MAT 3103: Computational Statistics and Probability

Chapter 1: Data Representation

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 or items or units which are 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).

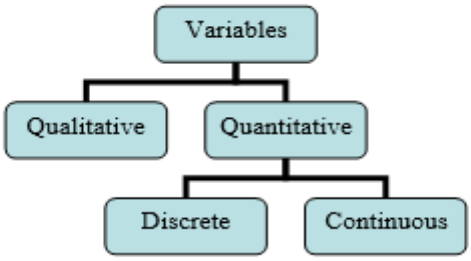
Example 1.1: We are interested to study the average number of signals sent from a station in different days of a year. There are two ways to measure average: one way to collect the information from the station for every day of the year. This process of collection of data is known as **census** and using 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:

- **Qualitative Variable:** The variable which cannot be measured by numerical figure is called a qualitative variable or categorical variable. E.g., gender, religion, color, name, letter grade, blood group etc.
- **Quantitative Variable:** The variable which is measured by numerical value is called a quantitative variable. E.g., age, weight, time, height, 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., age, weight, time, height, speed etc.

We can summarize types of variables in the following diagram:



Example 1.2:

Variable	Quantitative	Qualitative	Discrete	Continuous
Number of members in a family	√		√	
A person's marital status		√		
Length of a person's arm	√			√
Color of cars		√		
Number of errors on a math test	√		√	

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:

- **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 purpose (with the permission from AIUB), then it will be secondary data for that person.

Sources of data:

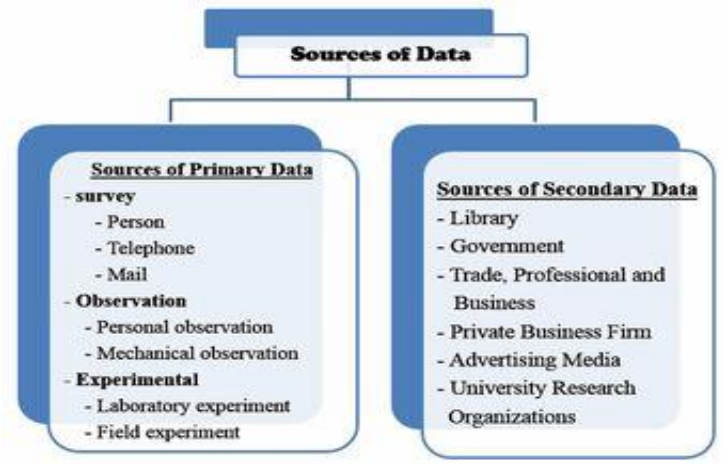


Image source: https://computing2k16.fandom.com/wiki/Primary_vs_Secondary_Sources

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 variable.

Example 1.6

In the following table the length of 40 laurel leaves are recorded to the nearest millimeter. Construct a frequency distribution.

136	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	118	154	165
146	173	142	147	135	153	140	135
161	145	135	142	156	156	145	128

The resulting table for grouping the length of 40 laurel leaves:

Class Interval of length	Tally	Frequency
118-128	///	3
128-138	/// //	7
138-148	/// /// ///	13
148-158	/// ////	9
158-168	///	5
168-178	///	3
Total		40

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.

Example 1.7:

Length	Frequency	Midpoint	Cumulative frequency
118-128	3	123	3
128-138	7	133	10
138-148	13	143	23
148-158	9	153	32
158-168	5	163	37
168-178	3	173	40
Total	40		

- Find the number of leaves which length is less than 158mm.
  Ans: There are 3+7+13+9 = 32 leaves whose length are less than 158mm.
- Find the percent of leaves which length is above and 148mm.
  Ans: There are $\frac{9+5+3}{40} \times 100 = 42.5\%$ leaves whose values are above and 148mm.

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.

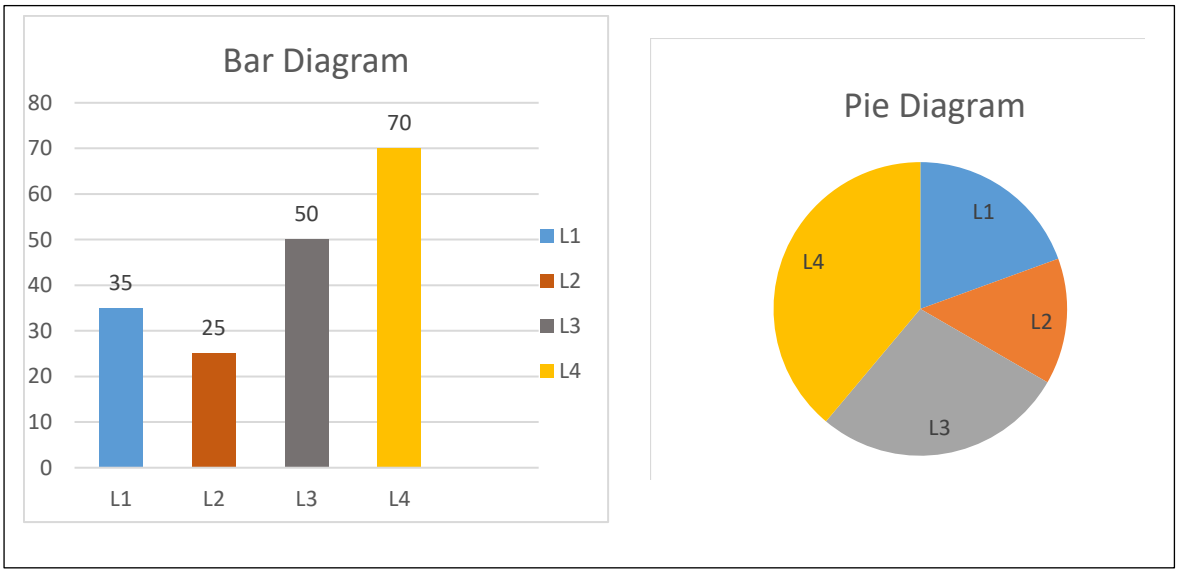
Pie diagram: Pie diagrams can be defined as a circle drawn to represent the totality of a given data. The circle is also divided into sectors with each sector proportional to the components of the variable it represents. Pie diagram is very useful in drawing comparison among the various components or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

Example 1.8

The following are the number of computers available in different laboratories.

Laboratory	Number of computers	Angels $= \frac{x \times 360^\circ}{Total}$
L1	35	70
L2	25	50
L3	50	100
L4	70	140
Total	180	360

1. Draw a bar diagram of the data.
2. Represent the data by a pie diagram.



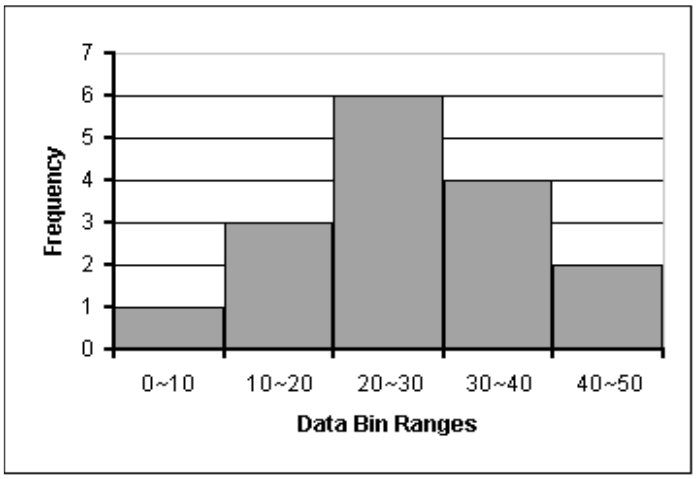
Histogram: Consists of set of rectangles having: (a) bases on a horizontal axis with centers at the class marks and length equal to the class interval sizes, and (b) areas proportional to the class frequencies. It is the graphical representation of continuous classes of frequency distribution.

Example 1.9: A frequency distribution the changing the size of the bin is given as.

Class Interval	**Frequency**
0-10	1
10-20	3
20-30	6
30-40	4
40-50	2
Total	16

Draw a histogram using the given data set.

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



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:

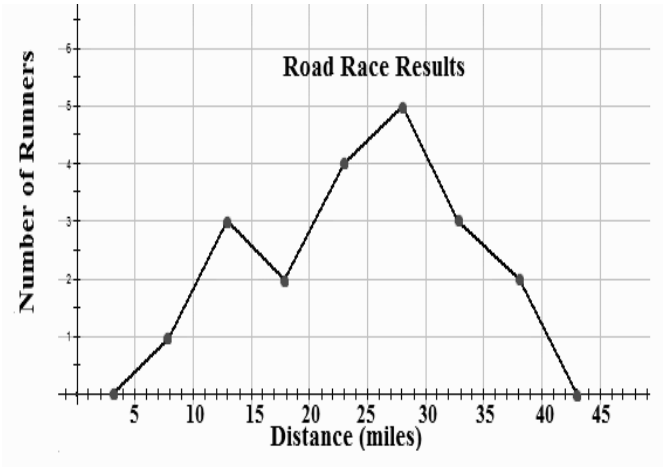
It is a smooth graph of the class frequency plotted against the mid value. It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram by free hand/ smooth hand.

Example 1.10

The following distribution table represents the number of miles run by 20 randomly selected runners during a recent road race. Represent the data by frequency curve.

Distance	Frequency	Mid value
6-11	1	8.5
11-16	3	13.5
16-21	2	18.5
21-26	4	23.5
26-31	5	28.5
31-36	3	33.5
36-41	2	38.5
Total	20	

Solution: Frequency curve of selected runners during a recent road race



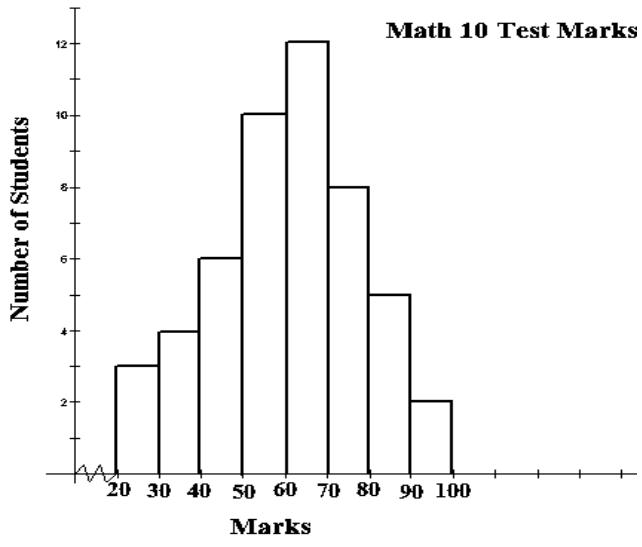
Example 1.11

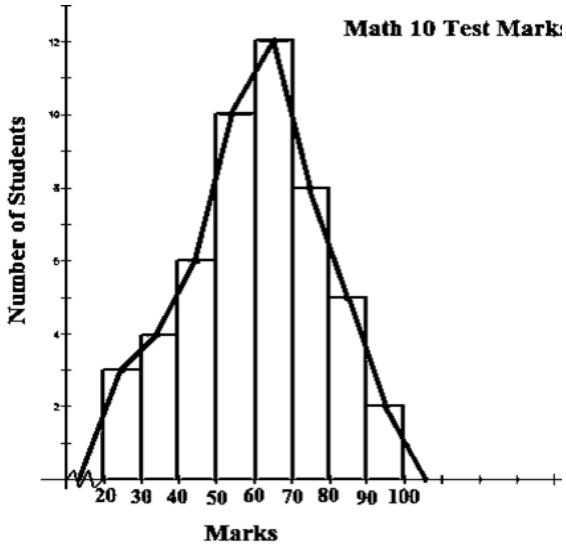
The following histogram represents the marks made by 40 students on a math 10 test.

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.





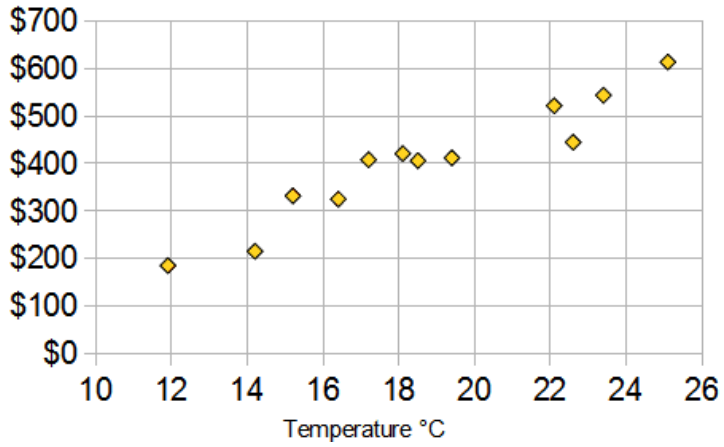
Scatter diagram: A scatter (XY) Plot has points that show the relationship between two sets of data. To construct a **Scatter plot**, Label the x- and y- axis. Choose a range that includes the maximums and minimums from the given data.

Example 1.12

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C (X)	**Ice Cream Sales (Y)**
14.2°	215
16.4°	325
11.9°	185
15.2°	332
18.5°	406
22.1°	522
19.4°	412
25.1°	614
23.4°	544
18.1°	421
22.6°	445
17.2°	408

Solution: Here is the same data as a Scatter Plot. In this example, each dot shows one person's weight versus their height.



It is now easy to see that **warmer weather leads to more sales**, but the relationship is not perfect.

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 variable as qualitative or quantitative.

a.	The number of consumers who refuse to answer a telephone survey.	
b.	The winning time for a horse running in a race.	
c.	Gender of an employee of a garment factory.	
d.	Ethnic origin of a candidate for a public office.	
e.	Brands of soft drinks sold in a café.	

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

a.	Number of persons in a family	
b.	Color of cars	
c.	Marital status of people	
d.	Number of errors in a person's credit report	
e.	Number of typographical errors in newspapers	
f.	Monthly TV cable bills	

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 number of calls received by person in different days:

Days	:	1	2	3	4	5	Total
Number of calls:		7	8	5	15	10	45

Represent the number of calls of different days by bar diagram and by pie diagram.

1.5 The following is the distribution of weights (in kg) of 50 persons:

Weight (in kgs)	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90	Total
Number of persons	4	8	5	12	7	5	6	3	50

Draw a histogram for the above data.

1.6 Frequency distribution of the resting pulse rate in healthy volunteers (N = 63)

Pulse/min	No. of volunteers
60-65	2
65-70	7
70-75	11
75-80	15
80-85	10
85-90	9
90-95	6
95-100	3
Total	63

a) Represent the data by histogram and frequency curve.

b) Find the percentage of volunteers in whose less than 85 pulses are counted.

c) Find the percentage of volunteers in whose above and 70 pulses are counted.

d) Find the number of volunteers in whose less than 90 pulses are counted.

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, the midpoints of the class are:

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