**Measures of Central Tendency**

A **descriptive statistic** is a summary statistics that quantitatively describes or summarizes features from a collection of information / dataset. Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study. It helps analysts to understand the data better.

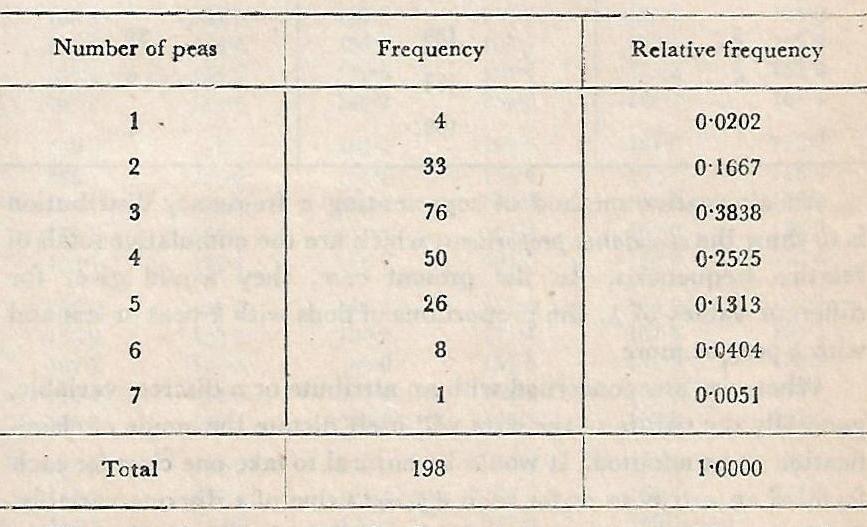
Descriptive statistics is distinguished from inferential statistics (or inductive statistics) by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance). Other important measures are ***quantiles***, ***coefficient of variation***, ***moments***, ***skewness***, ***kurtosis*** etc.

**Measures of central tendency or average**

Quite often it is found in the data to cluster around a central value (This will be evident by examining the frequency distribution). In such a case, it would be legitimate to use a ***single value*** (central value) to represent the whole set of figures.

Table 1. Frequency distribution of #peas in pea-pods



***A numerical value which represents (approximately) the entire statistical data is called a measure of central tendency or an average of that data.***

In connection with a frequency distribution, it is referred to as a measure of location because it determines the position of the distribution on the axis of the variable.

**Desirable properties of a measure of central tendency**

* It should be rigidly defined
* It should be based on all observations
* It should be readily comprehensible
* It should be affected as little as possible by sampling fluctuations
* It should readily lend itself to algebraic treatment

**There are three measures of central tendency which are used generally.** These are:

1. Mean (Arithmetic mean)
2. Median
3. Mode

**Mean: When we are telling simply 'mean', it implies** arithmetic mean, which is obtained by dividing the sum of the given values by their number.

* **Mean of ungrouped data:** The mean (arithmetic mean) of observations is
* **Mean of discrete data arranged in frequency table:** If the variates have frequencies respectively, then the mean is
* **Mean of grouped data when the frequency distribution is given in the form of classes:** In this case, it is assumed that the frequency in each class is centred at its class mark. Thus, if there are classes and , denote the frequency and the class mark respectively of the class, then the mean is given by the formula:

It may be noted that the above formula will give only an approximate value of the mean. The error of approximation will, however, be negligible provided the range of is very large compared to the width of the class-interval.

**Some important properties of the arithmetic mean**

* The total of a set of observations is equal to the product of their number and the mean, i.e.
* The sum of the deviations of the given values of a variable from its mean is necessarily zero, i.e.

***Proof:***

* If two variables and are so related that , where are constants, then the arithmetic means of and are related in the same way as and themselves are, i.e.

In particular, if , where are constants, then

***Proof:*** , or , or

Now summing over all values of we get

=

Hence, , or

This relation implies if each of the observation is increased, decreased, multiplied or divided by a constant, then the mean also will be similarly affected.

* Let there be two groups of values of , containing and values and having means and respectively, then the grand mean is

***Proof:*** Let (=1, 2,...,) and (=1, 2,...,) denote the values in the two sets respectively. The sum of values in the two sets taken together will then be equal to the sum of values in the first set plus the sum of values in the second set, i.e.

But this sum must be equal to () times the grand mean of , i.e.

or,

This can be generalized to any number of groups, i.e.

* Suppose the values of two variables and are given for each of individuals. If a new variable is formed as , then the mean of the new variable will be

***Proof:***  ( = 1, 2,...,)

Summing over all individual we have,

or,

or,

* is the least when , where are the observations, is any arbitrary constant and is the arithmetic mean.

***Proof:***

Therefore,

Both terms on the right hand side are positive. We have only to choose the value of which makes minimum possible, and this will be achieved when the second term on the right hand side has the minimum possible value 0, i.e. , or .

**Advantages of arithmetic mean**

Arithmetic mean is the simplest and easiest to understand and fulfils all the desirable criteria of a measure of the central tendency.

* It is rigidly defined based on all observations and also easy to calculate.
* If the arithmetic mean and the number of observations in each of several groups are given, they can be algebraically combined to find the arithmetic mean of the composite group.
* Arithmetic mean is also be affected very little by sampling fluctuations. If many samples are drawn from the same statistical population, arithmetic mean will be found to fluctuate very little.

**Disadvantages of arithmetic mean**

* It is affected by the presence of extreme values, i.e. extremely large or small observations.
* It cannot be calculated from frequency distributions with open-end classes.

**Computation of AM from group frequency data**

**Problem 1:** Given the following frequency distribution, calculate the mean (arithmetic mean):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Wages | 12.5-17.5 | 17.5-22.5 | 22.5-27.5 | 27.5-32.5 | 32.5-37.5 | 37.5-42.5 | 42.5-47.5 | 47.5-52.5 | 52.5-57.5 | Total |
| #Workers | 2 | 22 | 10 | 14 | 3 | 4 | 6 | 1 | 1 | 63 |

**Table 2: Calculations for arithmetic mean**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class interval | Frequency  ( | Class-mark () |  |  |
| 12.5-17.5 | 2 | 15 | -3 | -6 |
| 17.5-22.5 | 22 | 20 | -2 | -44 |
| 22.5-27.5 | 10 | 25 | -1 | -10 |
| 27.5-32.5 | 14 | 30 | 0 | 0 |
| 32.5-37.5 | 3 | 35 | 1 | 3 |
| 37.5-42.5 | 4 | 40 | 2 | 8 |
| 42.5-47.5 | 6 | 45 | 3 | 18 |
| 47.5-52.5 | 1 | 50 | 4 | 4 |
| 52.5-57.5 | 1 | 55 | 5 | 5 |
| Total | 63 | - | - | -22 |

**Exercise 1:** Find the missing frequencies in the following frequency distribution, when it is known that A.M. =11.09.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class limits | 9.3-9.7 | 9.8-10.2 | 10.3-10.7 | 10.8-11.2 | 11.3-11.7 | 11.8-12.2 | 12.3-12.7 | 12.8-13.2 | Total |
| Frequency | 2 | 5 |  |  | 14 | 6 | 3 | 1 | 60 |

**Exercise 2:** The following table gives the monthly income of 45 employees in a factory. The total monthly income of 5 employees in the class “Rs 400 & above” is Rs. 6000. Find the mean.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Income | 220-249 | 250-279 | 280-309 | 310-339 | 340-369 | 370-399 | 400 & above | Total |
| Frequency | 6 | 8 | 12 | 5 | 5 | 4 | 5 | 45 |

**Exercise 3:** Show that if be the arithmetic mean of the values weighted by ( =1, 2, ..., ), then .

**Exercise 4:** Put the following information into a frequency distribution and obtain the arithmetic mean (assuming the range of salary is Rs. 0 - Rs. 500):-

“For a group of wage earners, 20%, 40%, 70% and 80% of the wage earners receive less than Rs. 50, 120, 300 and 350 respectively; and 5% are receiving Rs. 400 and over.”

**Exercise 5:** The mean monthly salary paid to all employees in a certain company was rupees 500/-. The mean monthly salaries paid to male and female employees were 520 and 420 rupees respectively. Obtain the percentage of male to female employees in the company.

[**Ans.** E1: =12, =17; E2: 399.78; E4: 196.25; E5: Male=80%, Female=20%]

**Median**

If the given values of are arranged in an increasing or decreasing order of magnitude, then the middle-most value in this arrangement is called median.

**Median from simple series:** If there are observations arranged in ascending or descending order, then

**Median from simple frequency distribution**: When the observations are grouped into a simple frequency distribution, the median can be obtained on the basis of the cumulative frequencies.

**Table 3:** Cumulative frequency distribution of number of peas

|  |  |  |  |
| --- | --- | --- | --- |
| Number of peas | Frequency | Cumulative frequency | |
| Lass-than | More-than |
| 1 | 4 | 4 | 198 |
| 2 | 33 | 37 | 194 |
| 3 | 76 | 113 | 161 |
| 4 | 50 | 163 | 85 |
| 5 | 26 | 189 | 35 |
| 6 | 8 | 197 | 9 |
| 7 | 1 | 198 | 1 |

Here, total frequency, , which is even number

**Median from grouped frequency distribution**

The median here may be supposed to be the value for which the cumulative frequency is . Then the median value can be calculated using any one of the following two methods.

1. ***By the application of simple interpolation in a cumulative frequency distribution***

Let us denote the lower and upper class-boundaries of the class containing the median by and and the corresponding cumulative frequencies by and respectively. If we assume that cumulative frequency is a linear function of between and , then the median ( which is the value with cumulative frequency , will satisfy the following relation

where are the width and frequency of the class-interval containing the median (). The class-interval containing the median value is called as median class.

**Table 4:** Frequency distribution of height for 177 Indian adult males

|  |  |  |  |
| --- | --- | --- | --- |
| Height (cm)  Class interval | Frequency | Cumulative frequency | |
| Lass-than | More-than |
| 144.55-149.55 | 1 | 1 | 177 |
| 149.55-154.55 | 3 | 4 | 176 |
| 154.55-159.55 | 24 | 28 | 173 |
| 159.55-164.55 | 58 | 86 | 149 |
| 164.55-169.55 | 60 | 146 | 91 |
| 169.55-174.55 | 27 | 173 | 31 |
| 174.55-179.55 | 2 | 175 | 4 |
| 179.55-184.55 | 2 | 177 | 2 |

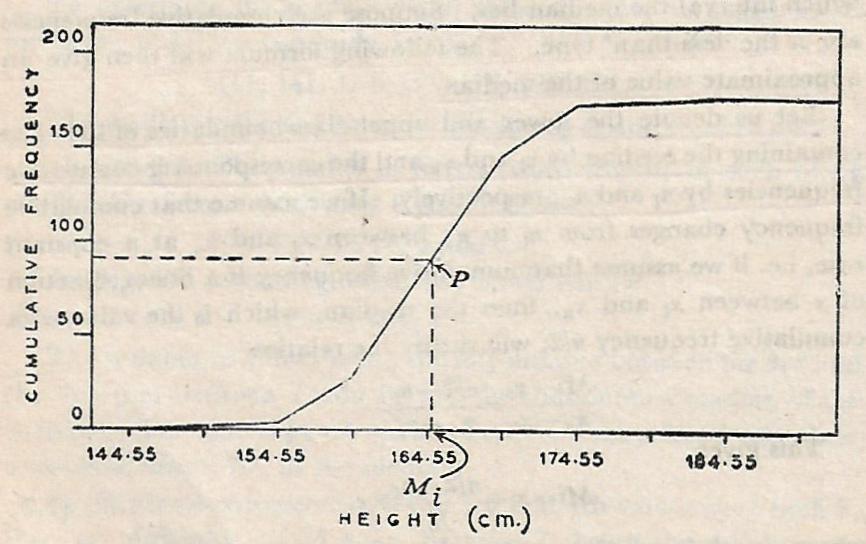
Here, total frequency and so

Examining less-than type cumulative frequency, it is found that the median lies between cm and cm, for which the cumulative frequencies are and . Here and . Therefore,

cm

1. ***Graphical method***

The median value may also be obtained geometrically, from the ogive of the frequency distribution. The median will be given by the abscissa of the point on the ogive for which the ordinate is .



**Fig. 1**: Less-than type ogive for height of 177 Indian adult males

**Advantages of median**

* It is not difficult to understand, although it is not as popular as the arithmetic mean.
* It is easy to calculate.
* It can be calculated from grouped frequency distributions with classes of unequal width or with open-end classes.
* It is unaffected by extreme values, i.e. it is not affected by very large or very small values (outliers).
* It remains the same under any transformation that leaves the order of the values unchanged.

**Disadvantages of median**

* For calculation of median data must be arranged.
* It cannot be treated algebraically. Given the medians of several groups of observations, median of the composite group cannot be determined.
* The calculation of median from a grouped frequency distribution is based on simple interpolation, which assumes that the observations in the median class are uniformly distributed. But in reality, this may not be true.
* Median is affected more by sampling fluctuations than the arithmetic mean.

**Uses of median**

* In situations, where the data contain a few extreme values widely different from the majority of the values, the median (instead of mean) would be the appropriate measure of the central tendency.
* In general, the median is the most informative measure of central tendency for skewed distributions (e.g. income distributions) or distributions with outliers.
* When the values of the variable are given in the form of a frequency table of which one or both of the terminal classes are open. In such cases, computation of mean is impossible because class-marks of those terminal classes are indeterminate. But this will generally be no bar to the computation of the median or mode.
* Median is applicable to qualitative data is psychological and social studies, where numerical measurements may not be available, but it is possible to rank the objects in some order.

**Exercise 1:** Find the median from the following simple frequency distribution:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|  | 7 | 44 | 35 | 16 | 9 | 4 | 1 | 116 |

**Exercise 2:** The following is the table which gives you the distribution of marks secured by some students in an examination:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Marks | 0-20 | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | 71-80 |
| #Students | 42 | 38 | 120 | 84 | 48 | 36 | 31 |

Find median marks.

**Exercise 3:** The table below gives diastolic blood pressure of 250 men. The readings were made to the nearest millimetre and the central value of each group is given below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Blood pressure (mm) | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 |
| #Men | 4 | 5 | 31 | 39 | 114 | 30 | 25 | 2 |

Calculate from the data the mean and the median.

**Exercise 4:** You are given the following incomplete frequency distribution. It is known that the total frequency is 1000 and that the median is 413.11. Estimate by calculation the missing frequencies.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Values | 300-325 | 325-350 | 350-375 | 375-400 | 400-425 | 425-450 | 450-475 | 475-500 |
| Frequency | 5 | 17 | 80 | ? | 326 | ? | 88 | 9 |

[**Ans.** E1: 2; E2: 40.46; E3: Mean = 79.08, Median = 79.52; E4: 227, 248]

**Mode**

The mode of a set of observations is that value which occurs with the maximum frequency. This definition, properly speaking, applies to a discrete variable only.

For a continuous variable, the above definition needs to be modified. The mode here is the ***value of the variable with the highest frequency-density*** corresponding to the ideal distribution which would be obtained if the total frequency were increased indefinitely and if, at the same time, the width of the class-intervals were decreased indefinitely. Graphically, it may be looked upon as the abscissa corresponding to the highest ordinate in the frequency curve (the limiting form of histogram).

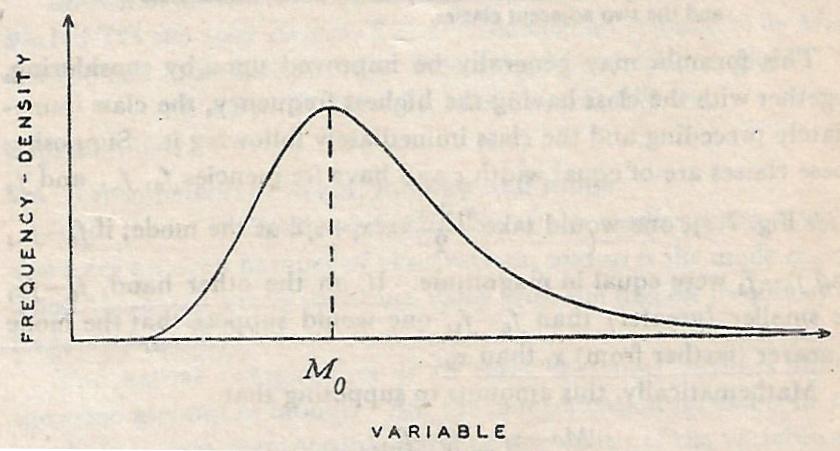


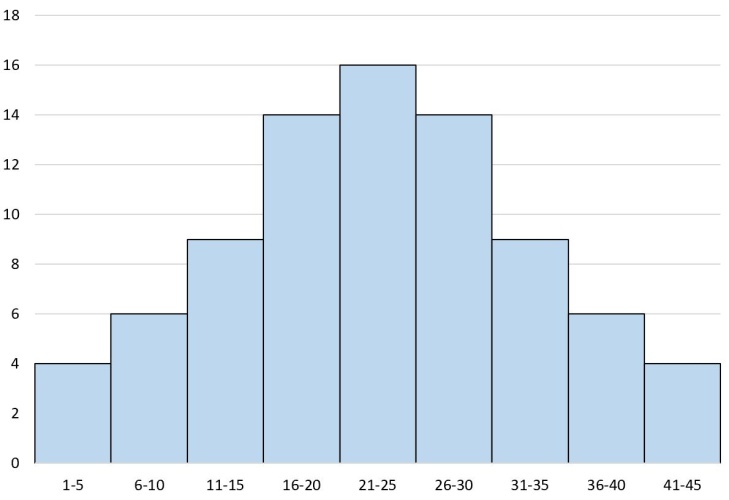
Fig. 2: Mode of a continuous variable

**Three approaches for computation of mode from grouped frequency distribution**

1. By the application of simple interpolation in a cumulative frequency distribution
2. Graphical method
3. From empirical relation

**Computation of mode by application of simple interpolation in a cumulative frequency distribution**

Suppose, the frequency distribution is perfectly symmetric as follows:



Then, **t**he mid-value of the class interval having the highest frequency (i.e. modal class) may be approximately taken to be the mode.

Thus, if the lower and upper class-boundaries of the class containing the highest frequency are and respectively, then the mode will be approximately given by

, where is the width of the interval ......................(1)

But in reality the frequency distribution will **rarely be symmetric.** Suppose, the frequency distribution is as follows:

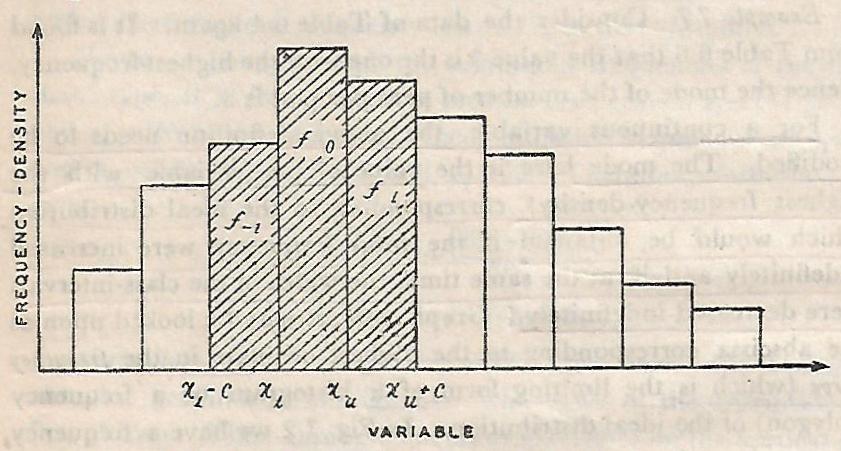


Fig. 3: Frequency distribution of a variable

In this case, one would not expect the mode to be in the middle of the modal class, i.e. . Rather, the mode will be nearer to or depending on the frequencies in the classes immediately preceding and immediately following the modal class.

Suppose the modal class, preceding class and the following class have frequencies , and respectively. Then, if (-) be smaller (greater) than (-), one would suppose that the mode is nearer (further from) than . Mathematically, this amounts to supposing that

where, is the difference of frequencies in the modal class and the preceding class and is the difference of frequencies in the modal class and the following class.

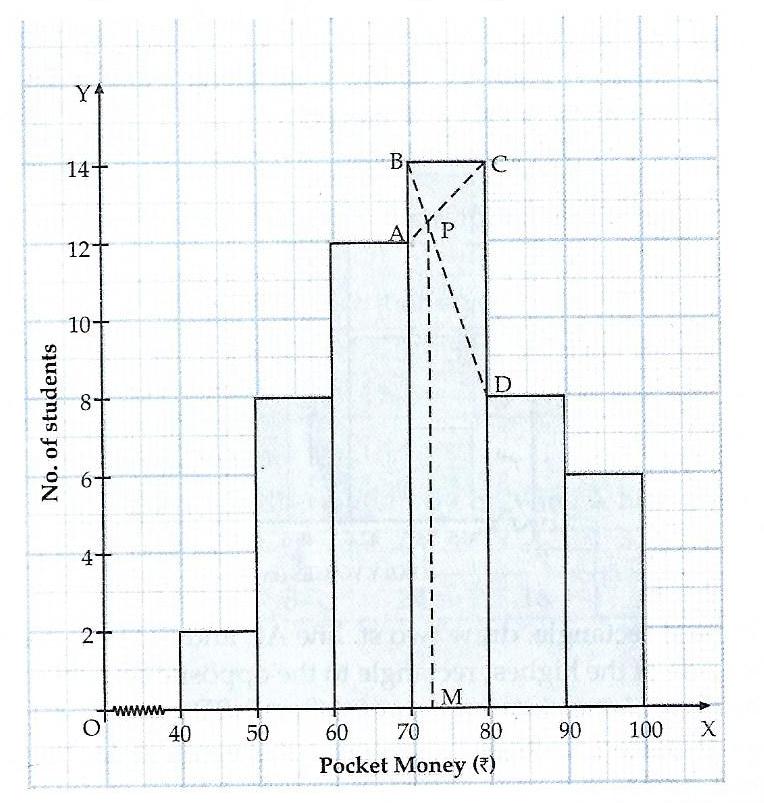
**Determination of mode using Graphical method**

The mode value may also be obtained geometrically from the histogram. Inside the highest rectangle two straight lines from the corners of the rectangles on either side of the highest rectangle to the opposite corners of the highest rectangle are drawn. The measure of mode is given by the abscissa of the point of intersection of the two straight lines.

**Example:** In a school, the weekly pocket money of 50 students is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Weekly pocket money (in Rs.) | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
| No. of students | 2 | 8 | 12 | 14 | 8 | 6 |

Draw a histogram and find the mode from the graph.



**Relation between Mean, Median and Mode**

* For unimodal distributions of moderate skewness the following approximate relation has been found to hold:
* When the distribution is symmetrical, mean, median and mode coincide, i.e.

In particular, for the ***normal distribution*** mean, median and mode are all equal.

* In most frequency distributions, it has been observed that the three measures of central tendency (mean, median and mode) obey the approximate relation provided the distribution is not very skew.

**Determination of mode using the Empirical Relation**

Another method of determining the mode is to make use of the above mentioned empirical relation, i.e. , which implies that

Mode =

Given the mean and the median, an approximate value of the mode may be obtained using the above relation.

**Advantages of mode**

* From a simple frequency distribution, mode can be obtained only by inspection. Also, for a simple series with small number of observations, mode can be determined without any calculation.
* It is unaffected by the presence of extreme values
* It can be calculated from frequency distributions with open-end classes.

**Disadvantages of mode**

* Mode has no significance unless a large number of observations are available.
* It is a peculiar measure of central tendency. When all values occur with equal frequency, there is no mode. On the other hand, if two or more values have the same maximum frequency, there is more than one mode.
* It cannot be treated algebraically
* It is quite difficult to calculate mode accurately for continuous variable
* It is less reliable and less stable as regards to sampling fluctuations

**Uses of mode:**

* Mode is most useful as a measure of central tendency ***when examining categorical data***, such as models of cars or flavours of soda, for which a mathematical average or median value based on ordering cannot be calculated.
* ***In some cases, mode is desirable than mean or median***. For example,
* A dealer of shoes will be more interested in the mode of sizes of shoes he sells than arithmetic mean or median.
* Modal wage may be considered as the representative wage of a group of workers. Modal wage is that wage which the largest number of workers receives.

**Exercise 1:** Calculate the Median and Mode from the following data:

|  |  |
| --- | --- |
| **Value** | **Frequency** |
| Less than 10 | 4 |
| Less than 20 | 16 |
| Less than 30 | 40 |
| Less than 40 | 76 |
| Less than 50 | 96 |
| Less than 60 | 112 |
| Less than 70 | 120 |
| Less than 80 | 125 |

[**Ans.** Median=36.25, Mode=34.29]