs and diagrams are used for presenting qualitative data and which are used for quantitative data?

1.4 The following are the size of shoes purchased by person in an outlet:

Size of Shoes	:	5	6	7	8	9	10	11
Number of persons:		10	20	25	40	22	15	6

Represent the size of shoes purchased by different persons by bar and by pie diagram.

1.5 Given below is the table showing the approximate lengths, in mm, of 40 leaves taken from different parts of a certain species.

Weight (in kgs)	25-30	30-35	35-40	40-45	45-50	50-55	55-60	Total
Number of persons	1	4	8	10	8	7	2	40

Is the histogram appropriate to represent the given data? If yes, draw a histogram for the above data.

1.6 The daily profits (in Taka) of 100 shops are distributed as follows:

Daily profit	No. of Shops
0-50	12
50-100	18
100-150	27
150-200	20
200-250	17
250-300	6
Total	100

a) Represent the data by histogram and frequency curve.

b) Find the percentage of shops in which daily profit less than 150 Taka.

c) Find the percentage of shops in which daily profit is above and 200 Taka.

d) Find the number of shops in which daily profit is less than 100 Taka.

1.7 The following are the number of e-mails received in different days by different organizations:

Days (x)	:	5	8	3	10	15
No. of mails received (y)	:	54	65	42	107	89

Draw a scatter diagram of the data.

Sample MCQs

1. A graph that uses vertical bars to represent data is called a ____.

a. Line graph b. Bar graph c. Scatterplot d. Vertical graph

2. _____ are used when you want to visually examine the relationship between two quantitative variables.

a. Bar graphs b. Pie graphs c. Line graphs d. Scatterplots

3. A frequency distribution can be:

a. Qualitative b. Discrete c. Continuous d. Both (b) and (c)

4. The number of classes in a frequency distribution is obtained by dividing the range of variable by the:

a. Total frequency b. Class interval c. Mid-point d. Relative frequency

5. The largest and the smallest values of any given class of a frequency distribution are called:

a. Class Intervals b. Class marks c. Class boundaries d. Class limits

6. The lower- and upper-class limits are 20 and 30, respectively; the midpoints of the class are:

a. 20 b. 25 c. 30 d. 50