**Chapter 1**

**Probability and Statistics**

**Introduction and Descriptive Statistics**

**Introduction**

In the modern world of computers and information technology, the importance of statistics and probability is very well organized by all discipline. Statistics has originated as a science of statehood and found applications slowly and steadily in agriculture, engineering, economics, commerce, biology, medicine, industry, planning, education and so on. As on date there is no other human work of life, where statistics can not applied.

Probability and statistics are concerned with events, which occur by chance. Examples include occurrence of accidents, errors of measurements, production of defective and non defective items from a production line, and various games of chance, such as drawing a card from a well-mixed deck, flipping a coin, or throwing a symmetrical six-sided die. In each case, we have some knowledge of likelihood of some possible result, but we can not predict with certainty the outcome of any particular trial.

**Definition of statistics**

"Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data from the logical analysis." It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one.

**Descriptive Statistics**

The branch of statistics devoted to the summarization and description of the data (population or sample) is called descriptive statistics. It is used to organize and describe the characteristic of a collection of data in tabular, graphical or numerical form. Descriptive statistics has no hypothesis and does not analyze data.

**Inferential statistics**

If it may be too expensive to obtain or it may be impossible to acquire every measurement in the population, then we will want to select a sample of data from the population and use the sample to infer the nature of population. The branch of statistics which is used to draw inferences about the population of data from a sample data drawn from the population is called inferential Statistics. It is also known as sampling statistics.

**Application of Statistics in engineering**

An engineer is someone who solves problems of interest to society by the efficient application of scientific principles. Engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets customer's needs. The engineering, or scientific, method is the approach to formulating and solving these problems. The steps in the engineering method are as follows:

1. Develop a clear and concise description of the problem.

2. Identify, at least tentatively, the important factors that affect this problem or that may play a role in its solution.

3. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. State any limitations or assumptions of the model.

4. Conduct appropriate experiments and collect data to test or validate the tentative

model or conclusions made in steps 2 and 3.

5. Refine the model on the basis of the observed data.

6. Manipulate the model to assist in developing a solution to the problem.

7. Conduct an appropriate experiment to confirm that the proposed solution to the problem is both effective and efficient.

8. Draw conclusions or make recommendations based on the problem solution.

**Functions of statistics**

1. Statistics simplifies complexity
2. Statistics present fact in a definite form
3. Statistics facilitates comparisons
4. To help in formulations of policies
5. Statistics helps in forecasting
6. Statistics helps in formulating and testing hypothesis

**Limitation of Statistics**

1. Statistics does not deal with individuals
2. Statistics does not study qualitative phenomena
3. Statistical laws are not exact
4. Statistics is only a means
5. Statistics is liable to be misused

**Frequency Distribution**

Frequency distribution is simply a table in which the collected and classified data are presented in to different classes and the numbers of cases which fall in each class are recorded. There are two kinds of frequency distribution. Univariate and Bivariate frequency distribution. Univariate frequency distribution may be of the following three types.

Individual series

Individual series is a series where items are listed singly after observations as distinguished from listing them in groups. If the temperature of a week in a particular month is given individually, it will form individual series.

Example of Individual series

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Day | S | M | T | W | Th | F | Sat |
| Temp(0c) | 32 | 31 | 35 | 26 | 25 | 30 | 25 |

**Discrete series**

The series formed from a discrete variable is known as discrete series. It is also known as discrete frequency distribution. In a discrete series the data are presented in away that exact measurements of units are clearly indicated.

Example of Discrete series

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No of goals in a match | 0 | 1 | 2 | 3 | 4 |
| No of match | 27 | 9 | 8 | 5 | 2 |

**Continuous series**

The series formed from a continuous variable is termed as a "continuous series". It is also known as "continuous frequency distribution". In continuous series the data or variate values always lie between two numbers known as interval.

Example

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Wages (in Rs.) | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
| No of workers | 10 | 12 | 15 | 7 | 6 |

**Classification according to class interval**

To classify the data according to class-interval the following terms are important to define:

1. Class: different groups in to which the data can be classified are known as classes or class intervals.
2. Class limits: The two values specifying the class are called the class limit. The smaller value is said to be lower limit and the greater value the upper limit. In the class 30-40, 30 is lower limit and 40, the upper limit.
3. Class-size: The difference between the upper and lower limit of a class is called magnitude of the class or class size. In the class 30-40, the class size = 40-30=10.
4. Class frequency: The number of observations falling within a particular class is known as the class frequency. In the given in continuous series, In the example given in continuous series 10is the frequency of the class 30-40. That is, there are 10 persons whose wage lies in between 30-40.
5. Mid-value of the class: it is the value lying in the half-way between two extreme limits of a class. It is also known as mid-point or class mark. It is determined by dividing the sum of the lower and upper limit by two. The mid value of the class interval 30-40 is =35.

**Method of Forming class interval**

There are two ways of forming class-interval:

* Exclusive class
* Inclusive class

**Exclusive class**

The formation of class-interval by this method is that the upper limit of one class is the lower limit of next class so as to make continuous without any gap. This type of method is mostly useful in case of continuous variable. Since rainfall is continuous variable so for such variable, we form classes by this method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rainfall( in mm): | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
| No of days: | 12 | 7 | 10 | 4 | 2 |

**Inclusive Method**

The formation of the class interval by this method is that both lower and upper limits are included in a particular class. This method is mostly used in case of discrete variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rainfall(in mm) | 10-19 | 20-29 | 30-39 | 40-59 | 60-69 |
| No of days: | 12 | 7 | 10 | 4 | 2 |

**Method of changing Inclusive class in to Inclusive class**

When the class intervals are formed by "Inclusive method" there is a gap between the upper limit of any class and the lower limit of the succeeding class. To change the classes of "Inclusive type" in to that of "Exclusive type" we adjust and get the real limits of the classes by using the correction factor. The correction factor denoted by Cf is given by

Cf =

Then the real upper limit = Upper limit + Cf

And the real lower limit = lower limit - Cf

The real upper limit and lower limits are said to be the class boundaries.

**Pictorial Representation of Data (Pie-Chart, Histogram and Ogive Curves)**

**Pie Diagram (Circular diagram or Angular Diagram)**

Pie Diagram is used for depicting the components of a single factor. In such diagram both the total and the component parts or sectors can be shown. The area of a circle is proportional to the square of its radius. It is divided in to different sectors by radial lines such that the area of each of the sectors representing the component value of the total value.

**Procedure of construction of Pie chart**

1. Compare the total value of a variable to 3600. Then find the angle corresponding to the component value of the total ie

Angle corresponding to component value = (3600 /total value)\* given value

1. Draw a circle of appropriate radius. For a single pie chart the radius may be any value, which makes the chart attractive. However, two or more pie charts at a time the radii of circles of the pie charts must be proportional to the square root of their total value, ie

**r1 : r2**

3. Take any radius as base line and draw and angle represented by the first component at centre using protractor. Similarly draw other sectors to represent remaining component values.

4. Shade each sector differently by lines, dots or with different colors

**Limitations of Circular Diagram**

1. Negative data (profit or loss) can not be shown in pie chart.
2. If there are more than six component pie diagram is not suitable
3. Pie diagram are less effective than other (i.e.bar ) due to difficulty in comparing the areas of the sectors

Example 1. Express the following using Angular diagram.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | Food | Clothing | Housing | Fuel | Education | Miscellaneous |
| %Expenditure | 50 | 15 | 10 | 5 | 10 | 10 |

Solution: Angle at the centre of circle = 3600, total value = 3600

Let x = ((360/100)\*x)0 =(3.6 \* x)0

|  |  |  |
| --- | --- | --- |
| Item | % of expenditure | Values in angle = (3.6\*x) |
| Food | 50 | 1800 |
| Clothing | 15 | 540 |
| Housing | 10 | 360 |
| Fuel | 5 | 180 |
| Education | 10 | 360 |
| Miscellaneous | 10 | 360 |
| Total | 100 | 3600 |

Example 2.The number of B.E. students in two colleges A and B are given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | No of students | | | |
| College | Civil | Computer | Electronics | Industrial |
| A | 40 | 60 | 50 | 30 |
| B | 60 | 80 | 50 | 50 |

Compare the student's number with the help of angular diagram.

Solution: Number of students in college A is 180. The angle at the centre of circle is 3600.

Total Value = 3600 and 180 % = 360, Then x= ((360/180)\*x)0 = (2\* x)0

Number of students in college B =240 then x = ((360/240)\*x)0 = (1.5\*x)0 [ 240 % =3600]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| For college A | | | | For college B | | | |
| Program | No of students | (in0) | Cum0 | Program | No of students | (in0) | Cum0 |
| Civil | 40 | 800 | 800 | Civil | 60 | 90 | 90 |
| Computer | 60 | 1200 | 2000 | Computer | 80 | 120 | 210 |
| Electronics | 50 | 1000 | 3000 | Electronics | 50 | 75 | 285 |
| Industrial | 30 | 600 | 3600 | Industrial | 50 | 75 | 360 |
| Total | 180 |  | | | 240 |  | |
| Square root | 13.41 | 15.49 |  | |
| Ratio of circle | 1 | 1.16 |  | |

=

Pie chart of the BE students in two colleges A and B

**Histogram**

Histogram is a bar diagram which is suitable for frequency distribution with continuous classes. The width of bars is equal to class interval and heights of bars are proportional to the frequency of the respective classes. The bars touch each other but bar never overlap the other. A line bar diagram is also known as frequency bar diagram.

**Remarks**

1. If only mid points are given, we evaluate, upper and lower limits of different classes the draw histogram as in the above cases.
2. The inclusive frequency must be changed in to exclusive type.
3. We can not construct a histogram for distribution with open-end classes.
4. It is also misleading if the distribution has unequal intervals and suitable adjustment in frequencies are not made.

Example 1

Draw histogram for the following data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Marks: | 0-4 | 4-8 | 8-12 | 12-16 | 16-20 |
| No of Students: | 4 | 6 | 10 | 8 | 4 |

Figure 1

Example 2

Represent the following data by means of Histogram:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Weekly wages(in Rs): | 10-15 | 15-20 | 20-25 | 25-30 | 30-40 | 40-60 | 60-80 |
| No of workers: | 7 | 20 | 27 | 15 | 24 | 20 | 8 |

Solution

This is the case of unequal class- interval, so adjustment of frequencies must be made. Minimum class size is 5. Since the magnitude of class interval 30-40 is double of 10-15, so the frequencies 30-40 must be divided by 2. Similarly, the frequencies of 40-60 and 60-80 must be divided by 4.

Figure 2

**Cumulative frequency curve or ogive**

Cumulative frequency curve which is also known as " Ogive" is a graphical representations of cumulative frequency distribution of a continuous variable. In drawing Ogive , points are plotted with cumulative frequency(c.f.)along the y-axis and the corresponding class boundaries along the x-axis and joining them freely. There are two Ogives

1. Less than Ogive
2. More than Ogive

**Less than ogive**

First prepare less than cumulative frequencies in which case the frequencies are added serially from top to bottom. Then taking upper limits of the class intervals as the x-coordinates and the corresponding less than cumulative frequencies as the y- coordinates, points are plotted. The points thus obtained, joined freely will give the less than Ogive. Obviously, "less than Ogive" is an increasing curve sloping upward from left to right and has the slope of an elongated S.

**More than Ogive**

First prepare more than cumulative frequencies in which case frequencies are added serially from bottom to the top. Then taking lower limits of the class-intervals as the x- coordinates and the corresponding more than cumulative frequencies as the y-coordinates, points are plotted. The points thus obtained are joined freely will give the "more than ogive". It is a decreasing curve and slopes downwards from left to right and the shape of the elongated S upward down.

One important thing is to be remembered is that unless and until it is stated, Ogive or cumulative frequency curve means the less thanogive.

**Utility of Ogives**

1. Ogives are useful for graphical computation of partition values such as quartiles, deciles etc.
2. Ogive can also be used to have graphically the number of observations below or above a given value of the variable or lying between certain intervals of the variable.
3. The comparative study of two or more distributions can betterly be studied with the help of Ogive as different Ogives relating to different distributions can be drawn on the same paper.

**Example**

Draw a less than Ogive from the following data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Income (in Rs.): | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
| No of persons: | 5 | 10 | 18 | 23 | 7 | 6 |

1. Find the number of persons having income s less than equal to Rs. 35
2. Find the median

Solution

First we prepare the less than cumulative frequencies which are as follows

|  |  |  |
| --- | --- | --- |
| Income (In Rs.) | Frequencies | Less tha C.f. |
| 0-10 | 5 | 5 |
| 10-20 | 10 | 5+10=15 |
| 20-30 | 18 | 15+18=33 |
| 30-40 | 23 | 33+23=56 |
| 40-50 | 8 | 56+8=64 |
| 50-60 | 6 | 64+6= 70 |

Now points are plotted with less than c.f. against upper limits of the class intervals

Figure

1. and b) are left for students as home work.

Example

Following is the distribution of marks in management obtained by 50 students

|  |  |
| --- | --- |
| Marks | No of students |
| More than 10 | 50 |
| More than 20 | 43 |
| More than 30 | 32 |
| More than 40 | 18 |
| More than 50 | 8 |
| More than 60 | 5 |

Solution : More than cumulative frequencies have already been given. So we plot the more than c.f. against the lower limit of the class intervals.

Figure

Example:

Draw both "less than" and "more than" Ogives from the following data and compute the value of the median.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rain fall(mm.): | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
| No of Days: | 4 | 6 | 10 | 20 | 18 | 2 |

Solution: First we prepare less than and more than cumulative frequencies

|  |  |  |  |
| --- | --- | --- | --- |
| For less than Ogive | | For more than Ogive | |
| Rainfall | Less than c.f. | Rainfall | More than c.f. |
| Less than 20 | 4 | More than 10 | 60 |
| Less than 30 | 10 | More than 20 | 56 |
| Less than 40 | 20 | More than 30 | 50 |
| Less than 50 | 40 | More than 40 | 40 |
| Less than 60 | 58 | More than 50 | 20 |
| Less than 70 | 60 | More than 60 | 2 |

Figure

Finding the median value is left for students for home work

**1.3 Measures of location (Measures of central Tendency)**

The averages are the measures which condense a huge mass of data in to single value representing whole data. Averages are the typical values around which the most of the data tend to cluster. These are the values which lie between two extreme observations of the entire data and give us the idea about the concentration of the values in the central part of the distribution. A measure of such single values is known as measures of central tendency.

**Objects of central tendency**

The objects of central tendency are

1. To facilitates the comparisons
2. To present the salient features of a mass of complex values
3. To know about universe from a sample
4. To trace mathematical relation
5. To help in decision making

**Requisite of a good average**

To be an ideal average, the following characteristics should be satisfied.

1. It should be rigidly defined and its value should be definite.
2. It should be simple to understand.
3. It should be based on all the observations.
4. It should be easy to calculate.
5. It should suitable for further mathematical treatment.
6. It should be least affected by fluctuation of sample.
7. It should not be affected by extreme observations

**Types of Averages**

The measure of central tendency is designed to measure central value around which most of the data tend to concentrate. The following are the measures of central tendency or measures of location:

**Arithmetic mean**

Arithmetic mean or simply a 'mean' of a set of observations is the sum of all the observations divided by the number of observations. Arithmetic mean is also known as the arithmetic average:.

1 Individual series

Let x1,x2, x3,………,xn  be the n values of the variable then their arithmetic mean denoted by is defined by

=

= ……………..(1)

Where n is the number of observations. This method of finding A.M. is known as 'Direct method'.

**Short-cut method**

When the number of observations bee too large and when the size of variate values be very big, we find their arithmetic mean by taking the deviations of the items from any arbitrary. This method is known as **short-cut method.**

When the deviations are taken from any arbitrary number known as assumed mean the formula (1) takes the following form:

= ……………..(2)

Where a = assumed mean, = sum of the deviations of the items taken from the assumed mean 'a', n = number of observations. d = x-a.

**2 Discrete series**

Let f1, f2, f3, ……………fn be the frequencies of the variate values x1,x2,x3…….., xn respectively, then their arithmetic mean denoted by is defined by

=

= = ……………….. (3)

Where N = = total frequency

This method of finding arithmetic mean is known as "Direct method".

**Short- cut method**

The formula for arithmetic average in the short cut method is of the following form

= ………………(4)

Where a = assumed mean,

d = x – a = deviation of the item taken from the assumed mean 'a',

N=total frequencies.

**Continuous series**

The same formula (3) can be used to find the arithmetic average in case of continuous frequency distribution, the only difference in continuous frequency distribution is that x is the middle value of the corresponding class.

Step-deviation method

The formula for the arithmetic mean in step deviation method can be put in the following form

= a +…………………… (5)

Where a = assumed mean, N = = Total frequency d' = and h = class size

**Example**: The expenditure of 6 families in rupees are given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Family | A | B | C | D | E | F |
| Expenditure (in Rs.) | 300 | 325 | 375 | 400 | 425 | 500 |

Calculate the arithmetic mean by (a) direct method (b)Short- cut method

**Solution**

**(a)**

Calculation of Arithmetic mean

|  |  |
| --- | --- |
| Family | (Expenditure in Rs.) |
| A | 300 |
| B | 325 |
| C | 375 |
| D | 400 |
| E | 425 |
| F | 500 |
|  | =2325 |

Here =2325, n=6 = ?

Now,= = = 387.5

Monthly average expenditure = Rs. 387.5

Example

Find the arithmetic mean from the data given below by using by using (a) direct method (b) Step deviation method.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class boundaries | 15-25 | 25-35 | 35-45 | 45-55 | 55-65 | 65-75 |
| Frequency | 4 | 11 | 19 | 14 | 0 | 2 |

**Solution:**

1. By direct method

|  |  |  |  |
| --- | --- | --- | --- |
| Class boundaries | Mid value(x) | Frequency(f) | fx |
| 15-25 | 20 | 4 | 80 |
| 25-35 | 30 | 11 | 330 |
| 35-45 | 40 | 19 | 760 |
| 45-55 | 50 | 14 | 700 |
| 55-65 | 60 | 0 | 0 |
| 65-75 | 70 | 2 | 140 |
|  |  | N= 50 | = 2010 |

Here

= 2010 N = 50 =?

= = 40.2

1. Step deviation method

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class boundaries | Mid value  (x) | Frequency(f) | d=x-40 | = | f |
| 15-25 | 20 | 4 | -20 | -2 | -8 |
| 25-35 | 30 | 11 | -10 | -1 | -11 |
| 35-45 | 40 | 19 | 0 | 0 | 0 |
| 45-55 | 50 | 14 | 10 | 1 | 14 |
| 55-65 | 60 | 0 | 20 | 2 | 0 |
| 65-75 | 70 | 2 | 30 | 3 | 6 |
|  |  | N=50 |  |  | = 1 |

Here a=40 = 1 N =50 h= 10 =?

= a + \*h = 40 +\*10 = 40.2

Discrete case: Class discussion

Home work

Compute the arithmetic mean of the following frequency distribution by (a) direct method (b) Short cut method

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Heights in cm (x) | 10 | 20 | 30 | 40 | 50 | 60 |
| No of plants (f) | 2 | 3 | 9 | 21 | 11 | 5 |

Weighted Arithmetic mean

In the simple arithmetic mean all the items are assumed equally important in the distribution. But in practice this may not be so, the importance of some items may be greater than the other. So in such cases proper weighted should be given to various items. Now we define the following weighted arithmetic average in which proper weight is considered.

Let w1, w2, w3,…………., wn be the weights given to the variate values x1, x2 ,x3, …………………..xn respectively then their weighted arithmetic mean denoted by w = = ………………(6)

Example

Computed the weighted arithmetic mean of the index number from the data given below.

|  |  |  |
| --- | --- | --- |
| Group | Index no | weights |
| Food | 125 | 7 |
| Clothing | 133 | 5 |
| Fuel and light | 141 | 4 |
| House rent | 173 | 1 |
| Misc. | 182 | 3 |

Solution

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Index no (x) | Weights (w) | w.x. |
| Food | 125 | 7 | 875 |
| Clothing | 133 | 5 | 665 |  |
| Fuel and light | 141 | 4 | 564 |
| House rent | 173 | 1 | 173 |
| Misc. | 182 | 3 | 546 |
|  |  | = 20 | =2823 |  |

Here = 20 =2823 w = ?

Now w = = =141.5

Properties of Arithmetic Mean

1. The algebraic sum of the deviations of the items taken from their arithmetic mean is zero.
2. The sum of the squares of the deviations of the items is minimum when taken from the arithmetic mean.
3. Combined mean: If n1 and n2 be the sizes and 1 and 2 be the arithmetic means of two component series, then the means 12 of the combined series of size n1 + n2  is given by 12 =
4. Since = so = n so it is clear that any two given values will enable us to compute the remaining value.

Example The mean height of 35 male workers in a certain factory is 162 cms. and the mean height of 25 female workers in the same factory is 158 cms. Find the combined mean height of 60 workers in the factory.

Solution:

Left for students as a home work

Merits and Demerits Of Arithmetic means.

Merits:

1. It is rigidly defined.
2. It is based on all observations.
3. It is simple to understand and easy to calculate
4. It is suitable for further mathematical treatments.
5. It is least affected by fluctuation of sampling.

Demerits:

1. It is very much affected by fluctuation of sampling.
2. It can be computed accurately in case of open end classes.
3. It gives some times fallacious conclusion.
4. It can not be determined by inspection or by graphical method.
5. It can not be used if we are dealing with qualitative characteristics which can not be measured quantitatively.

**Median**

The variate values dividing the total number of observations into two equal parts is said to be the median. The median divides the whole observations in to two equal halves, one half comprising all the values greater and the second half, all the values less than median. It is the positional average i.e. its value depends on the position occupied by a value in the frequency distributions. It is denoted by Md. Determination of Median depends upon the series given.

**Individual series**

First we arrange the set of observations according to their ascending or descending order. If the number of observation is odd, the middle value gives the median. Again if the number of observation is even, there will be two middle values, so the arithmetic average of two middle values gives the median. The formula for computing the median in case of individual series is

Median = value of ()th  item ……………….(1)

Where n is the no of observations.

Example:

Find the median from the following set of observations:

5, 2, 3, 4, 10, 7, 12

Solution

Arranging the given set of observations according to their ascending order, we have 2, 3, 4, 5, 7, 10,12

N = no. of observation = 7

Median = value of ()th  item

=Value of ()th  item

= value of 4th item = 5

Example

Find the median wage from the following set of monthly wages (in Rs.); 900, 1000, 650, 1200, 860, 920

Solution

Arranging the set of monthly wages according to ascending order

We have

650, 860, 900, 920, 1000, 1200

n = number of observation = 6

Median = value of ()th  item

= value of ()th item

= value of (3.5)th  item =

= = 910 Median wage = Rs 910

Discrete series:

In finding the median in case of discrete series, we use the following procedure:

1. Arrange the data according to their ascending order of their magnitudes.
2. Form the cumulative frequency table
3. Use the formula median = value of ()th item…………….(2)
4. Now see the cumulative frequency column and note the value corresponding to the cumulative frequency either equal to or greater than. This gives value of median.

Example

Find the median from the following data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Monthly saving (Rs.) | 250 | 300 | 450 | 600 | 1000 |
| No of families | 25 | 16 | 20 | 24 | 15 |

Solution:

|  |  |  |
| --- | --- | --- |
| Monthly saving | Frequency | Cumulative frequency |
| 250 | 25 | 25 |
| 300 | 16 | 25+16=41 |
| 450 | 20 | 41+20=61 |
| 600 | 24 | 61+24=85 |
| 1000 | 15 | 85+15=100 |

Here N= total frequency=100

Median = value of ()th item = value of ( )th  item = value of (50.5)th item

= 450

Monthly median wage = Rs 450

Continuous series

The following steps are to be used in finding the median in case of continuous series:

1. Prepare less than cumulative frequency distribution.
2. Find .
3. See cumulative frequency equal to or just greater than N/2 and note.
4. The corresponding class contains the median value and is called the median class.
5. The value of the median is now computed by the formula

Median = L + ……………….(3)

Where L = lower limit of the median class

N = total frequency

c.f. =cumulative frequency preceding the median class

f = frequency of the median class

h = size of the median

Example

Find the median from the following distribution:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Consumption of Electricity(units) | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
| No of families | 10 | 20 | 10 | 25 | 10 | 50 | 40 | 30 |

Solution

|  |  |  |
| --- | --- | --- |
| Consumption of electricity | Frequency | Cumulative frequency(c.f) |
| 20-30 | 10 | 10 |
| 30-40 | 20 | 30 |
| 40-50 | 10 | 40 |
| 50-60 | 25 | 65 |
| 60-70 | 10 | 75 |
| 70-80 | 50 | 125 |
| 80-90 | 40 | 165 |
| 90-100 | 30 | 195 |

Here = = 97.5 Median lies in the class 70-80

L =70 =97.5 c.f.=75 f = 50 h = 10 Median =?

Now Median = L + = 70 +

Median = 70+ 4.5 = 74.5 median = 74.50 units

**Merits and Demerits of Median**

**Merits**

1. Median is rigidly defined.
2. It is simple to understand and easy to calculate.
3. Median is not affected by extreme observation.
4. Median can be computed even for open- end classes.
5. Median can sometimes located by inspection
6. Median can be obtained graphically.
7. Median is only the average to be used while dealing with qualitative characteristics such as intelligence, beauty etc.

Demerits

1. An arrangement of data according to magnitude is necessary.
2. Median is not based on all observations.
3. For an ungrouped data, if the number of observation is even, median can not be determined exactly.
4. Median is not suitable for further mathematical treatment.
5. For a small size sample, median is affected by fluctuation of sampling.

Mode

Mode is that variate value which repeats maximum number times. That is, it is the variate value which occurs most frequently in the set of observations and around which the other items of the set concentrate densely.

Computation of mode of continuous series

In case of individual and discrete series, the mode can easily be found out by inspection. But in case of continuous series we use the following formula to calculate the mode.

Mo = Mode = L +

Where, L = lower limit of the modal class

,

= frequency preceding the modal class

= maximum frequency

Frequency following the modal class

h = size of the modal class Mo = Mode

Example Calculate the mode of the following set of observations:

2, 5, 7, 5, 3, 5, 4, 5, 2

Solution

|  |  |
| --- | --- |
| Value | No of Occurrence |
| 2 | 2 |
| 3 | 1 |
| 4 | 1 |
| 5 | 4 |
| 7 | 1 |

Since the value 5 repeats maximum number of times i.e. 4 times so mode =5

Example Calculate the mode for the distribution of weights of 150 students from the data given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wts.(in K. g.) | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 |
| Frequency | 18 | 37 | 45 | 27 | 15 | 8 |

Solution Since the distribution is regular and 45 is maximum frequency, 50-60 is the modal class.

Hence L =50 f1 = 45 f0 = 37 f2 = 27 h= 10 Mo= ?

f1 - f0 = 45-37 = 8 = f1 – f2 =45- 27 =18

Now

Mode = Mo = L + = 50 + = 50+3.07 =53.07

While applying the above formula for mode, it is necessary to see that the class intervals are uniform throughout. If they are not uniform, they must first be made equal on the assumption that the frequencies are equally distributed throughout the class.

If there are irregularities in the distribution i.e. if the highest values occurs in the very beginning or very last or if the highest frequency occurs in the middle and is not supported by relatively small items, then mode will be calculated by method of grouping.

Method of grouping

Example: Determine the mode of the following data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X: | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| F: | 6 | 9 | 4 | 2 | 10 | 8 | 7 | 5 | 1 | 3 |

Solution

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Size of item | Col I | Col lI | Col III | Col IV | Col V | Col VI |
| 20 | 6 | 15 |  | 19 |  |  |
| 21 | 9 | 13 | 15 |  |
| 22 | 4 | 6 | 16 |
| 23 | 2 | 12 | 20 |
| 24 | 10 | 18 | 25 |
| 25 | 8 | 15 | 20 |
| 26 | 7 | 12 | 13 |
| 27 | 5 | 6 | 9 |
| 28 | 1 | 4 |  |
| 29 | 3 |  |  |  |

Analysis table

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| I |  |  |  |  | 1 |  |  |  |  |  |
| II |  | 1 | 1 |  |  |
| III |  |  | 1 | 1 |  |
| IV | 1 | 1 | 1 |  |  |
| V |  | 1 | 1 | 1 |  |
| VI |  |  | 1 | 1 | 1 |
| Total |  |  |  | 1 | 4 | 5 | 3 | 1 |  |  |

Since the variate value repeats maximum number of times i.e. 5 times , so mode =25

Computation of mode using empirical relation

A distribution with only one mode is said to be unimodal. If a distribution has two equal maximum frequencies, then the distribution is said to be bimodal distribution. If there are three or more equal maximum frequencies, then the distribution is said to be multimodal distribution. In this situation mode is said to be ill defined and can be calculated by the following empirical relation. Mode =3 Median – 2 mean

Merits and demerits of mode

Merits:

1. It is easy to calculate and simple to understand
2. Mode is not at all affected by extreme observations.
3. It can be obtained even in case of open end classes.
4. It can be obtained by inspection or by graph.

Demerits

1. Mode is not rigidly defined.
2. It is not based on all the observations.
3. Mode is not suitable for further mathematical treatment.
4. Mode is affected to a greater extent by fluctuation of sampling.

**Partition values**

Partition values are those variate values which divide the total no of observation in to equal number of parts. The equal number of parts may be of four, ten or hundred etc.

Variate values dividing the total number of observations in to 4 equal parts are known as "Quartiles". Median divides the total number of observations in two equal halves: lower half and upper half.

Deciles are those variate value which divide the total number of observations in to 10 equal parts. So, there are 9 deciles. They are denoted by D1, D2, D3……….., , D9 such that D1<D2, <D3<……….., <D9.

The variate values dividing the total number observations in to 100 equal parts are known as percentiles. There are 99 percentiles P1, P2, P3………….., P99 such that P1 < P2 < P3<…………..,< P99.

**Computation of partition values**

The procedure for computing quartiles, deciles, and percentiles in case of individual, discrete and continuous series are same as in case of median.

**Individual series**

If n be the number of observations arranged according to their ascending order then the quartiles, deciles and percentiles can be computed by the following formulae:

Q1 =Value of item, Q3 =Value of item

D2 =Value of item D3 =Value of item

P3 =Value of item P4 =Value of item

Discrete series

The formulae for quartiles, Deciles and percentiles in case of discrete series are same as in individual series; the only difference is that n should be replaced by N the total frequency.

Continuous series

The following formulae are used to find the quartiles, deciles and [ercentiles in case of continuous series.

Q1 = L + Q3 = L +

D6 = L + D3 = L +

P3 = L + P4 = L +

Where

L= lower limit of the class in which the particular partition value lies.

N = Total frequency

c.f. = cumulative frequency preceding the class in which the particular partition values lies.

H = class size

In particular Q1 = P25 Md = Q2 = D5 = P50 and Q3 = P75

All the solved examples and exercise problems are left for the students as homework.

**Measures of variability (Measure of dispersion)**

**Introduction and meaning**

Averages gives us the idea of concentration of the idea of concentration of the items around the central part of the distribution. But the averages only do not give the clear pictures about the distribution because two distributions with same averages may differ in the scatterness of the items from the central value. For this we have the following example:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | Mean | Mode | Median |
| Series A | 25 | 26 | 27 | 27 | 27 | 28 | 29 | 27 | 27 | 27 |
| Series B | 0 | 10 | 18 | 27 | 27 | 27 | 80 | 27 | 27 | 27 |

From the above table, we see that mean, mode and median of the two series A and B are same. Only with these results we can not say that the two series A and B are similar. Because, the difference of the items from the averages in B is more in comparison to A. So in series A, items are concentrating more around the central value but the scatterness of the items from central value in series B is more. Hence, though two series A and B A and B have same average, they can not be said similar because they may differently constituted.

Actually the meaning of dispersion is the scatterness of the items from the central value. So dispersion is defined as the measure of the variation in the items from the central value.

The main objects of measuring variability are as follows:

1. To determine the reliability of an average.
2. In devising a system of quality control.
3. To compare two or more series with regard to their variability
4. To help in using other statistical tools.

Requisite of a good measure of dispersion

The following are the necessary characteristics to be satisfied by different measures of dispersion to become a good measure.

1. It should be rigidly defined and its value should be definite.
2. It should be simple to understand.
3. It should be based on all the observations.
4. It should be easy to calculate.
5. It should suitable for further mathematical treatment.
6. It should be least affected by fluctuation of sample.
7. It should not be affected by extreme observations

Absolute and relative measures

Those measure of dispersion whose units are same as the units of the given series are known as the absolute measure of dispersion. These types of dispersion can be used only in comparing the variability of the series (or distribution) having same units.Comparasion of two distributions with different units can not be made with absolute measures. On the other hand, the relative measures of dispersion are obtained as the ratio of absolute measure of dispersion to suitable average and are thus a pure number of independent of units. Two distributions with different units can be compared with the help of relative measures of dispersion.

**Method of measuring dispersion**

The following are the methods of measuring dispersions:

1. Range
2. Semi interquartile range or quartile deviation
3. Mean deviation or Average deviation
4. Standard deviation
5. Lorenz's curve

**Range**

Range is defined as the difference between the largest item and the smallest item in the set of observation. So in a set of observation if L is the is the largest item and S is the smallest item, then the range is defined by Range = L-S

In a grouped frequency distribution, range is the difference between the upper limit of the largest class and lower limit of the smallest class.

Range is the absolute measure of dispersion. It can not be used to compare two distributions with different units. So, the relative measure corresponding to the range known as coefficient of range is defined by

Coefficient of range =

Example

Find the range from the following data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Day: | Sun | Mon | Tue | Wed | Th | Fr | Sat |
| Temp(0c) | 32 | 34 | 31 | 25 | 21 | 35 | 36 |

Solution

L = Largest item =36 S = Smallest item =21

= 36-21 = 150C.

**Merits and Demerits of Range**

Merits

1. It is rigidly defined.
2. Range is simple to understand and easy to calculate.
3. Only minimum time is required to know the variability with the help of range.

Demerits

1. It is not based on all the observations.
2. Range is affected by extreme by fluctuation of sampling.
3. Range is affected by extreme Values.
4. Range can not be calculated in case of open end classes.
5. Range is not suitable for further mathematical treatments.

Semi- interquartile range or Quartile deviation

The measure of dispersion depending upon the lower and upper quartiles is known as the 'Quartile deviation'. The difference between the upper and lower quartiles is known as "interquartile range". Half the interquartile range is known as known as semi- interquartile range or quartile deviation.

Quartile deviation =

Q. D =

Less the quartile deviation, more will be uniformity or less will be the variability in the central 50% of the items. Again greater the Q.D. less will be the uniformity or greater will be the variability in the central 50% of the items, Q. D is absolute measure of dispersion.

The relative measure based on the lower and upper quartiles known as coefficient of quartile deviation is given by

Coeff. Of Q.D. =

Example

From the following table giving the heights of students, calculate the quartile deviation and the coefficient of Q.D.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Height(c.m.) | 153 | 155 | 157 | 159 | 161 | 163 |
| No of students | 25 | 21 | 28 | 20 | 18 | 24 |

Solution

Class discussion: Ans (3, 0.02)

**Merits and demerits of Q.D**.:

**Merits**

1. It is rigidly defined.
2. It is simple to understand and easy to calculate.
3. It is better measure of dispersion in comparison to range as it is based on 50% of central items.
4. It is not affected by extreme values.
5. It can be calculated even when end classes are open.

Demerits

1. It is not based on all the observations.
2. Q.D is affected by fluctuation of sampling
3. Q.D. is not suitable for further mathematical treatments.

**Mean deviation (Average Deviation)**

Mean deviation is defined as the arithmetic mean of the deviation of the items from mean, median or mode, when all deviations are considered positive.

**Individual Series**

If , Md , Mo be the arithmetic mean, median and mode of variate values x, then the mean deviations (M.D.) are computed by the following formulae:

Mean deviation from mean = =

Where read as modulus of d or x- or the absolute value of the deviations taken from mean (ignoring the negative sign) , n= no of observations.

Mean deviation from median and mode can similarly be defined by replacing by Md or Mo in the deviation.

**Discrete series**

Mean deviation in the discrete series is computed by the formula

M.D. from mean = where N = total frequency

M. D. from median and mode can similarly be obtained by replacing by Md  and Mo  respectively.

**Continuous series**

The formula for finding the mean deviations in continuous series are same as in case of discrete series, the only difference in this case is that x is mid values of the classes

There is an advantage of finding the mean deviations from median as the sum of the deviations of the items from median is least when signs are ignored. However in practice mean deviation from the mean is mostly used.

The relative measure of mean deviation is defined as follows:

Coefficient of mean deviation from mean =

Coefficient of mean deviation from Median =

Coefficient of mean deviation from Mode =

**Example** Calculate the mean deviation from the mean for the following frequency distribution.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wages (in Rs.) | 0-4 | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 |
| No of persons: | 7 | 7 | 10 | 15 | 7 | 6 |

**Solution**

**Computation of mean deviation from mean**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wages | Mid value (x) | f | d' = (x -14)/4 | fd' |  | f |
| 0-4 | 2 | 7 | -3 | -21 | 10 | 70 |
| 4-8 | 6 | 7 | -2 | -14 | 6 | 42 |
| 8-12 | 10 | 10 | -1 | -10 | 2 | 20 |
| 12-16 | 14 | 15 | 0 | 0 | 2 | 30 |
| 16-20 | 18 | 7 | 1 | 7 | 6 | 42 |
| 20-24 | 22 | 6 | 2 | 12 | 10 | 60 |
|  |  | 52 |  | -26 |  | 264 |

Here, a=14, N=52, = -26 h =4 = ?

Now = a + = 14 + = 12

For Mean deviation: N= 52 = 264 M.D.from mean = ?

Now, Mean Deviation from Mean = = = 5.08

Mean Deviation from Mean = 5.08

**Merits and Demerits of M.D**.

**Merits**

1. It is easy to understand and calculate.
2. It is based on all the observation.
3. It is less affected by extreme values in comparison to Standard deviation.
4. As the deviations are taken from the central values, so comparison of two distributions about their formation can easily be made.

**Demerits**

1. The greatest drawback of mean deviation is that the algebraic signs are ignored.
2. It is not capable of algebraic treatments.
3. It can not be computed in case of open end classes.
4. It is affected by fluctuation of sampling.
5. It does not give satisfactory result when deviations are taken from mode as mode is ill-defined.

**Standard deviation**

Standard deviation (s.d.) is defined as the positive square root of the mean of the square of the deviations taken from the arithmetic mean. It is denoted by.

**Individual series**

If x be the variate values and , their arithmetic mean , then the s.d. is given by

=

Where n = no of observations.

In short cut method, s.d. is computed by the formula

Where d = x - a, a = assumed mean.