UNIT 1

Introduction, Definition, objectives

 # What is Statistics?

Statistics is the discipline that concerns the **collection, organization, analysis, interpretation, and presentation of data.**

Data are individual facts or items of information, may be **qualitative or quantitative**.

 # Primary & Secondary Data

Primary data are the original data derived from your research endeavors. **Secondary data** are data derived from your primary data. Primary data is information collected through original or first-hand research. For example, surveys and focus group discussions. On the other hand, secondary data is information which has been collected in the past by someone else. For example, researching the internet, newspaper articles and company reports.

 # Population & Sample

Population A population consists of all the items or individuals or subjects about which you want to draw a conclusion. So, the population is the "large group" in which you are interested.

Sample A sample is the portion of a population selected for analysis. The sample is the "small group" for whom we have (or plan to have) data, often randomly selected.

 # Sample and Parameter

Parameter is a **numerical measure** that describes **a characteristic** of **a population.**

Statistic is a **numerical measure** that describes **a characteristic** of **a sample**

 # BRANCHES OF STATISTICS

Descriptive Statistics: The branch of statistics that focuses on **collecting, summarizing, and presenting** a set of data.

Inferential Statistics: The branch of statistics that analyzes sample data to draw conclusions about a population.

 # Variable: A **characteristic of an individual** that will be analyzed using statistics

Categorical (qualitative) variables have values that can only be placed into categories, such as "yes" and "no"; major; architectural style; etc.

Numerical (quantitative) variables have values that represent quantities.

- Discrete variables arise from a counting process

Examples: Number of printing errors per page on a book. Number of customers arriving at a restaurant

- Continuous variables arise from a measuring process

Examples: Height of a person, Weight of a person, Time a customer waits in a bank queue.

UNIT 2

Data Summarization

Data summarization is the first step in **statistics**, it is aimed at extracting useful information. Summary statistics are used to summarize a set of observations, to communicate the largest amount of information as simply as possible.

Data can be summarized **numerically as a table** (tabular summarization), or **visually as a graph** (data visualization).

Frequency Distribution

Frequency is how often something repeats, and a **frequency distribution** is a representation, either in a graphical or tabular format, that **displays the number of observations** within a given interval. t gives a visual display of the frequency of items or shows the number of times they occurred.

Example 1

Tally marks are often used to make a frequency distribution table. For example, let's say you survey a number of households and find out how many pets they own. The results are 3, 0, 1, 4, 4, 1, 2, 0, 2, 2, 0, 2, 0, 1, 3, 1, 2, 1, 1, 3. Looking at that string of numbers boggles the eye; a frequency distribution table will make the data easier to understand.

Number of Pets (x)	Tally	Frequency (f)
0	𝍸	4
1	𝍸 I	6
2	𝍸	5
3	III	3
4	II	2

Types of frequency distribution

Ungrouped frequency distribution: It shows the frequency of an item in each separate data value rather than groups of data values.

Grouped frequency distribution: In this type, the data is arranged and separated into groups called class intervals. The frequency of data belonging to each class interval is noted in a frequency distribution table. The grouped frequency table shows the distribution of frequencies in class intervals.

 # Steps for constructing Frequency distribution

- Sort the data in ascending order
- Calculate the range of data
- Decide on the number of intervals in the frequency distribution
- Determine the intervals.
- Decide the starting point
- Tally and count the observations under each interval.

 # Exercise:

100 schools decided to plant 100 tree saplings in their gardens on world environment day. Represent the given data in the form of frequency distribution and find the number of schools that are able to plant 50% of the plants or more?

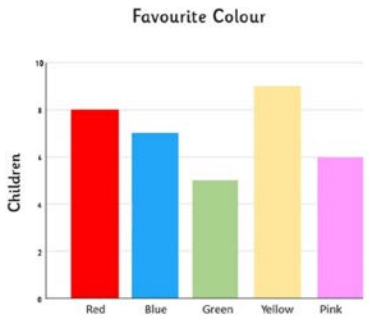
95, 67, 28, 32, 65, 65, 69, 33, 98, 96, 76, 42, 32, 38, 42, 40, 40, 69, 95, 92, 75, 83, 76, 83, 85, 62, 37, 65, 63, 42, 89, 65, 73, 81, 49, 52, 64, 76, 83, 92, 93, 68, 52, 79, 81, 83, 59, 82, 75, 82, 86, 90, 44, 62, 31, 36, 38, 42, 39, 83, 87, 56, 58, 23, 35, 76, 83, 85, 30, 68, 69, 83, 86, 43, 45, 39, 83, 75, 66, 83, 92, 75, 89, 66, 91, 27, 88, 89, 93, 42, 53, 69, 90, 55, 66, 49, 52, 83, 34, 36

Frequency Distribution Graphs

There is another way to show data that is in the form of graphs and it can be done by using a frequency distribution graph. The graphs help us to understand the collected data in an easy way. The graphical representation of a frequency distribution can be shown using the following:

 # Bar Graph:

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.



 # Pie Chart:

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. Or

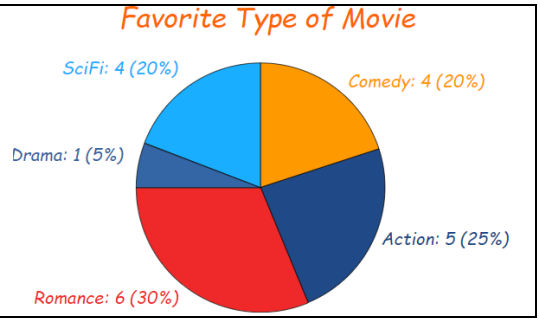
A Pie Chart is a type of graph that displays data in a circular graph. The pieces of the graph are proportional to the fraction of the whole in each category. In other words, each slice of the pie is relative to the size of that category in the group as a whole. The entire "pie" represents 100 percent of a whole, while the pie "slices" represent portions of the whole.

Imagine you survey your friends to find the kind of movie they like best:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

You can show the data by this Pie Chart:



Histograms: A histogram is a graphical presentation of data using rectangular bars of different heights. In a histogram, there is no space between the rectangular bars.

A two-dimensional graphical representation of a continuous frequency distribution is called a histogram. In histogram, the bars are placed continuously side by side with no gap between adjacent bars. That is, in histogram rectangles are erected on the class intervals of the distribution. The areas of rectangle are proportional to the frequencies.

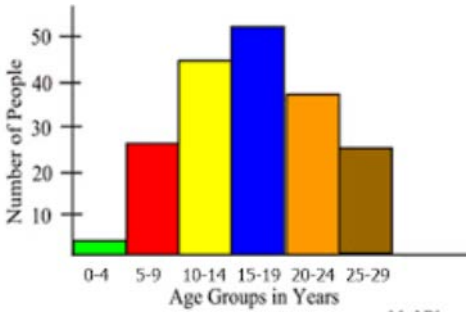
Steps of constructing Histogram:

Step 1 : Represent the data in the continuous form if it is in the discontinuous form.

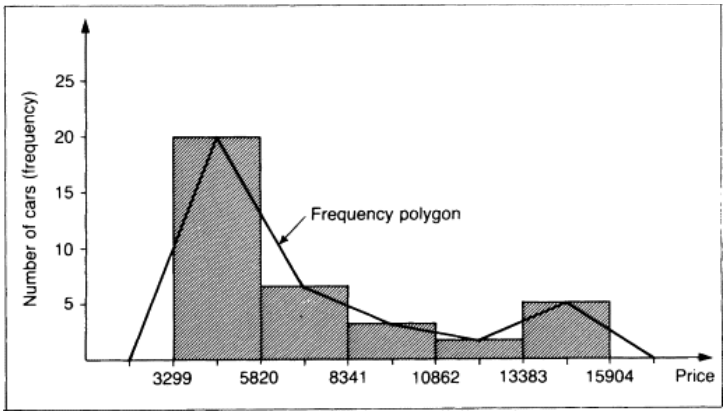
Step 2 : Mark the class intervals along the X-axis on a uniform scale.

Step 3 : Mark the frequencies/Frequency densities along the Y-axis on a uniform scale.

Step 4 : Construct rectangles with class intervals as bases and corresponding frequencies/f.d. as heights.



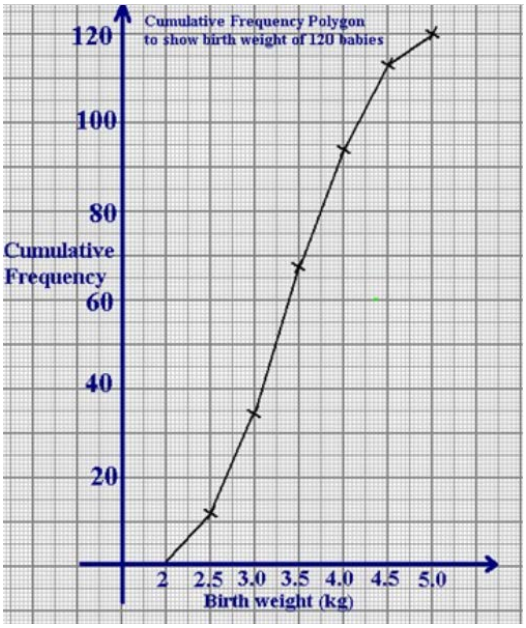
 # Frequency Polygon: A frequency polygon is drawn by joining the mid-points of the bars in a histogram.



 # Cumulative Frequency Polygon (Ogive curve):

A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results.

Birth Weight (kg)	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0	4.0-4.5	4.5-5.0
Frequency	12	22	33	27	18	8
Cumulative Frequency	12	34	67	94	112	120



UNIT 3

Measures of Location/ Central Tendency

A measure of central tendency/ Location

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called **measures of central location**. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the **median and the mode.**

The **mean, median and mode** are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Mean

The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

The three classical Pythagorean means are

- The arithmetic mean(AM)
- The geometric mean(GM), and
- The harmonic mean(HM).

The Arithmetic Mean:

The arithmetic mean is calculated by adding all of the numbers and dividing it by the total number of observations in the dataset.

For example: Arithmetic Mean of 4 + 10 + 7 is 21/3 = 7

For raw data $\mathbf{Arithmetic\ Mean} = \frac{\sum x}{n}$, where $\sum x$ is the sum of all individual's data and ***n*** is the total number of data/observation.

For frequency distribution $\mathbf{Arithmetic\ Mean\ A.M.} = \frac{\sum fx}{\sum f}, where\ f\ is\ the\ frequeny$

For Example:

For Ungrouped Data

x	5	10	15	20	25	30
f	4	5	7	4	3	2

x	f	f x x = fx
5	4	20 (4x5)
10	5	50 (5x10)
15	7	105
20	4	80
25	3	75
30	2	60
Total	**N=25**	**Σ fx =390**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{390}{25} = 15.6$$

For Grouped Data

Class interval	f1	Class mark(x1)	f1x1
0-5	4	2.5	10
5-10	6	7.5	45
10-15	10	12.5	125
15-20	16	17.5	280
20-25	12	22.5	270
25-30	8	27.5	220
30-35	4	32.5	130
TOTAL	∑ f1=60		∑ f1x1=1080

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1080}{60} = 18$$

Note

The arithmetic mean works well when the data is in an additive relationship between the numbers, often when the data is in a 'linear' relationship which when graphed the numbers either fall on or around a straight line. *i.e*. when they are **clustered.**

Geometric Mean

Not all datasets establish a linear relationship, sometimes you might expect a multiplicative or exponential relationship and, in those cases, arithmetic mean is ill-suited and might be misleading to summarize the data.

The Geometric Mean (GM) is the average value or mean which signifies the central tendency of the set of numbers by taking the root of the product of their values. Basically, we multiply the '**n**' values altogether and take out the **nth root** of the numbers, where n is the total number of values.

Geometric Mean $G.M. = \sqrt[n]{x_1\ x_2\ \ x_3 \ldots . x_n}$

i.e. The geometric mean of 5, 7 and 10 is $= \sqrt[3]{5 \times 7 \times 10} = 7.04$

Note

The geometric mean works well when the data is in an multiplicative relationship or in cases where the data is compounded; hence you multiply the numbers rather than add all the numbers.

For example

Suppose you invested $500 initially which yielded 10% return the first year, 20% return the second year and 30% return the third year. After three years, you have $500 * 1.1 * 1.2 * 1.3 = $858.00.

Whereas if you taking arithmetic mean, it's 10+20+30 = 20% return on average per year, so after three years you would have $500 * 1.2 * 1.2 * 1.2 = $864. As we can see, arithmetic mean overestimates earnings by nearly $6 which is not right since we applied an additive operation to a multiplicative process.

Investors usually consider using geometric mean over arithmetic mean to measure the performance of an investment or portfolio.

Harmonic Mean

The Harmonic Mean (HM) is defined as the reciprocal of the arithmetic mean of the reciprocals of the data values.

i.e. $H.M. = \frac{1}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+---+\frac{1}{x_n}\right)/n} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+---+\frac{1}{x_n}}$

For the numbers 4, 6 and 8 $H.M. = \frac{1}{\left(\frac{1}{4}+\frac{1}{6}+\frac{1}{8}\right)/3} = \frac{3}{\left(\frac{1}{4}+\frac{1}{6}+\frac{1}{8}\right)} = 5.54$

Note

Harmonic mean is used when we want to average units such as speed, rates and ratios.

For example: I drove at an speed of 60km/hr to Seattle downtown and returned home at a speed of 30km/hr and the distance from my house to Seattle is 20 km. What was my average speed for the whole trip?

Average speed = $\frac{1}{\left(\frac{1}{60}+\frac{1}{30}\right)/2} = 40\ km/h$ **NOT** (60+30)/2 = 45 km/hr.

Relationship among AM, GM and HM

For two Number a and b

Arithmetic mean (A.M.) = $\frac{a+b}{2}$

Geometric mean (G.M.) = $\sqrt{a \times b}$

Harmonic mean (H.M.) = $\frac{1}{(\frac{1}{a}+\frac{1}{b})/2} = \frac{2ab}{a+b} = \frac{ab}{(a+b)/2} = \frac{(GM)^2}{AM}$

$$HM = \frac{(GM)^2}{AM}$$

The harmonic mean has the least value compared to the geometric and arithmetic mean and $AM \geq GM \geq HM$

Median

Like mean median is a measure of central tendency. Median determines the middle value of a dataset listed in ascending order (i.e., from smallest to largest value). The measure divides the lower half from the higher half of the dataset.

How to Find the Median

The median can be easily found. In some cases, it does not require any calculations at all. The general steps of finding the median include:

- Arrange the data in ascending order (from the lowest to the largest value).
- Determine whether there is an even or an odd number of values in the dataset.
- If the dataset contains an odd number of values, the median is a central value that will split the dataset into halves.
- If the dataset contains an even number of values, find the two central values that split the dataset into halves. Then, calculate the mean of the two central values. That mean is the median of the dataset.

For Example

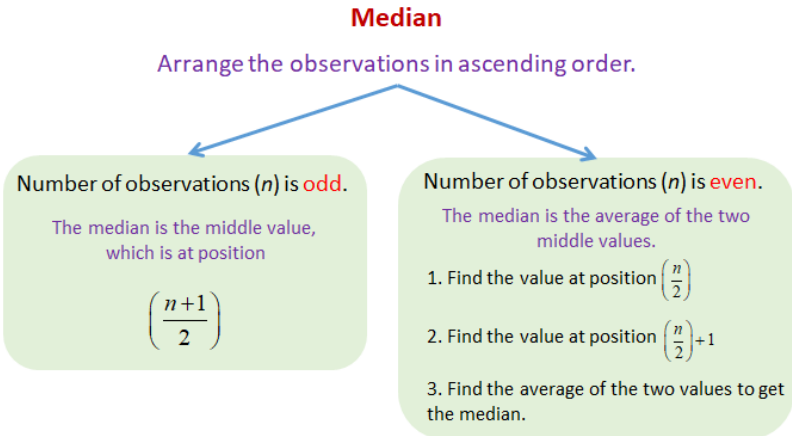
1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5



Median Class

To find the median class, we have to find the cumulative frequencies of all the classes and n/2. After that, locate the class whose cumulative frequency is greater than (nearest to) n/2. The class is called the median class.

data :

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
Frequency	6	7	15	16	4	2

Class	Frequency	Cumulative frequency	Mid-point x_i
0 – 10	6	6	5
10 – 20	7	7 + 6 = 13	15
20 – 30	15	13 + 15 = 28	25
30 – 40	16	28 + 16 = 44	35
40 – 50	4	44 + 4 = 48	45
50 – 60	2	48 + 2 = 50	55
	50		

$N = \sum f_i = 50$

Median Class = $\left(\frac{N}{2}\right)^{th}$ term

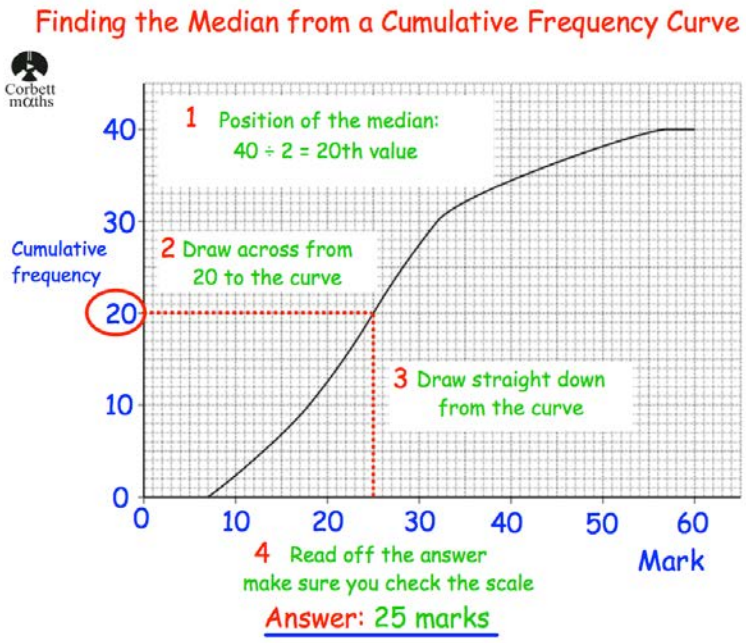
= $\left(\frac{50}{2}\right)^{th}$ term

= 25^{th} term

In above data, cumulative frequency of class 20 - 30 is 28 which is slightly greater than 25.

∴ Median class = 20 - 30

Finding Median Using Cumulative Frequency Graph



Note

As Median does not get influenced by extreme values (mean does get influenced by extreme value), so when dataset is highly fluctuating or deviating from the central value, median can be used as an appropriate measure of central tendency.

Mode:

The mode is the value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all.

When the data set has one mode, we call it **Unimodal**

For example, the mode (unimodal) in the following dataset is 19:

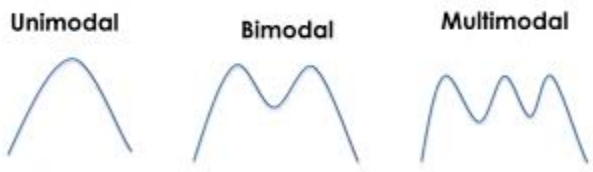
Dataset: 3, 4, 11, 15, 19, 19, 19, 22, 22, 23, 23, 26

When the data set has two modes, we call it **bimodal**

For example, the modes in the following dataset are 11 and 19:

Dataset: 3, 7, 4, 11, 15, 11, 14 19, 19, 19, 22, 20, 11, 22, 23, 23, 26

When the data set has more than two modes, we call it **multi-modal**



Note

The mode tells us the most common value in categorical data when the mean and median can't be used.

Unit - 04

Measures of Dispersion

The measures of location alone does not provide a complete or sufficient description of data. In this section, we present descriptive numbers that measures the variability or spread of the data set. Dispersion (variability, scatter, or spread) characterizes how stretched or squeezed a set of data is.

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.

Example: Let us consider a simple example to show why a measure of dispersion is so important. Consider two groups each of 6 students with their scores in a particular examination:

Group-I:	48	50	52	51	49	50
Group-II:	1	2	100	99	98	0

The arithmetic mean for each group is 50. It is very much apparent from the data that the first group consists of average or near average intelligent students and the second group is made up of very bright and very dull students.

There are many types of dispersion measures:

- Range
- Inter Quartile Range (IQR)
- Mean Deviation (MD)
- Variance/Standard Deviation
- Coefficient of variation (CV)

Range

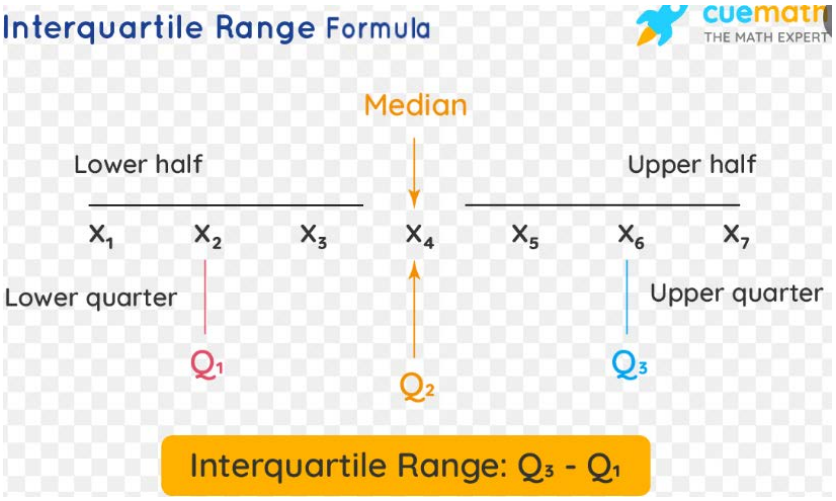
Range is the difference between the largest and smallest observations.

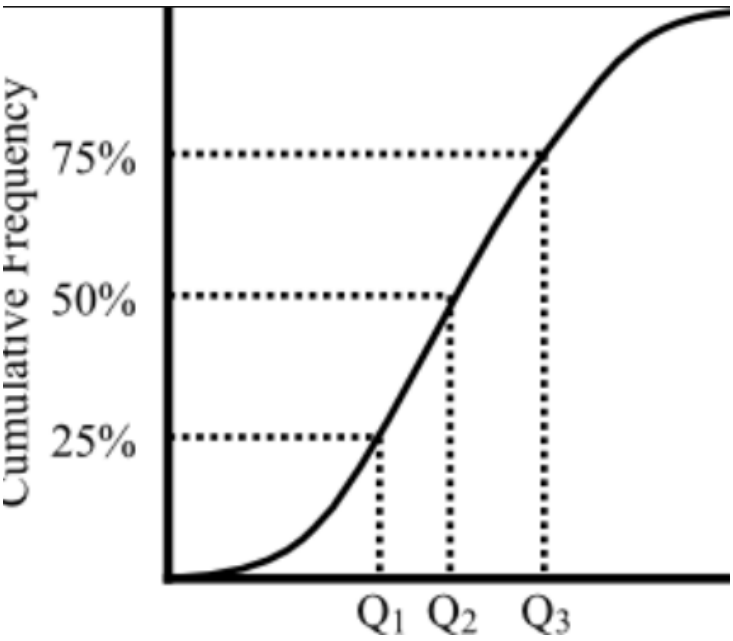
i.e. Range = Largest value – Lowest value

The greater the spread of the data from the center of the distribution, the larger the range will be. Since the range takes into account only the largest and smallest observations, it is susceptible to considerable distortion if there is an unusual extreme observation.

The interquartile range (IQR) measures the spread in the middle 50% of the data; it is the difference between the observation at Q3, the third quartile (or 75th percentile), and the observation at Q1, the first quartile (or 25th percentile).

Thus, interquartile range IQR = Q3 – Q1





Mean Deviation (MD) or Mean Absolute deviation (MAD)

Mean deviation is used to compute how far the values in a data set are from the center point. the mean deviation is used to calculate the average of the absolute deviations of the data from the central point.

MD or MAD $= \frac{\Sigma|x-\mu|}{n}$

Example

You and your friends have just measured the heights of your dogs (in millimeters):

The heights are: 600mm, 470mm, 170mm, 430mm and 300mm

Find the mean deviation

The heights are: 600mm, 470mm, 170mm, 430mm and 300mm

The **mean**: $\mu = (600 + 470 + 170 + 430 + 300)/5 = 1970/5 = 394$ mm

x	$\lvert x - \mu \rvert$
600	206
470	76
170	224
430	36
300	94
	$\sum \lvert x - \mu \rvert = 636$

$$\text{MD or MAD} = \frac{\sum |x-\mu|}{n} = \frac{636}{5} = 127.2\ mm$$

Standard Deviation or Variance

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by s in case of sample and Greek letter σ (sigma) in case of population. The formula for calculating standard deviation is as follows:

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \quad \text{or } \sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

S = Sample standard deviation

σ = Population standard deviation

Variance is the square of standard deviation.

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1} \quad \text{or } \sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

For frequency or grouped frequency distribution

$$S = \sqrt{\frac{f(x-\bar{x})^2}{\sum f - 1}} \qquad \sigma = \sqrt{\frac{f(x-\mu)^2}{\sum f}}$$

Note that throughout the course we will use

$$\sigma = \sqrt{\frac{(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}$$

OR $$\sigma = \sqrt{\frac{f(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum fx^2}{n} - (\bar{x})^2}$$ for frequency distribution

Coefficient of variation

Coefficient of variation is a type of relative measure of dispersion. It is expressed as the ratio of the standard deviation to the mean. The coefficient of variation is a dimensionless quantity and is usually given as a percentage. It helps to compare two data sets on the basis of the degree of variation. If there are data sets that have different units then the best way to draw a comparison between them is by using the coefficient of variation. The higher the CV, the greater the dispersion.

$$C.V. = \frac{\sigma}{\mu} \times 100\% = \frac{\sigma}{\bar{x}} \times 100\%$$

$$CV\ (\%) = \left(\frac{Standard\ deviation}{Mean}\right) \times 100$$

Five Number Summary & Box and Whisker Plot

A five-number summary is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value.

These values have been selected to give a summary of a data set because each value describes a specific part of a data set: the median identifies the centre of a data set; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations provide additional information about the actual dispersion of the data. This makes the five-number summary a useful measure of spread.

A five-number summary can be represented in a diagram known as a box and whisker plot. In cases where we have more than one data set to analyze, a five-number summary with a corresponding box and whisker plot is constructed for each.