Population: A population is the entire group that you want to draw conclusions about.

Sample: The sample is an unbiased subset of the population that best represents the whole data.

The process of collecting data from a small subset of the population and then using it to generalize over the entire set is called **Sampling**.

Descriptive measures.

Here our objective is to develop measures that can be used to summarize a dataset.

These descriptive measures are quantities whose values are determined by the data.

Most commonly used descriptive measures can be categorized as

Measures of central tendency: These are the measures that indicate the most typical value or center of a dataset.

Measures of dispersion: These measures indicate the variability or spread of a dataset.

A measure of central tendency tells us that where the data is concentrated or what is the most typical value of a data set.

The most commonly used measures of central tendency is the mean.

Mean: The mean of a dataset is the sum of the observations divided by the no of observations.

→ The mean is usually referred to as average.

→ For discrete observations $x_1, x_2 ----- x_N$.

$$\text{Sample mean } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + -- + x_n}{n}$$

where $x_1, x_2 --- x_n$ are the elements in the sample

Population mean $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{x_1 + x_2 + \cdots + x_N}{N}$

Consider the examples

Dataset 1. 2, 12, 5, 7, 6, 7, 3 $\bar{x} = 6$

Dataset 2 2, 105, 5, 7, 6, 7, 3 $\bar{x} = 19.285$

Dataset 3 2, 105, 5, 7, 6, 7 $\bar{x} = 21.33$

When we observe Dataset I and 2, there is so much difference in means even though all observations are same except one.

i.e Mean was very sensitive to outliers

Mean for grouped data : (Discrete single value data)

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}$$

Example : observations are 2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

x_i	f_i	$f_i x_i$
1	2	2
2	3	6
3	5	15
4	4	16
5	1	5
	15	

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{2+6+15+16+5}{15} = \frac{44}{15} = 2.93.$$

Mean for grouped data : Continuous data.

Class interval	Frequency	Midpoint (m_i)	$f_i m_i$
30 – 40	3	35	105
40 – 50	6	45	270
50 – 60	18	55	990
60 – 70	17	65	1105
70 – 80	4	75	300
80 – 90	2	85	170
	50		2940

$$\text{Mean} = \frac{\sum f_i m_i}{\sum f_i} = 58.8.$$

Thus 58.8 is not the actual mean. 8 approximating only with the midpoint, we are not taking the actual values. (The best approximation). But for discrete single value data, we get exact mean.

It may be noted that if the values of x or (and) f are large the calculation of mean is quite time consuming and tedious. The arithmetic is reduced to a greatest extent by using the following method.

Let $d_i = \frac{x_i - A}{H}$ where A is any arbitrary point.

Then $f_i d_i = \frac{f_i(x_i - A)}{h} = \frac{f_i x_i - A f_i}{h}$

Then $\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} \frac{f_i x_i - A f_i}{h} = \frac{1}{h}\left(\sum_{i=1}^{n} f_i x_i - \sum_{i=1}^{n} A f_i\right)$

$\Rightarrow \frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{N}\cdot\frac{1}{h}\left(\sum_{i=1}^{n} f_i x_i - \sum_{i=1}^{n} A f_i\right)$

$\Rightarrow h\cdot\frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{\sum_{i=1}^{n} f_i x_i}{N} - \frac{A}{N}\sum_{i=1}^{n} f_i$

$\Rightarrow \frac{h}{N}\sum_{i=1}^{n} f_i d_i = \bar{x} - \frac{A}{N}\cdot N = \bar{x} - A$

$\Rightarrow \bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i$

Example : Calculate the mean for the following frequency distribution

Class interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

Solution : Here we take $A = 28$ & $h = 8$

Class interval	Mid value (x_i)	Frequency (f_i)	$d = \frac{x_i - A}{h}$	$f_i d_i$
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16
24-32	28	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
		77		-25

$$\bar{x} = A + \frac{h}{N}\sum f_i d_i = 28 + \frac{8\times(-25)}{77} = 28 - \frac{200}{77} = 25.404$$

Properties of Arithmetic Mean

Property 1. Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If x_i are the observations and f_i their respective frequencies, then $\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$. $\bar{x}$ being the mean of distribution.

Proof:
$$\sum_i f_i(x_i - \bar{x}) = \sum_i f_i x_i - \bar{x}\sum_i f_i = \sum_i f_i x_i - \bar{x}\cdot N$$

Also $\bar{x} = \frac{\sum f_i x_i}{N} \Rightarrow \sum f_i x_i = N\bar{x}$

$$\sum_{i=1}^{n} f_i(x_i - \bar{x}) = N\bar{x} - \bar{x}N = 0$$

Property: 2: The sum of the squares of the deviations of a set of values is minimum when taken about mean.

Proof: Let $Z = \sum_{i=1}^{n} f_i(x_i - A)^2$ where A is an arbitrary point.

We have to prove that Z is minimum when $A = \bar{x}$

Z would be minimum for $\frac{dZ}{dA} = 0$ and $\frac{d^2Z}{dA^2} > 0$

$$\frac{dZ}{dA} = \sum_{i=1}^{n} f_i\, 2(x_i - A)(-1)$$
$$= -2\sum_{i=1}^{n} f_i(x_i - A) = 0 \Rightarrow \sum_{i=1}^{n} f_i x_i - A f_i = 0$$
$$\Rightarrow \sum f_i x_i = A\sum_i f_i \Rightarrow A = \frac{1}{N}\sum f_i x_i = \bar{x}$$

$$\frac{d^2Z}{dA^2} = -2\sum_i f_i(-1) = 2\sum f_i = 2N > 0$$

Hence Z is minimum at the point $A = \bar{x}$

Property 3: (Mean of the Composite Series): If $\bar{x}_i$ $i = 1, 2, 3 \ldots$ are the means of (K composite series) sizes n_i $(i = 1, 2, 3 \ldots k)$ respectively then the mean $\bar{x}$ of the composite series obtained on combining the composite series given by the formula $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}$

Proof:

$$\bar{x}_1 = \frac{1}{n_1}(x_{11} + x_{12} + \cdots + x_{1n_1})$$

$$\bar{x}_2 = \frac{1}{n_2}(x_{21} + x_{22} + \cdots + x_{2n_2})$$

$$\vdots$$

$$\bar{x}_k = \frac{1}{n_k}(x_{k1} + x_{k2} + \cdots + x_{kn_k})$$

$$\bar{x} = \frac{(x_{11} + x_{12} + \cdots + x_{1n_1}) + \cdots + (x_{k1} + x_{k2} + \cdots + x_{kn_k})}{n_1 + n_2 + \cdots + n_k}$$

$$= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum n_i \bar{x}_i}{\sum n_i}$$

Example: The average weekly salary of male employees in a firm was 5,200/- and that of females was 4,200/-. The mean salary of all the employees was 5000/-. Find the percentage of male and female employees