0302301 - STATISTICS FOR DATA ANALYSIS

UNIT	MODULE	WEIGHTAGE
1.	STATISTICS: OVERVIEW	20%
2.	MEASURE OF DISPERSION	20%
3.	CORRELATION AND REGRESSION	20%
4.	FUNDAMENTALS OF PROBABILITY	20%
5.	STATISTICAL ANALYSIS USING R PROGRAMMING	20%

0302301 - STATISTICS FOR DATA ANALYSIS

Text Book: Business statistics by Padmalochan hazarika

Related Programming Tool: R

UNIT - 1 STATISTICS: OVERVIEW

- Introduction
- Meaning of statistics
- Function of statistics
- Scope and Importance of statistics
- Limitations of Statistics
- Measure of central tendency
  - Mean
    - Arithmetic Mean
    - Arithmetic Mean of grouped frequency distribution
    - Combined Arithmetic Mean
    - Advantages ,disadvantages of Arithmetic Mean

UNIT - 1 STATISTICS: OVERVIEW

- Median
  - Individual frequency distribution
  - Ungrouped frequency distribution
  - Grouped frequency distribution
  - Advantages ,disadvantages of Median
- Mode
  - Individual frequency distribution
  - Ungrouped frequency distribution
  - Grouped frequency distribution
  - Advantages ,disadvantages of Mode

INTRODUCTION

- The Science of Kings
- Indicates Quantities
  - No. of soldiers in a state, Volume of arms, Volume of Text.
- Modern Age - Statistics extended
  - Agriculture, economics, sociology, psychology, business, management.

MEANING OF STATISTICS

- **STATISTICS**
  - **Singular Sense**
    - Subject statistics such as physics.
  - **Plural Sense**
    - Numerical Statements of fact or data.

DATA: Observations expressed in numerical figures obtained by measuring or counting are called data.

STATISTICS DEFINED IN PLURAL SENSE

- ❑ Aggregative
- ❑ Multiplicity of factore
- ❑ Numerically expressed
- ❑ Enumerated or estimated according to a reasonable standard of accuracy
- ❑ Collected in a systematic manner for a predetermined purpose
- ❑ Placed in relation to each other

“Aggregate of facts affected to marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.”

STATISTICS DEFINED IN SINGULAR SENSE

- ❑ Statistics is the science which deals with the method of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.
- ❑ Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty.
- ❑ Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data.
- ❑ Statistics is the science and art of handling aggregate of facts - observing, enumeration, recording, classifying and otherwise systematically treating them.

FUNCTIONS OF STATISTICS

- To present fact in proper form
- To simplify raw data
- To facilitate comparison
- To help formulating policies
- To study relationship between different phenomena
- To forecast future values
- To measure uncertainty
- To test a hypothesis
- To draw valid inference

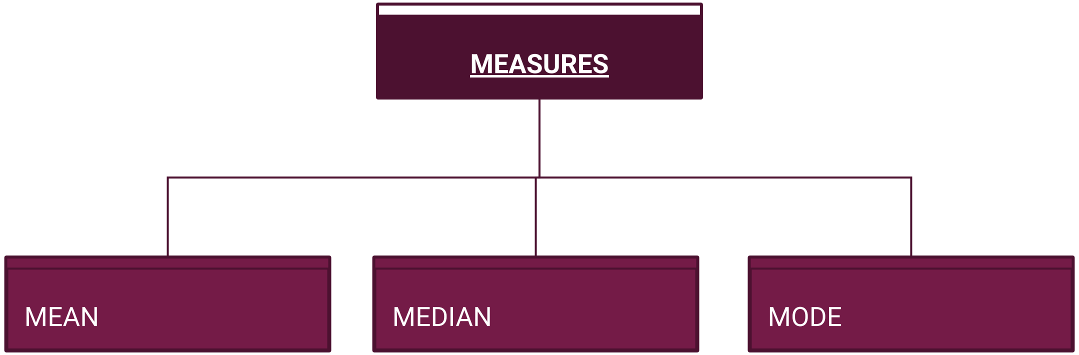
SCOPE/IMPORTANCE OF STATISTICS

- Statistics in Economics
- Statistics in Industry, Business and Commerce
- Statistics and State

LIMITATION OF STATISTICS

- Deals only with quantitative characteristics
- Does not deal with single object
- May not provide the best solution
- Can be misused.

MEASURE OF CENTRAL TENDENCY



MEAN

- ❑ **<u>Arithmetic Mean (Only in Syllabus)</u>**
- ❑ Geometric Mean
- ❑ Harmonic Mean

➢ **<u>Arithmetic Mean:</u>** The A.M. of a variable x is defined to be the sum of the values of x and divided by the number of values of x.

$x : x_1, x_2, x_3, \ldots, x_n$, $\bar{x}$ will be :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ or, } \frac{\sum x}{n}$$

i.e., $$\bar{x} = \frac{\sum x}{n}$$

MEAN

Mean

- Arithmetic Mean
- Arithmetic Mean of ungrouped frequency distribution
- Arithmetic Mean of grouped frequency distribution
    - Step-Deviation Method
- Combined Arithmetic Mean

Arithmetic mean →

ILLUSTRATIVE EXAMPLES

Example 1. (*i*) ***Find the A.M. of the following numbers:*** **5, 8, 10, 15, 24 *and* 28.**

(*ii*) ***Find the A.M. of the following series :*** ***x*** **: 4, – 2, 7, 0 *and* – 1.**

Solution: (*i*) The required A.M. $= \frac{5+8+10+15+24+28}{6} = \frac{90}{6} = \mathbf{15}.$

(*ii*) The required A.M. $\bar{x} = \frac{4+(-2)+7+0+(-1)}{5}$

$$= \frac{4-2+7+0-1}{5} = \frac{8}{5} = \mathbf{1.6}$$

MEAN

Arithmetic Mean of ungrouped frequency distribution

Example 2. Find A.M. of the following frequency distribution :

x :	1	2	3	4	5	6	7	8	9
y :	7	11	16	17	26	31	11	1	1

Solution: First of all we shall prepare the following frequency table :

Table 5.1: Calculations for A.M.

x	f	fx
1	7	7
2	11	22
3	16	48
4	17	68
5	26	130
6	31	186
7	11	77
8	1	8
9	1	9
	N = 121	$\sum fx = 555$

$$\text{A.M.}\ (\bar{x}) = \frac{\sum f x}{N} = \frac{555}{121} = \mathbf{4.59}\ (approx.)$$

MEAN

Arithmetic Mean of grouped frequency distribution

Example 3. Determine mean of the following distribution:

Daily wages (in ₹)	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of workers	:	6	5	8	15	7	6	3

Solution: Since there are three methods of obtaining mean, namely arithmetic mean (A.M.) method, geometric mean (G.M.) method and harmonic mean (H.M.) method, we shall apply A.M. method to find mean of the given distribution. (We shall see subsequently that A.M. is the most frequently used measure of central tendency.)

To find mean by applying arithmetic mean technique we form the following table :

Table 5.2: Calculations for A.M.

Wages	*Mid Value* $x = \frac{l_1 + l_2}{2}$	*No. of* **workers** (f)	fx
0 – 10	5	6	30
10 – 20	15	5	75
20 – 30	25	8	200
30 – 40	35	15	525
40 – 50	45	7	315
50 – 60	55	6	330
60 – 70	65	3	195
		$N = 50$	$\Sigma fx = 1670$

$$\text{Mean } (\bar{x}) = \frac{\Sigma f x}{N} = \frac{1670}{50} = 33.40$$

$\therefore$ The required mean (*i.e.*, average) wage = ₹ 33.40.

MEAN

Arithmetic Mean of grouped frequency distribution

- **Step - Deviation Method**

Example 4. Applying step-deviation method find A.M. of the following distribution:

Marks obtained:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students:	6	5	8	15	7	6	3

Solution: Let assumed mean A = 35

Table 5.3: Calculations for A.M.

Marks	*Mid value (x)*	*No. of students (f)*	$d = x - 35$	$d' = \frac{d}{10}$	fd'
0–10	5	6	– 30	– 3	– 18
10–20	15	5	– 20	– 2	– 10
20–30	25	8	– 10	– 1	– 8
30–40	35	15	0	0	0
40–50	45	7	10	1	7
50–60	55	6	20	2	12
60–70	65	3	30	3	9
		N = 50			$\sum fd' = -8$

Here $A = 35,\ h = 10$

Now, A.M. $(\bar{x}) = A + \frac{\sum f d'}{N} \times h$

$$= 35 + \frac{-8}{50} \times 10$$

$$= 35 - 1.6$$

$$= 34.4$$

33.4

MEAN

CLASS EXERCISE

Example 5. Determine arithmetic mean of the following distribution:

Height (in cm.) :	130-134	135-139	140-144	145-149
Frequency :	5	15	28	24
Height (in cm.) :	150-154	155-159	160-164	
Frequency :	17	10	1	

MEAN

CLASS EXERCISE-SOLUTION

Solution: We denote height by the variable x and we take the assumed mean A of x to be 147.

Table 5.4: Calculations for A.M.

Class interval	*Mid value* (x)	*Frequency* (f)	$d = x - A$, $A = 147$	$d' = \frac{d}{10}$, $h = 5$	fd'
130–134	132	5	– 15	– 3	– 15
135–139	137	15	– 10	– 2	– 30
140–144	142	28	– 5	– 1	– 28
145–149	147	24	0	0	0
150–154	152	17	5	1	17
155–159	157	10	10	2	20
160–164	162	1	15	3	3
		N = 100			$\sum fd' = -33$

$$\text{Now, A.M.} \quad (\bar{x}) = A + \frac{\sum f d}{N} \times h = 147 + \frac{-33}{100} \times 5$$

$$= 147 - \frac{33}{20} = 147 - 1.65 = 145.35$$

i.e., the required A.M. = **145.35 cm.**

MEAN

Combined/Composite Arithmetic Mean

Example 7. The average marks obtained by two groups of students in an examination are 75 and 85. If the average marks of all the students is 80, find the ratio of students in the two groups.

Solution : Let x denote marks of all the students, x_1 denote marks of the first group, x_2 denote marks of the second group, n_1 denote no. of students of the first group and n_2 denote number of students of the second group. Then we have, $\bar{x}_1 = 75$, $\bar{x}_2 = 85$, $\bar{x} = 80$. We are to determine $n_1 : n_2$.

Now, $$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

or, $$80 = \frac{75n_1 + 85n_2}{n_1 + n_2}$$

or, $80(n_1 + n_2) = 75n_1 + 85n_2$

or, $80n_1 + 80n_2 = 75n_1 + 85n_2$

or, $80n_1 - 75n_1 = 85n_2 - 80n_2$

or, $5n_1 = 5n_2$

$$\therefore \frac{n_1}{n_2} = \frac{5}{5} = \frac{1}{1}$$

i.e., $\boldsymbol{n_1 : n_2 = 1 : 1}$.

MEAN

Combined/Composite Arithmetic Mean

CLASS EXERCISE

Example 8. The average daily wage of 100 workers in a factory is ₹ 72.00. The average daily wage of 70 male workers is ₹ 75.00. Find the average daily wage of female workers.

MEAN

Combined/Composite Arithmetic Mean **CLASS EXERCISE-SOLUTION**

Example 8. The average daily wage of 100 workers in a factory is ₹ 72.00. The average daily wage of 70 male workers is ₹ 75.00. Find the average daily wage of female workers.

Solution: Let x denote the wages of all the workers, x_1 denote the wages of male workers, x_2 denote the wages of female workers, n_1 denote no. of male workers and n_2 denote no. of female workers. Then we have,

$$n_1 = 70,\ \bar{x}_1 = ₹\,75,\ n_1 + n_2 = 100,\ x = ₹\,72$$

$$\therefore \quad n_2 = 100 - 70 = 30;\ \text{i.e., number of girls} = 30$$

Now we are to determine the average marks of 30 female students.

We know that,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\Rightarrow 72 = \frac{70 \times 75 + 30\bar{x}_2}{100} \quad \text{[Since the average wages of 70 male workers } (\bar{x}_1) = 75]$$

$$\Rightarrow 72 = \frac{5250 + 30\bar{x}_2}{100} \qquad \Rightarrow 5250 + 30\bar{x}_2 = 7200$$

$$\Rightarrow 30\bar{x}_2 = 7200 - 5250 = 1950$$

$$\Rightarrow \bar{x}_2 = \frac{1950}{30} = 65$$

MEAN

<u>Advantage</u>

- ❑ It is easy to determine and understand A.M.
- ❑ The A.M. of a distribution is based on all the values or observations of the distribution.
- ❑ It can be used for further algebraic treatment.
- ❑ The formula for A.M. is rigidly defined implying that for a given series, A.M. is unique whosoever is calculate.
- ❑ It provides a good basis for comparison.
- ❑ For obtaining A.M. of series, its values need not be arranged in a given order.
- ❑ It the A.M. and the number of observations of a distribution are known then the sum of the observations of the distribution can be known.

MEAN

Disadvantage

- ❑ Unduly affected by extreme values.
- ❑ In case even a single observation of a series is missing, one cannot determine the A.M. of the series.
- ❑ The determination of A.M. of a grouped frequency distribution is based on the unrealistic assumption that the observation of each class is concentrated at the center of that class.

MEDIAN

- The median of distribution in ascending or descending order is that observation of the distribution which divides the distribution into two equal parts.

- Values should be in ascending or descending order.

- If there are odd number of values in the series, the median will be :

$$\left(\frac{n+1}{2}\right) \text{th value}$$

- If there are even number of values in the series, the median will be :

$$\frac{n}{2} \text{th value and the } \left(\frac{n}{2}+1\right) \text{th value}$$

MEDIAN

1. 77, 73, 72, 70, 75, 79, 78

Solution: (*i*) Arranging the values of the series in ascending order, we get,

70, 72, 73, 75, 77, 78, 79

No. of terms in the series = 7 = An odd number

$\therefore$ The required median = $\frac{7+1}{2}$ th term = 4th term = 75.

2. 94, 33, 86, 68, 32, 80, 48, 70

(*ii*) Arranging the data (values of observations) in ascending order, we get,

32, 33, 48, 68, 70, 80, 86, 94

No. of terms in the series = 8 = An even number.

Now, $\frac{n}{2}$ th term = $\frac{8}{2}$ th term = 4th term = 68 and $\left(\frac{n}{2}+1\right)$ th term = 5th term = 70.

$\therefore$ The required median = $\frac{68+70}{2}$ = 69 .

Note : By arranging the terms in descending order also we will get the same result.

MEDIAN

Median of an ungrouped frequency distribution:

Example 22. *Determine median for the following distribution:*

Wages (₹) :	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
No. of workers :	**8**	**10**	**11**	**16**	**20**	**25**	**19**	**9**	**6**

Solution:

Table 5.11: Table for determining median

Wages (₹)	*No. of workers* (f)	*Cumulative frequency* (f_c)
20	8	8
21	10	18
22	11	29
23	16	45
24	20	65
25	25	90
26	19	109
27	9	118
28	6	124
	N = 124	

Here total frequency (*i.e.*, total no. of observations) = 124 which is even. Hence the A.M. of the $\frac{N}{2}$ th and the $\left(\frac{N}{2}+1\right)$ th terms will be the median.

Now $\frac{N}{2} = \frac{124}{2} = 62$ and $\frac{N}{2} + 1 = 63$. We find from the cumulative frequency column that 62nd term and 63rd term lie between 45 and 65. Since 65 is the cumulative frequency of 24 hence each of the 62nd and the 63rd terms will be 24.

Hence the required median = **₹ 24.**

MEDIAN

Median of an ungrouped frequency distribution:

CLASS EXERCISE

Example 23. Find the median marks for the following distribution of marks obtained by 18 students :

Marks obtained	:	*5*	*10*	*15*	*20*	*25*	*Total*
No. of Students	:	*3*	*4*	*2*	*5*	*4*	*18*

Solution: To find the median marks we form the following cumulative frequency table.

MEDIAN

Median of an ungrouped frequency distribution: **CLASS EXERCISE-SOLUTION**

Table 5.12: Table for determining median

Marks	*Frequency* (f)	*Cumulative Frequency* (f_c)
5	3	3
10	4	7
15	2	9
20	5	14
25	4	18
Total	N = 18	

Since $N = 18$ which is an even number, hence the A.M. of the $\frac{N}{2}$ th value and the $\left(\frac{N}{2}+1\right)$ th value will be the median.

Now, $\frac{N}{2} = \frac{18}{2} = 9$, and $\frac{N}{2} + 1 = 10$.

The value whose cumulative frequency is 9 [since $\frac{N}{2} = 9$ lies between 7 and 9 (excluding 7 and including 9) in the cumulative frequency column] is 15 and the value whose cumulative frequency is 14 is 20 (since 10 lies between 9 and 14 in the cumulative frequency column). Hence the required

A.M. $= \frac{15+20}{2}$ marks $=$ **17.5 marks.**

MEDIAN

Median of an grouped frequency distribution:

N is odd, then the median class will be that class which will contain the $\frac{N+1}{2}$ th observation. Again, if N is even then the median class will be that class which will contain the $\frac{N}{2}$ th observation. The procedure of detecting the median class is similar to the procedure of detecting the median of an ungrouped frequency distribution. After the detection of the median class the particular median value is determined by using the following formula :

$$\text{Median } (M_e) = L + \frac{\frac{N}{2} - f_c}{f} \times I \qquad \text{... (5.11)}$$

Where

L = Lower class limit (lower class boundary) in case of exclusive (inclusive) classification,

f = Frequency *i.e.*, simple frequency of the median class,

f_c = Cumulative frequency of the class preceding the median class,

N = Total frequency,

I = Length of the median class (The symbol h may be used instead of I)

MEDIAN

Median of an grouped frequency distribution:

Example 24. Determine median for the following distribution :

Daily wages (₹)	:	**50–55**	**55–60**	**60–65**	**65–70**	**70–75**
No. of workers	:	**6**	**10**	**22**	**30**	**16**
Daily wages (₹)		**75 – 80**	**80–85**			
No. of workers	:	**12**	**15**			

Solution :

Table 5.13: Table for determining median

Weekly wages (₹)	*No. of workers* (f)	*Cumulative frequency* (f_c)
50–55	6	6
55–60	10	16
60–65	22	38
65–70	30	68
70–75	16	84
75–80	12	96
80–85	15	111
	N = 111	

$\frac{N+1}{2}$ th term $= \frac{111+1}{2}$ th term = 56th term. From the cumulative frequency table we find that the 56th term lies in the class 60-70.

65-70 60-70 is the median class. Now, median $= L + \dfrac{\frac{N}{2} - f_c}{f}$

Here $L = 65,\ f = 30,\ f_c = 38,\ N = 111,\ l = 5$

$$\therefore \quad \textbf{Median} = 65 + \frac{\frac{111}{2} - 38}{30} \times 5 = 65 + \frac{55.5 - 38}{30} \times 5$$

$$= 65 + \frac{17.5}{30} \times 5 = 65 + \frac{17.5}{6}$$

$$= 65 + 2.92 = 67.92$$

i.e., the required median = **₹ 67.92.**

MEDIAN

Advantage

- ❑ Extreme values do not affect median.
- ❑ Median is easy to understand. It is also easy to determine.
- ❑ Median can also be determined graphically.
- ❑ Medians of individual distributions and ungrouped frequency distributions can be determined simply by observation.

MEDIAN

Disadvantage

- ❑ In order to determine the median of distribution, the distribution must be arranged in order. This is not needed in other measures of central tendency.
- ❑ Median of a distribution is not based on all the observations of the distribution.
- ❑ In comparison to mean it is more affected by fluctuations of sampling.

MODE

Mode of a distribution is that observation of the distribution whose frequency is the maximum.

Mode is not unique. Distribution may have more than one mode.

$$\text{Mode } (M_0) = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times I \qquad ...(5.12$$

where L = Lower limit/lower boundary of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

and I = Length of the modal class. (Instead of I, the symbol h may also be used.)

MODE

Example 25. Determine mode for the following distribution :

Marks	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30
No. of students	7	10	16	32	24	18

Marks	31 – 35	36 – 40	41 – 45
No. of students	10	5	1

Solution: Since the frequency of the class 16-20 is the maximum, hence this class is the modal class. The class intervals of the given distribution are as per the inclusive method of classification and hence in determining mode we must take the lower boundary of the modal class.

Now, Mode $= L + \dfrac{f_1 - f_0}{2f_1 - f_0 - f_2} \times l = 15.5 + \dfrac{32 - 16}{2 \times 32 - 16 - 24} \times 5$

$$= 15.5 + \frac{16}{64 - 40} \times 5 = 15.5 + \frac{16}{24} \times 5$$

$= 15.5 + 3.33 =$ **18.83 marks.**

MODE

i) 3,4,5,2,3,4,1,6,4

Ans: 4

ii) 7,9,11,7,6,5,9,13

Ans: 7 and 9

iii) 3,5,6,7,9,12,3,6,5,9,12,7

Ans: Since the frequency of each of observation is the same (being 2 in each case), hence the given series has no mode.

MODE

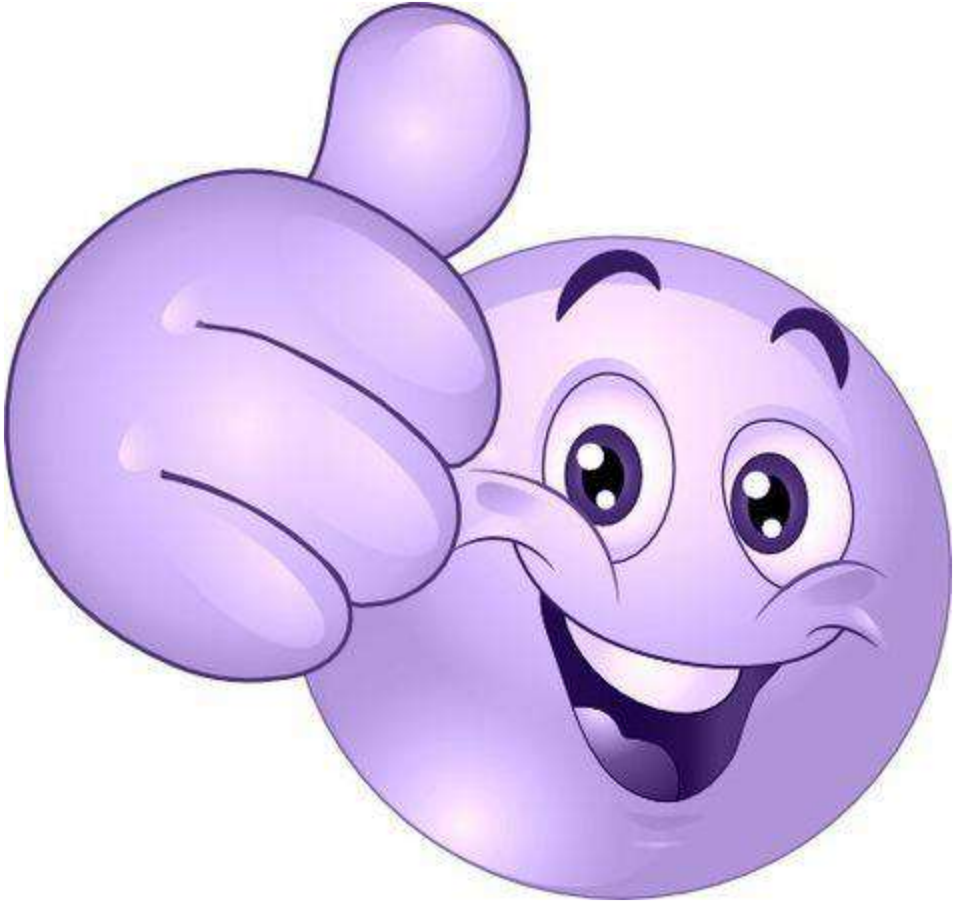
Advantage

- The mode of an ungrouped frequency distribution can be determined simply by observation.
- Mode is not affected by extreme values.
- Model is easy to understand.
- Mode can be determined graphically.

MODE

Disadvantage

- Mode is not based on all the observations.
- It is not suitable for further mathematical treatment.
- Like arithmetic mean we cannot know the sum of the observation of a distribution if we know the mode and the number of observations of the distributions.



THANK

YOU...