Introduction to Statistics and Data Analysis

DR. KAVITA

ASSOCIATE PROFESSOR

SCHOOL OF MATHEMATICS, TIET PATIALA

WHAT IS STATISTICS?

- Statistics deals with the collection, analysis, interpretation, and the presentation of the scientific data.
- The Japanese industrial miracle which began in the middle of 19th century is mostly because of use of statistical methods and statistical thinking among the management.
- In current days we can say 'DATA is SUPREME'. Its like a mine and you can dig information as valuable as gold out of it.

TYPES Of STATISTICS.

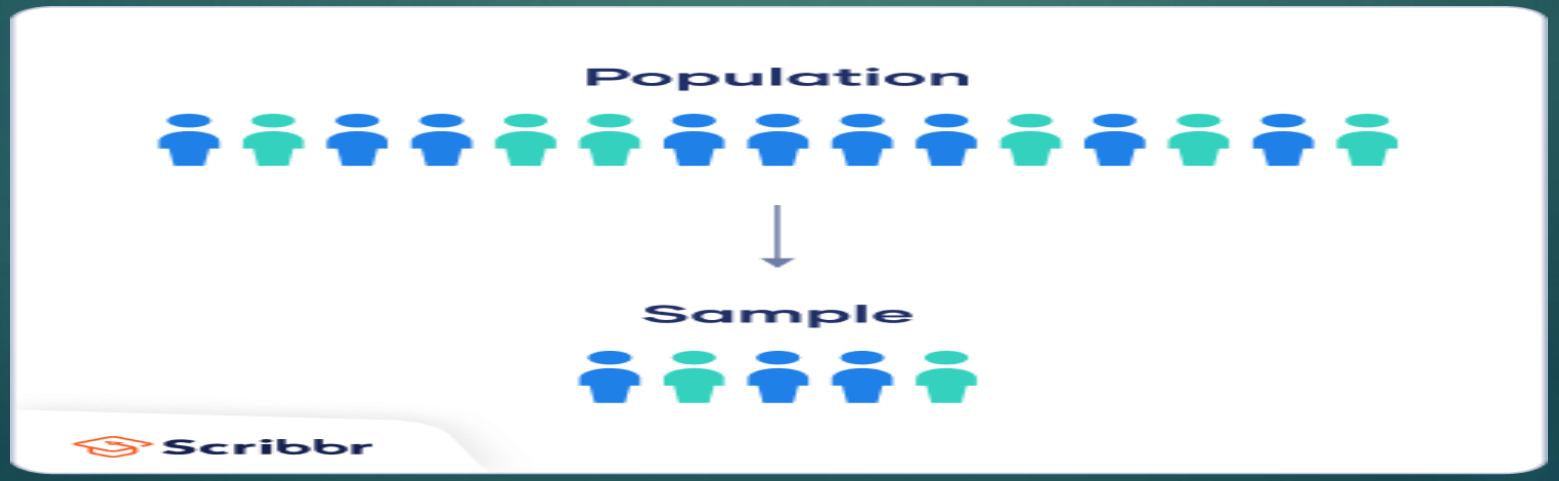
- Statistics can be classified into two different categories. The two different types of Statistics are:
  - Descriptive Statistics
  - Inferential Statistics
- In Statistics, descriptive statistics describes the data, whereas inferential statistics helps you make predictions from the data.

Comparison Chart

Basis for comparison	Descriptive statistics	Inferential statistics
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Function	It explains the data, which is already known, to summarize sample.	It attempts to reach the conclusion to learn about the population, that extends beyond the data available.

Population and Sample

- A population is entire group that you want to draw conclusions about.
- A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.
- You want to study political attitudes in young people. Your population is the 300,000 undergraduate students (POPULATION). Because it's not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers (SAMPLE) from three different universities- this is the group who will complete your online survey.



Population parameter vs. sample statistic

- When you collect data from a population or a sample, there are various measurements and numbers you can calculate from the data.
- A parameter is a measure that describes the whole population.
- A statistic is a measure that describes the sample.
- You can use estimation or hypothesis testing to estimate how likely it is that a sample statistic differs from the population parameter.
- Sampling error: A sampling error is the difference between a population parameter and a sample statistic

Experimental design

- Data for statistical studies are obtained by conducting either experiments or surveys.
- Experimental design is the branch of statistics that deals with the design and analysis of experiments.
- In an experimental study, variables of interest are identified. One or more of these variables, referred to as the factors of the study, are controlled so that data may be obtained about how the factors influence another variable referred to as the response variable, or simply the response.
- consider an experiment designed to determine the effect of three different exercise programs on the cholesterol level of patients with elevated cholesterol. Each patient is referred to as an experimental unit, the response variable is the cholesterol level of the patient at the completion of the program, and the exercise program is the factor whose effect on cholesterol level is being investigated. Each of the three exercise programs is referred to as a treatment.

Collection of Data: Sampling

- Sampling is the process of selecting a group of individuals (called sample) from a population to study them and characterize the population as a whole.
- A good sample should satisfy the below conditions-
  - Representativeness: The sample should be the best representative of the population under study.
  - Accuracy: Accuracy is defined as the degree to which bias is absent from the sample. An accurate (unbiased) sample is one that exactly represents the population.
  - Size: A good sample must be adequate in size and reliability.
- Different types of Sampling techniques: Probability sampling and Non-probability sampling.
- Probability sampling involves random selection, allowing you to make statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect initial data.

Collection of Data: Sampling

- There are four types of probability sampling techniques:
  - Simple random sampling
  - Cluster sampling
  - Systematic sampling
  - Stratified random sampling

Collection of Data: Sampling

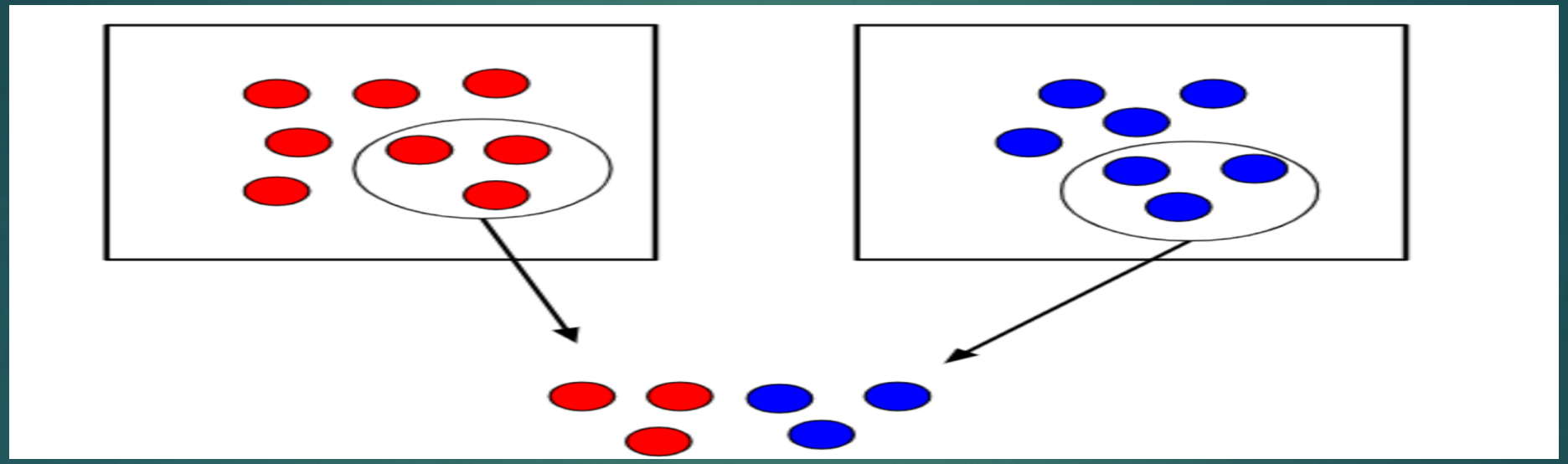
- **Simple Random Sampling:** Simple random sampling requires using randomly generated numbers to choose a sample. More specifically, it initially requires a sampling frame, a list or database of all members of a population. You can then randomly generate a number for each element and take the first n samples that you require.

Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

Id	Name	Random_num
001	Bob	6
002	Joe	3
003	Eric	4
004	Daniel	2
005	Ricky	1
006	Nathan	5

Collection of Data: Sampling

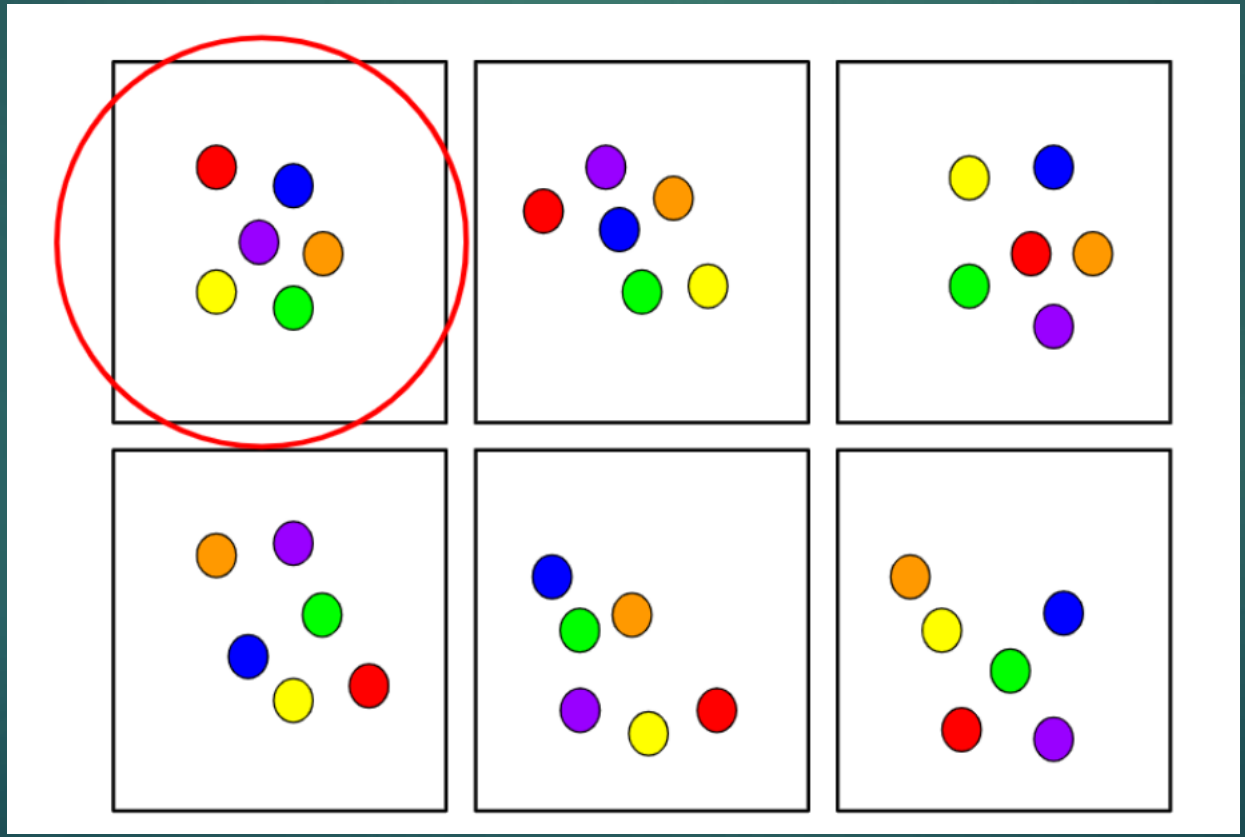
- **Stratified Random Sampling:** Stratified random sampling starts off by dividing a population into groups with similar attributes. Then a random sample is taken from each group.



- This method is used to ensure that different segments in a population are equally represented. To give an example, imagine a survey is conducted at a school to determine overall satisfaction. It might make sense here to use stratified random sampling to equally represent the opinions of students in each department.

Collection of Data: Sampling

- **Cluster Random Sampling:** Cluster sampling starts by dividing a population into groups, or clusters. What makes this different that stratified sampling is that each cluster must be representative of the population. Then, you randomly selecting entire clusters to sample.



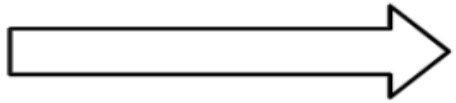
Collection of Data: Sampling

- **Systematic Random Sampling:** is a very common technique in which you sample every k'th element. For example, if you were conducting surveys at a mall, you might survey every 100th person that walks in, for example.
- If you have a sampling frame then you would divide the size of the frame, N, by the desired sample size, n, to get the index number, k. You would then choose every k'th element in the frame to create your sample.

Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

N = 6, desired sample size (n) = 2

N/n = 6/2 = 3



Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

Measures of location and variability

- Measures of location describe the central tendency of the data. They include the mean, median and mode.
- The (arithmetic) mean, or average, of n observations is simply the sum of the observations divided by the number of observations.
- The major advantage of the mean is that it uses all the data values, and is, in a statistical sense, efficient.
- The main disadvantage of the mean is that it is vulnerable to outliers. Outliers are single observations which, if excluded from the calculations, have noticeable influence on the results. or example, if we had entered '21' instead of '2.1' in the calculation of the mean of the data 1.2, 1.3, 1.4, 1.5, 2.1, then 21 is the outlier. It will change the mean from 1.50kg to 7.98kg.
- It does not necessarily follow, however, that outliers should be excluded from the final data summary, or that they always result from an erroneous measurement.

Measures of location and variability

- The median is defined as the middle point of the ordered data. It is estimated by first ordering the data from smallest to largest, and then counting upwards for half the observations. The estimate of the median is either the observation at the center of the ordering in the case of an odd number of observations, or the simple average of the middle two observations if the total number of observations is even.
- The median has the advantage that it is not affected by outliers, so for example the median for the data '1.2, 1.3, 1.4, 1.5, 2.1' would be unaffected by replacing '2.1' with '21'. However, it is not statistically efficient, as it does not make use of all the individual data values.

Measures of location and variability

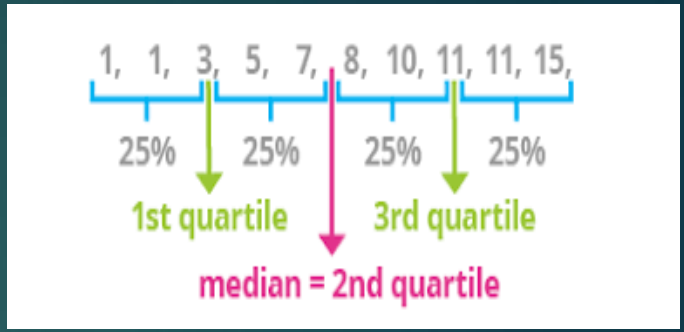
- A third measure of location is the mode. This is the value that occurs most frequently, or, if the data are grouped, the grouping with the highest frequency.
- For example, in the data 1, 1, 2, 2, 2, 3, 4, the mode is 2 as it is occurring the maximum number of times.
- It is not used much in statistical analysis, since its value depends on the accuracy with which the data are measured; although it may be useful for categorical data to describe the most frequent category.
- The expression 'bimodal' distribution is used to describe a distribution with two peaks in it. This can be caused by mixing populations. For example, height might appear bimodal if one had men and women on the population.
- If there is a single peak, then the data is called a unimodal data.

Measures of location and variability

- Measures of dispersion describe the spread of the data. They include the range, interquartile range, standard deviation and variance.
- The **range** is given as the smallest and largest observations. This is the simplest measure of variability.
- Suppose we have the data 1, 1, 3, 5, 7, 8, 10,11, 11, 15
- The range here is [1-15]
- If outliers are present it may give a distorted impression of the variability of the data, since only two observations are included in the estimate.

Measures of location and variability

- The quartiles, namely the lower quartile, the median and the upper quartile, divide the data into four equal parts
- The quartiles are calculated in a similar way to the median; first arrange the data in size order and determine the median.
- Now split the data in two (the lower half and upper half, based on the median).
- The first quartile is the middle observation of the lower half, and the third quartile is the middle observation of the upper half.
- The interquartile range is a useful measure of variability and is given by the lower and upper quartiles. In this case it is [3-11].
- The interquartile range is not vulnerable to outliers and, whatever the distribution of the data, we know that 50% of observations lie within the interquartile range.



Measures of location and variability

- The estimated average of the deviations (from mean) squared is called the variance.

$$s^2 = \frac{\sum (x - \mu)^2}{N}$$

- Standard deviation is the square root of the variance. It is is the average distance from the center(mean).

$$s = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Measures of location and variability

- Note that we use μ for the population mean and $\bar{x}$ for the sample mean.
- σ for the standard deviation of the population and 's' for the standard deviation of the sample.

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

- When calculating sample variance, we use (n – 1) in the denominator instead of n because this tends to produce better estimates.
- Reason: The sample mean encapsulates exactly one bit of information from the sample set, while the population mean does not. Thus, the sample mean gives one less degree of freedom to the sample set.

For the following sample of 6 fishes' length caught from the lake , Find the variance of the length of fish.

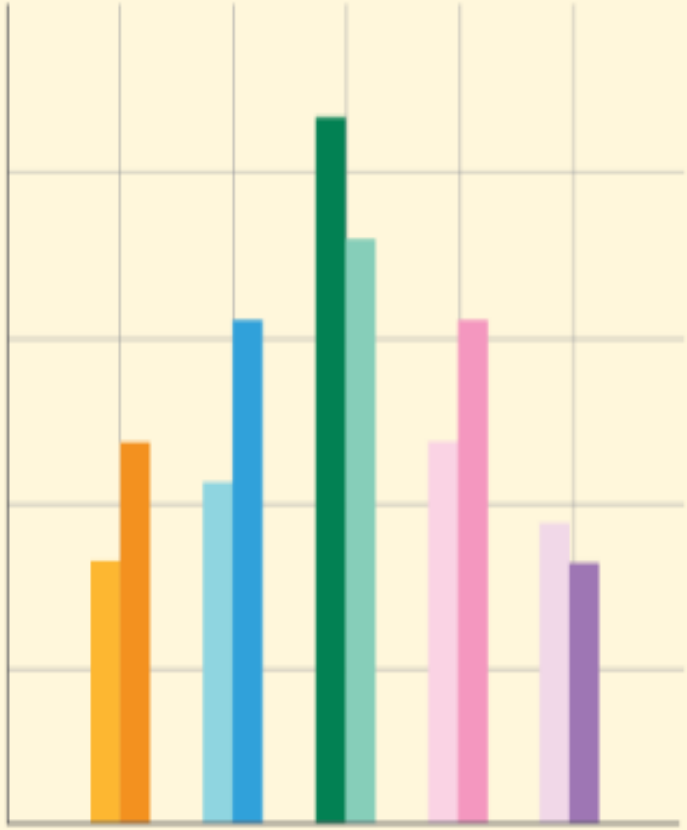
x	$(x - \bar{x})$	$(x - \bar{x})^2$
3	-3	9
4	-2	4
5	-1	1
6	0	0
8	2	4
10	4	16
Sum	0	34

First calculate the mean which is 6.

Divide this by 5.

$s^2 = 6.8$

Graphical representation of data

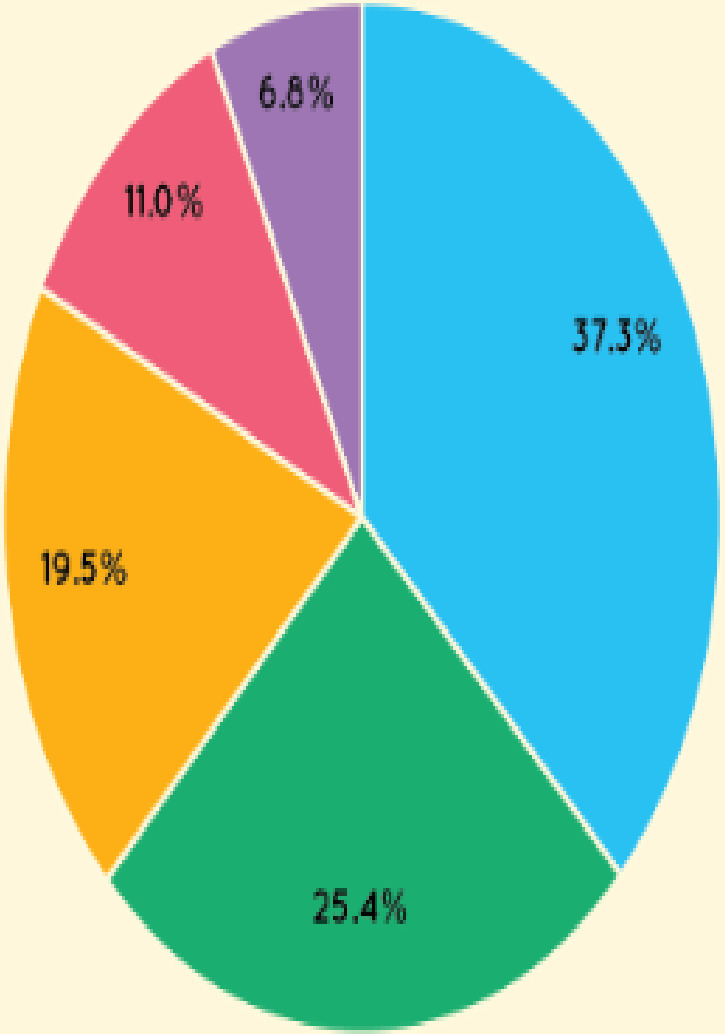


Bar Graph

A group of data represented with rectangular bars with lengths proportional to the values is a bar graph.

The bars can either be vertically or horizontally plotted.

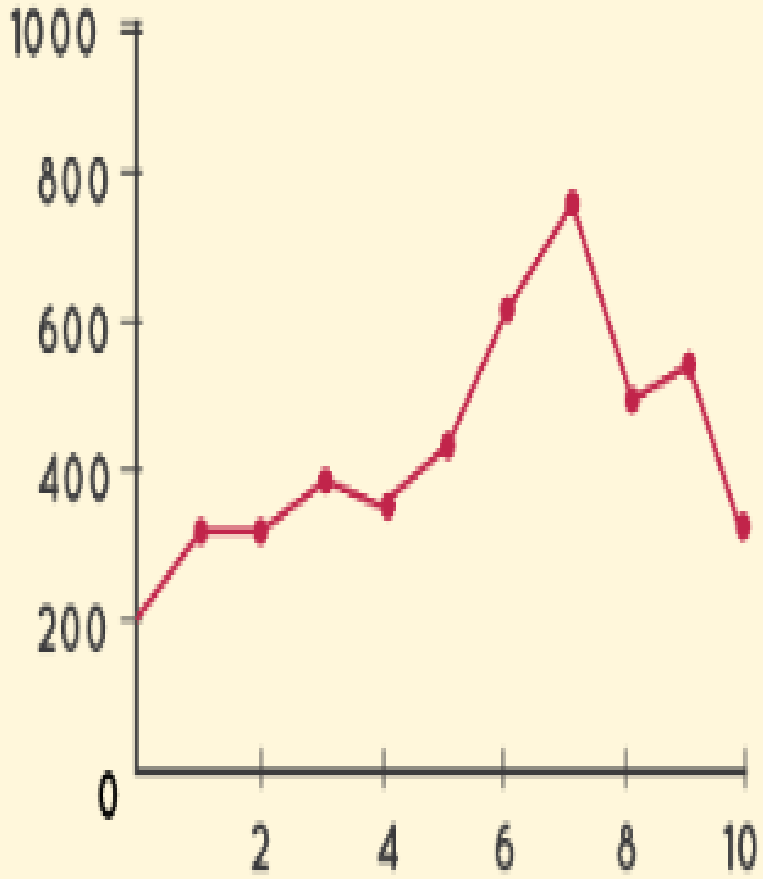
Graphical representation of data



The pie chart is a type of graph in which a circle is divided into Sectors where each sector represents a proportion of the whole. Two main formulas used in pie charts are:

- To calculate the percentage of the given data, we use the formula: (Frequency ÷ Total Frequency) × 100
- To convert the data into degrees we use the formula: (Given Data ÷ Total value of Data) × 360°

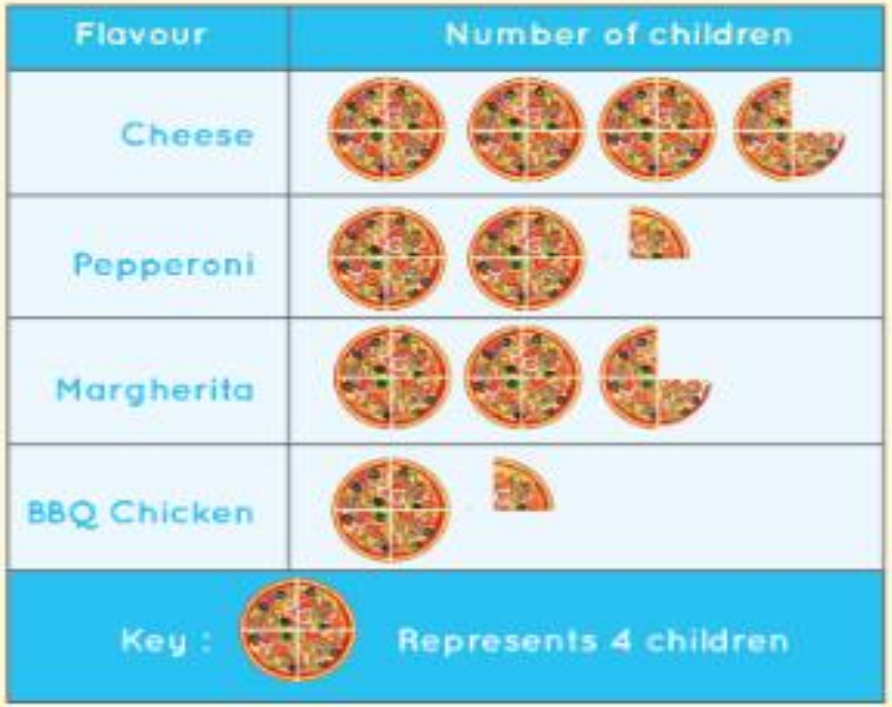
Graphical representation of data



Line graph

The line graph represents the data in a form of series that is connected with a straight line. These series are called markers.

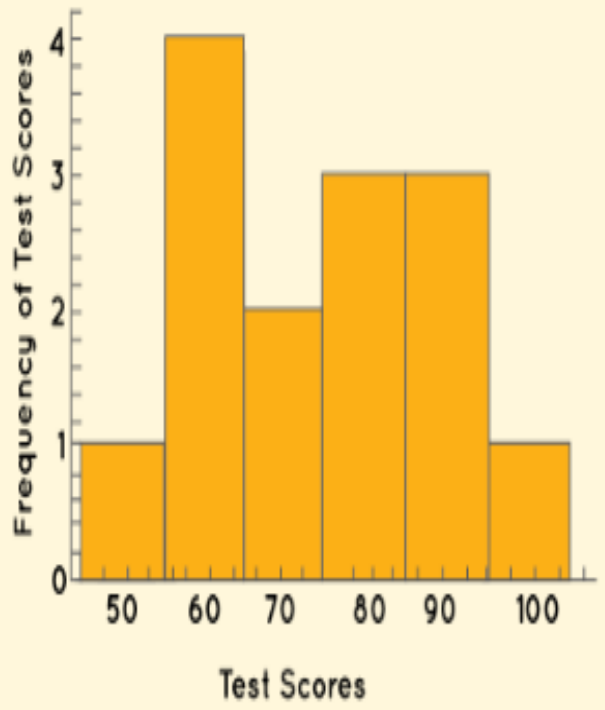
Graphical representation of data



Pictograph

Data shown in the form of pictures is a pictograph. Pictorial symbols for words, objects, or phrases can be represented with different numbers.

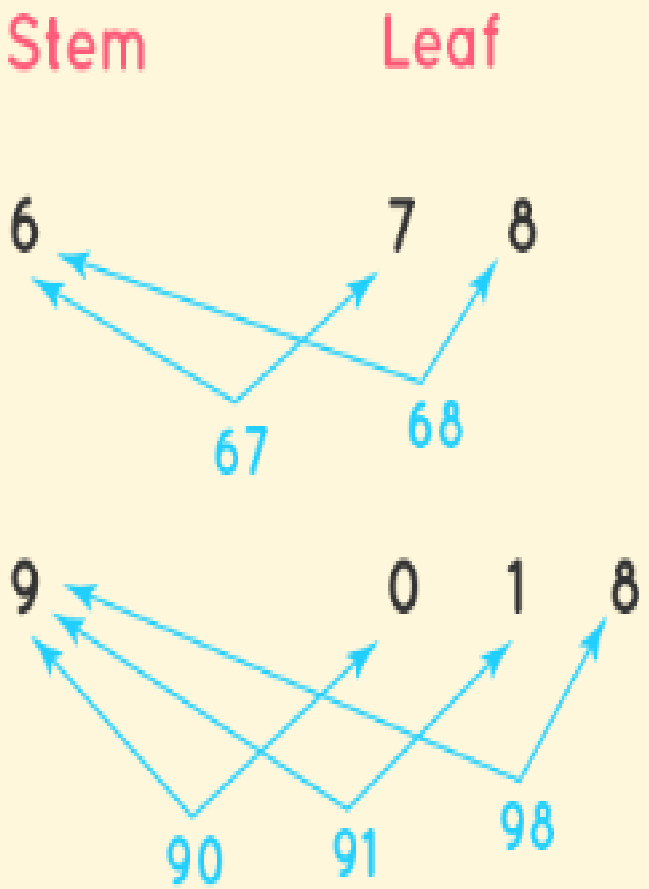
Graphical representation of data



Histogram

The histogram is a type of graph where the diagram consists of rectangles, the area is proportional to the frequency of a variable and the width is equal to the class interval. Here is an example of a histogram.

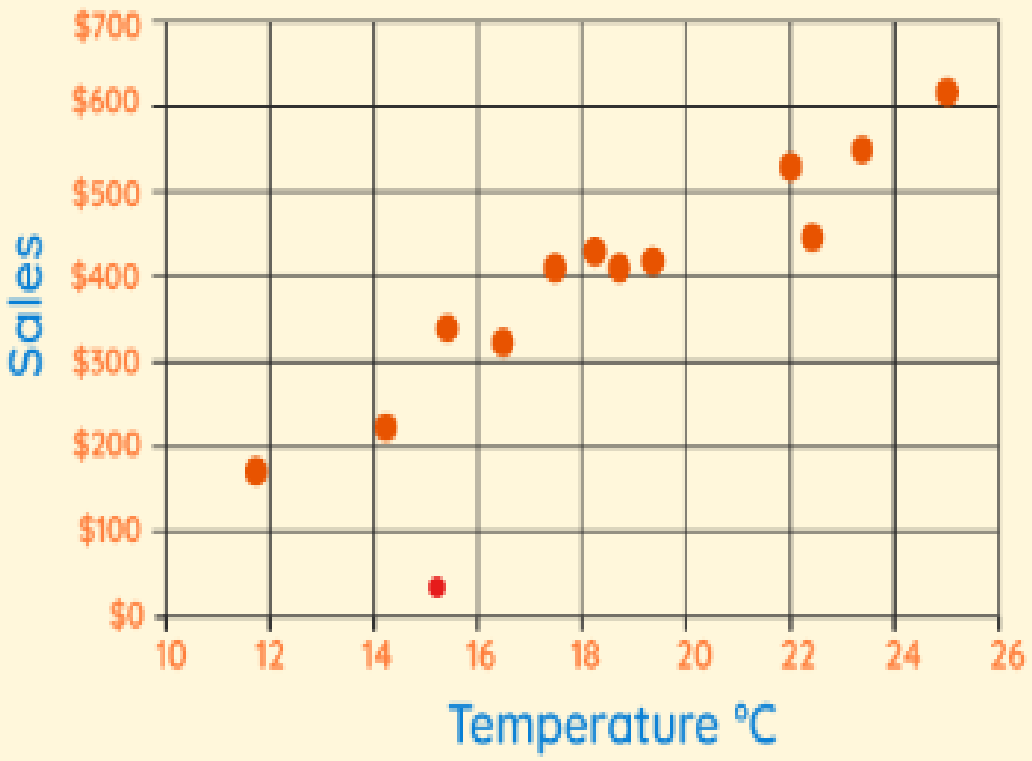
Graphical representation of data



Stem and Leaf Plot

The stem and leaf plot is a way to represent quantitative data according to frequency ranges or frequency distribution. It is a graph that shows numerical data arranged in order. Each data value is broken into a stem and a leaf.

Graphical representation of data



Scatter Plot

Scatter diagram or scatter plot is a way of graphical representation by using Cartesian coordinates of two variables. The plot shows the relationship between two variables.

Thank You!