Statistics

- It is the science of assembling, analyzing, characterizing, and interpreting the collection of data.
- Generally the data are characterized by,

1. **Measure of Central Tendency:** Data shows a tendency to concentrate at certain values
2. **Measure of Dispersion:** Data varie about a measure of central tendency

Measures of Central Tendency

- Mean
- Mode
- Median

Arithmetic Mean

Arithmetic Mean of a set of numbers X1, X2, X3,……..XN denoted by $\bar{x}$ and is defined as

$$Mean = \frac{Sum\ of\ the\ items}{Number\ of\ the\ items} = \frac{\sum X}{N}$$

Weighted Arithmetic Mean:

- **Direct Method :**

$$\overline{X} = \frac{\sum fm}{N}$$

Example : Calculated the Arithmetic Mean DIRC Monthly Users Statistics in the University Library

Month	No. of Working Days	Total Users
Sep-2011	24	11618
Oct-2011	21	8857
Nov-2011	23	11459
Dec-2011	25	8841
Jan-2012	24	5478
Feb-2012	23	10811
Total	**140**	**57064**

$$\text{Mean} = \frac{\text{Total number of users}}{\text{Total number of working days}}$$

$$= \frac{\sum X}{N} = \frac{57064}{140} = \mathbf{407.6}$$

Example: Find mean from the following data

Marks class	Frequency (f)	Mid value (m)	mf
0-10	2	5	10
10-20	18	15	270
20-30	30	25	750
30-40	17	35	595
40-50	3	45	135
	sum= 70		Sum= 1760

$$Mean = \frac{\sum mf}{\sum f} = 25.14$$

Advantages of Mean

- It is easy to understand & simple to calculate.
- It is based on all the values.
- It is rigidly defined.
- It is not based on the position in the series.

Disadvantages of Mean

- It is affected by extreme values.
- It cannot be located graphically.
- It gives deceptive (misleading) conclusions.

Geometric Mean

- It finds application in cases like populations where we are concerned with a quantity whose changes tend to be directly proportional to the quantity itself.
- Row data:

$$GM = \sqrt[n]{x_1 * x_2 * \ldots * x_n} \Rightarrow GM = Anti\log\left(\frac{\sum \log x_i}{n}\right)$$

- Frequency distribution:

$$GM = Anti\log\left(\frac{\sum f_i \log x_i}{\sum f_i}\right)$$

Harmonic Mean

The harmonic mean is useful in limited situations where time, rate or prices are involved.

- Row data: $$HM = \frac{n}{\sum \frac{1}{x}}$$

- Frequency distribution: $$HM = \frac{\sum f}{\sum \frac{f}{x}}$$

Median

Median is a central value of the distribution, or the value which **divides the distribution in equal parts**, each part containing equal number of items.

Calculation of Median –Discrete series :

i. Arrange the data in ascending or descending order. Then the median of this ordered set of values is the value x at (n+1)/2-th position if n is odd and average of x at (n/2)-th and (n/2)+1 positions if n is even.

ii. Discrete distribution: median in that value which corresponds to **((N+1)/2)-th cumulative frequency**

Calculation of median – Continuous series

For calculation of median in a continuous frequency distribution the following formula will be employed. Algebraically,

$$Median(M) = L1 + \frac{\frac{N}{2} - cf}{f} \times i$$

Example: Median of a set Grouped Data in a Distribution of Respondents by age

Age Group	Frequency (f)	Cumulative frequencies(cf)
0-20	15	15
20-40	32	47
40-60	54	101
60-80	30	131
80-100	19	150
Total	150	

$$\text{Median (M)} = 40 + \frac{\frac{150}{2} - 47}{54} \times 20$$

$$= 40 + \frac{75 - 47}{54} \times 20$$

$$= 40 + \frac{28}{54} X 20$$

$$= 40 + 0.52X20$$

$$= 40 + 10.37$$

$$= \mathbf{50.37}$$

Advantages of Median:

- Median can be understood even by common people
- Median can be determined even with the extreme items
- It can be located graphically
- It is most useful dealing with qualitative data

Disadvantages of Median:

- It is not based on all the values.
- It is not capable of further mathematical treatment.

Mode

- Mode is the most frequent value or score in the distribution.
- It is denoted by the capital letter Z.
- highest point of the frequencies

The exact value of mode can be obtained by the following formula.

$$Z = L_1 + \frac{f1 - f0}{2f1 - f0 - f2} \times i$$

Example: Calculate Mode for the distribution of monthly rent Paid by Libraries in Karnataka

Monthly rent (Rs)	Number of Libraries (f)
500-1000	5
1000-1500	10
1500-2000	8
2000-2500	16
2500-3000	14
3000 & Above	12
Total	65

$$Z=2000+\frac{16-8}{2(16)-8-14}\times 500$$

$$Z=2000+\frac{8}{32-8-14}X500$$

$$Z=2000+\frac{8}{10}\times 500$$

$$Z=2000+0.8\times 500=400$$

Z=2400

Advantages of Mode

- Mode is readily understandable and easily calculated
- It is not at all affected by extreme value
- The value of mode can also be determined graphically

Disadvantages of Mode

- It is not based on all observations
- It is not capable of further mathematical manipulation
- Mode is affected to a great extent by sampling fluctuations
- Choice of grouping has great influence on the value of mode

Measure of Dispersion

- Range
- Mean Deviation
- Variance
- Standard Deviation

Range

- Range is the difference of the greatest and the least values in the distribution.

- Simplest but crude measure of dispersion

Mean Deviation

- Mean Deviation (M.D.)= $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$

In frequency distribution,

$$M.D. = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - \overline{x}|, N = \sum_{i=1}^{n} f_i$$

Variance

Variance : $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

In frequency distribution,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^2, N = \sum_{i=1}^{n} f_i$$

Standard Deviation

- Positive square root of variance

Sample S.D.

- Any experimental data may be considered as a sample of the population;
- the statistics of a sample are used to express the variability of a subset
- and supply an estimate of the standard deviation of the population is known as the sample standard deviation and is denoted by ‘s’.

For Discreet Data

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, n > 100$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, n < 100$$

For Frequency Distribution

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2}\ , N > 100$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2}\ , N < 100$$

Notes:

- When a statistician selects a sample and makes a single measurement, he/she obtain at least a rough estimate of the mean of the parent population. This single observation, however, can give no hint as to the degree of the variability in the population.
- When a second measurement is taken, however a first basis for estimating the population variability is obtained. The statistician states this fact by saying that two observation supply one degree of freedom, and so on…

Coefficient of Variation (Relative Std. Deviation)

It is used to measure

1. The variation of the same character in two or more different series has to be compared

   e.g. growth in girls/boys, pulse rate in young/old

2. The variation of two different characters in one and the same series has to be compared

   e.g. pulse rate and blood pressure, height and blood pressure

$$C.V. = \frac{S.D.}{Mean} \times 100$$

Examples:

1. Find the standard deviation of IQ of 50 boys from the following table:

I.Q. (X)	0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160
No. of Boys (f)	3	4	3	4	13	12	8	3

Class	Frq.	Xi	Xi*fi	Xi-Mean	(xi-Mean)^2	fi*(xi-Mean)^2

Class	Frq.	Xi	Xi*fi	Xi-Mean	(xi-Mean)^2	fi*(xi-Mean)^2
0-20	3	10	30	-81.2	6593.44	19780.32
20-40	4	30	120	-61.2	3745.44	14981.76
40-60	3	50	150	-41.2	1697.44	5092.32
60-80	4	70	280	-21.2	449.4	1797.76
80-100	13	90	1170	-1.2	1.44	18.72
100-120	12	110	1320	18.8	353.44	4241.28
120-140	8	130	1040	38.8	1505.44	12043.52
140-160	3	150	450	58.8	3457.44	10372.32

- Mean = 91.2
- S.D.= 37.34

2. Calculate the mean, standard deviation, variance, the coefficient of variance, range, median of the following data of blood pressure measurement:

100, 98, 101, 94, 104, 102, 108, 108

3. Verify that the standard deviation of the values 1.19, 1.20 and 1.21 is ___. What is the standard deviation of the number 2.19, 2.20 and 2.21? Explain the result of the two calculation above.

Solution:

If a constant is added to each value, the S.D. is unchanged. S.D. depends on difference among the values, not the absolute magnitude.

4. An analysis of monthly wages paid to workers in two firms A and B belong to the same industry gave the following results.

	Firm A	Firm B
No. of wages earners	986	548
Average monthly wages	52.5	47.5
Variance of distribution of wages	100	121

(a) Which firms pays out larger amounts?

(b) In which firm is these greater variability?

Thank You