CHAPTER-3

SUMMARY STATISTICS

PART-A

PRESENTATION OF DATA

3.1.1 Frequency:-

Definition: The number of observations in a class is called as frequency or class frequency.

3.1.2 Frequency Distribution:-

Definition: A table containing class intervals along with frequencies is called as frequency distribution.

3.1.3 Frequency distribution of discrete variable:

Procedure:

[1] Find the smallest and the largest observation.

[2] Prepare first column of all possible values of variables from the smallest to the largest.

[3] Consider the observations one by one. Put a tally marks in second column.

[4] Count the number of tally marks and place them in the third column.

3.1.4 Frequency distribution of continuous variable:

Procedure:-

[1] Find the smallest and the largest observation.

[2] The number of classes be preferably between 7 to 20 and general formula is square root of number of observation.

[3] Prepare first column of table by entering the class intervals.

[4] Classify the observations one by one in the appropriate class by putting tally marks in the second column.

[5] Count the tally marks and enter the number in the third column.

3.1.5 Constant:-

Definition: The characteristic which does not change its value or (nature) is considered as constant.

For example*:* Height of a person after 25 years of age and altitude of a certain place from sea level, etc.

3.1.6 Attribute:-

Definition: A qualitative characteristic is called an **attribute.**

For examples: sex, nationality, literacy, religion, grade in examination, blood group, beauty, defective of an article produced by a machine.

3.1.7 Variable: A quantitative characteristic which changes its value is called a variable. **For examples:** weight of person, examination marks, population of a country, profit of a salesman.

3.1.7 (a) Discrete variable:-

Definition: A variable whose domain contains specific isolated values, generally zero and positive integers, is called as **discrete variable.**

For example: Number of students in a class, number of articles produced by a machine, population of a country, number of workers in a factory etc. are discrete variables. Most of the discrete variables have integral values.

3.1.7 (b) Continuous variable:-

Definition: A variable whose domain contains all values within certain limits is called as **continuous variable**.

For example: Weight of a person, height of students, temperature at a certain place, agricultural production , electricity consumption of a family, speed of a vehicle are the examples of continuous variable.

3.2.2 Classification:-

Raw data:-

Definition: The unprocessed data is called raw data. Data obtained on discrete variable are called discrete data whereas those obtained on a continuous variable are called continuous data.

In any statistical survey the information is in the form of raw data. It consists of many observations and from this unnecessary detail will be deleted and important things in the data will be brought into focus. Condensed the data and make comparison with similar data, next step is to classify the data into groups or classes.

3.2.3 Methods of Classification:

There are two methods of classification

(i) inclusive method (ii) exclusive method

[1] Inclusive method:- In this method the observation equal to upper limit is included in the same class. Therefore, the method is called as inclusive method. It can be observed that the upper limit of class is not the same as the lower limit of succeeding class. Therefore, a discontinuity is observed between the classes.

For example:

Class: 100-199, 200-299,300-399.

[2] Exclusive Method:- In this method the observation equal to upper limit does not belong to the same class. It is included in the next class. Therefore, the method is called exclusive method. For example , the observation 400 is included in 400-500. In other words, the observation equal to upper limit is excluded from the same class

For example:

Class: 100-200, 200-300,300-400.

In this case upper limit of one class is the lower limit of subsequent class. The classes are observed to be continuous without any gap in between them.

[3] Class-limits:- The two numbers designating the class- interval are called as class limits.

For example:

Class: 100-199, 200-299,300-399.

[4] Class Boundaries:- The class boundaries are the numbers up to which the actual magnitude of observation in the class can extend. The class boundaries are also called as actual limits or extended limits.

For example:

Class boundaries: 99.5-199.5, 199.5-299.5,299.5-399.5.

It can be clearly seen that in case of exclusive method of classification, class limit and class boundaries are same. Using class-boundaries the classes are made continuous, however, original frequency associated do not alter.

[5] Class mark or Mid Values:- It is the mid –point of class interval and the same can be obtained as follows:

$$\text{Mid-value} = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$

OR

$$\text{Mid-value} = \frac{\text{Upper boundary} + \text{Lower boundary}}{2}$$

[6] Class-width:- It is the actual length of the class interval .

Class width = Upper boundary – Lower boundary

OR

Class width = (Lower limit of the succeeding class) – (Lower limit of the considering class)

OR

Class width = (Upper limit of the succeeding class) – (Upper limit of the considering class)

[7] Open end class:- A class in which one of the limits is not specified is called an open end class.

For example:

Class: Below 200, 200-300,300- and above.

The class below 200 has no lower limit and the class 300 and above has no upper limit. Therefore, these classes are open end classes. **Whenever the extreme observations are widely spread, open end classes are used.** In case of income distribution or the classification of sales of a company, open end classes may be required. Open end classes create some problems in further analysis, therefore , as far as possible the open end classes should be avoided.

3.2.4 Cumulative Frequency Distribution:-

In many situations it is required to find the number of observation below or above certain value.

For example: in case of a frequency distribution of income, the number of persons below poverty line or in case of frequency distribution of examination marks number of candidates above 60 etc. is required to be found . In this case cumulate frequencies are much useful. There are two types of cumulative frequencies:

(i) Less than type cumulative frequency

(ii) More than type cumulative frequency.

(i) Less than type cumulative frequency: Less than type cumulative frequency of a class is the number of observations less than or equal to the upper limit of the corresponding class.

(ii) More than type cumulative frequency: More than type cumulative frequency is the number of observations more than or equal to the lower limit of corresponding class.

It is clear from the above explanation that the less than type cumulative frequencies can be obtained by computing cumulative sum of frequencies from the lowest class to highest class.

For example:

Marks	Frequency	Less than cumulative Frequency	More than cumulative frequency
0-10	5	5	4+4+15+45+5=40
10-20	12	5+12=17	4+4+15+12=35
20-30	15	5+12+15=32	4+4+15=23
30-40	4	5+12+15+4=36	4+4=8
40-50	4	5+12+15+4+4=40	4
Total	40		

It can be noted that the less than cumulative frequency is increasing in nature. Less than cumulative frequency of the lowest class is same as the usual frequency and the lass than type cumulative frequency of highest class is the total number of observations. In case of more than cumulative frequencies exactly reverse pattern will be seen.

3.2.5 Cumulative frequency distribution:-

A table containing upper limits along with less than type cumulative frequency or lower limits along with more than tsype cumulative frequency is called as cumulative frequency distribution.

3.2.6 Relative frequency:-

Two different frequency distributions may not have the same total frequency, hence for the purpose of comparison and interpretation, sometimes it is better to express the frequency of a class in terms of proportion (or percentage) of

the total number of observations. The proportion of number of observations in a class is the relative frequency.

Definition: Frequency of a class divided by the total frequency is called the relative frequency of the data.

$$\text{Relative frequency} = \frac{\text{Class Frequency}}{\text{Total Frequency}}$$

It can be noted that the relative frequency maintains the same pattern which is observed in class frequencies. The total of relative frequencies is always 1.

For example:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	**Total**
Frequency	5	25	27	32	6	5	100
Relative Frequency	0.05	0.25	0.27	0.32	0.06	0.05	1

3.2.7 Guidelines for the choice of classes:-

The classes should be chosen, so that it will condense the data and it will also maintain the patterns in the original data.

[1] The number of classes should not be too large, otherwise it will not serve the purpose of condensation.

[2] The number of classes should not be too small. If the number of classes is too small, it will not reveal the pattern in the original data. The number of classes should be between 7 to 20. However, according to the needs and requirements of the situation appropriate number of classes is chosen. If the number of observations is large, naturally the number of classes will be large.

3.2.8 Sturge's Rule:

If N is the total number of observations to be classified, then according to Sturge's rule, the number of classes is approximately 1+3.322 log N.

By the other approach as a thumb rule, the number of classes is

approximately $\sqrt{N}$.

Note:

[i] As far as possible, classes should be of uniform width.

[ii] As far as possible open end classes should be avoided.

[iii] The class width should be preferably 5 or multiple of 5.

[iv] The lower limit of the starting class be preferably multiple of 5. For example: The classes may be of the type 0-9, 10-19, or 10-20 , 20-30 …..etc.

3.2.9 Graphical presentation of statistical data

[i] Graphs are more effective way to serve the purpose of comparison and revealing the patterns.

[ii] Graphs are easy to understand and create an effect which lasts for a longer time.

[iii] They use voluminous, uninteresting, dry data and present the facts in an attractive and impressive manner.

[iv] They facilitate comparison and hence, conclusions can be drawn quickly, which is not possible with the help of a table or frequency distribution to the same extent.

[v] Therefore, graphs are found to be of immense use in several fields to emphasize the fact. LIC, banks, government agencies, industries use graphs to show their growth, development, extension activities etc.

Here we discuss various graphs associated with frequency distribution. Generally graphs are used to represent mathematical relationship between two variables.

3.2.10 Histogram:

[i] It is one of the popularly used graphs for the representation of frequency distribution.

[ii] It is a series of adjacent rectangles erected on X axis with class interval as base, hence width of rectangle is equal to class width.

[iii] Area of rectangle is taken as proportional to class frequency.

[iv] In inclusive method of classification, extended class interval is used as base, where, extended class interval is an interval designated by class boundaries.

[v] Base of rectangle is depend upon the class width of that class.

Histogram is classified into two classes.

Case (i) Classes of equal width: In this case height of rectangle is proportional to frequency.

Case (ii) Classes of unequal width: In this case height of rectangle is proportional to frequency density.

Where,

$$\text{Frequency density} = \frac{\text{Class Frequency}}{\text{Class width}}$$

Note:

(i) A serious drawback of histogram is that, it cannot be drawn for a frequency distribution with open end class.

(ii) In case of discrete variable, histogram need not contain adjacent rectangles; those may be separated like bar diagram.

(iii) Histogram is useful to find mode. Histogram remains same if class width is changed.

3.2.11 Frequency polygon:

[i] Generally a graph is expected to be in the form of a smooth curve.

[ii] Histogram does not fulfill this requirement. Therefore, another way of presentation of frequency distribution is frequency polygon or frequency curve.

[iii] Mid –values are taken on X-axis and frequencies on Y-axis to draw the graph.

[iv] Successive points are joined by the line segments. Further, to complete polygon we obtain closed figure by taking two more classes, one preceding to

first class and the other succeeding to last class. Frequency of these classes is taken to be zero.

[v] Mid- points of these classes are used to get closed figure. The figure so obtained is called as frequency polygon.

3.2.12 Frequency Curve:

[i] There is little difference in frequency polygon and frequency curve.

[ii] If the points are joined by smooth curves instead of straight lines we get a closed figure called as frequency curve.

[iii] While drawing frequency curve we should take care that the area under the curve is same as that of frequency polygon.

3.2.13 Cumulative Frequency Curve or Ogive:

[i] Cumulative frequency distribution is represented by cumulative frequency curve (or Ogive).

[ii] There are two types of cumulative frequencies hence, there are two types of cumulative frequency curves.

[iii] For less than type cumulative curve upper boundaries of classes are taken on X axis and less than cumulative frequencies on Y axis.

[iv] A preceding class before first class is also taken into consideration for drawing this curve. Cumulative frequency of this class is taken to be Zero.

[v] Similarly, to draw more than type cumulative frequency curve lower boundaries are taken on X-axis and more than cumulative frequencies on Y-axis.

[vi] In this case a succeeding class to the last class is taken with cumulative frequency zero. Those points are joined by smooth curve to get the cumulative frequency curve.

[vii] This type of curve is useful to find median.

3.2.14 Stem and Leaf Chart (Plot):-

It is similar to histogram. Stem and leaf is another way of representing data. It is developed by American statistician John Tukey of Princeton University.

For example: [1] The following are the mark obtained by 40 students in a class. Construct the stem and leaf diagram. Which steam has the maximum leaves?

31	37	02	41	12	38	29	22	15	17
09	50	47	34	34	19	21	24	22	23
12	08	26	32	33	28	36	33	37	13
11	32	15	31	48	22	17	28	33	49

Solution: The marks are two digits ranging from the minimum 02 and maximum 50. The unit places are to be treated as leaves and tenth place number is treated as stem. For example, 36 has stem 3 and leaf 6, 17 has stem 1 and leaf 7. To summarize the data we collect all the numbers having same stem. It is like classification. We make two columns the first one for stem and the second is for leaves. Instead of class interval we write stem and instead of tally marks we write leaves. Clearly in the above problem the stems are 0, 1, 2, 4, 5. Accordingly we prepare stem and leaf chart. The first number is 31, we select stem 3 and in the column of leaf as 1, the next number is 37 we select stem 3 and enter 7 in the column of leaf. In this manner we classify all the scores the final chart will look like as follows.

Stem	**Leaf**
0	2, 9, 8
1	2, 5, 7, 9, 2, 3, 5, 7
2	9, 2, 1, 4, 2, 3, 6, 8, 1, 2, 8
3	1, 7, 8, 4, 4, 2, 3, 6, 3, 7, 2, 1, 3
4	1, 7, 8, 9
5	0

In the above table first of all we have entered the leaves according to order of occurrence. However, we arrange them in the ascending order, thus the final stem and leaf chart will be as follows.

Stem	Leaf
0	2, 9, 8
1	2, 2, 3, 5, 5, 7, 7, 9
2	1, 1, 2, 2, 2, 3, 4, 6, 8, 9, 9
3	1, 1, 2, 2, 3, 3, 3, 4, 4, 6, 7, 7, 8
4	1, 7, 8, 9
5	0

Key: (i) "3.8" in the above stem and leaf plot stands for 38. Stem 3 has maximum leaves.

3.2.15 Comparison between stem-leaf chart and histogram:

Stem and leaf gives bar like's structure in front of stem, while in histogram bars are erected on class-intervals vertically. Like histogram, stem leaf plot also gives shape of frequency distribution. It also gives idea of spread as well as centering. Thus as far as shape spread and centering is concerned, histogram and stem-leaf serve the same purpose. However, in stem and leaf plot we have an advantage that all the original data is presented unlike frequency distribution. Thus there is no loss of data due to condensation.

Note:

[i] Number of stems should be reasonable.

[ii] In case the data items are there digit stem will be two digit and leaf will be single digit at unit place. For example to represent 316 we take 31 s stem and 6 as leaf.

[iii] In case leaves are more we can split to have two stems. For example, there are many leaves for stems. Then we can split the stems as 1, 1* 2, 2*,

3,3* …for stem 1 we include leaves 0, 1, 2, 3, 4 and 1* will have leaves 5, 6, 7, 8, 9 and similarly for 2, 2, 3, 3* etc. If the observations are of the type 102.3, 102.8, 103.5 ..then integers such as 102, 103 can be taken as stem and fractional part as leaf.

3.2.16 Advantages and Limitations of Graph:-

- **Advantage:**

[i] Information is presented in condensed form.

[ii] Facts are presented in more effective and impressive manner as compared to tables.

[iii] Easy to understand for a layman.

[iv] Create effect which lasts for longer time.

[v] Facilitate the comparison.

[vi] Help in revealing patterns.

Limitations:

[i] Using graphs we find the values approximately, while, tables give exact values.

[ii] Graphs give only a general idea about the phenomenon, which is not sufficient for further statistical analysis.

Fill in the blanks

[1] Classification is the …..of facts that are distinguished by some significant …..

Answer: **Grouping, Characteristics**

[2] For a good classification, the class should be ….. and ……..

Answer: **Exhaustive, Mutually Exclusive**

[3] Classification can be done according to…..

Answer: **Characteristics**

[4] Quantitative Classification leads to. …..

Answer: **Frequency**

[5] The difference between the upper and lower limit of a class is called …..

Answer: **Class Interval**

[6] The average of the upper and lower limit of a class is known as …..

Answer: **class mark**

[7] There is a general assumption that the class frequency is centered at the ……of the class.

Answer: **Mid-value**

[8] Diagrams are another form of …. …..

Answer: **Tabulation**

[9] The ogives of less than type and more than type distribution intersect at .. …..

Answer: **Median**

[10] In a histogram bars …… each other.

Answer: **Touch**

[11] A histogram is ….. for geographical classification.

Answer: **Not suitable**

[12] Frequency polygon can be drawn with the help of …..

Answer: **Histogram**

[13].…..diagram can be sketched for geographical series.

Answer: **Bar**

[14] Sub-divided bar diagram depicts the distribution of ….. of a factor.

Answer: **Components**

[15] Histogram can be drawn only for ….. distribution.

Answer: **Continuous frequency**

[16] Pie-chart is always …..

Answer: **Circular**

Choose correct alternative from the following

[1] A quantitative characteristics like weight of person, examination marks, is called ---

(a) variable (b) attributes

(c) frequency (d) none of these

Answer: (a) variable

[2] A qualitative characteristics like religion , nationality, sex is called

(a) variable (b) attributes

(c) frequency (d) none of these

Answer: (b) attributes

[3] A visual representation of the frequency distribution where the frequencies are represented by adjacent bars is called a

(a) frequency polygon (b) scatter gram

(c) histogram (d) tally chart

Answer: (c) histogram

[4] The horizontal axis on a graph is called the

(a) x-axis (b) y-axis

(c) z- axis (d) frequency axis

Answer: (a) x-axis

[5] The vertical axis on a graph is called the

(a) x-axis (b) y-axis

(c) z-axis (d) nominal axis

Answer: (b) y-axis

[6] If 35 is the upper limit of the class-interval of class size 10, then the lower limit of the class-interval is :

(a) 20 (b) 25

(c) 30 (d) none of these

Answer: (b) 25

[7] The curve drawn by taking upper limits along x axis and cumulative frequency along y-axis is :

(a) frequency polygon (b) more than ogive

(c) less than ogive (d) none of these

Answer: (c) less than ogive

[8] For ‘more than ogive’ the x-axis represents:

(a) upper limits of class-intervals

(b) mid-values of class-intervals

(c) lower limits of class-intervals

(d) frequency

Answer: (c) lower limits of class-intervals

[9] Ogive is the graph of:

(a) lower limits and frequency

(b) upper limits and frequency

(c) lower/upper limits and cumulative frequency

(d) none of these

Answer: (c) lower/upper limits and cumulative frequency

[10] The curve ‘less than ogive’ is always:

(a) ascending (b) descending

(c) sometimes ascending and sometimes descending

(d) none of these

Answer: (a) ascending

[11] For a given data with 50 observations the ‘less than ogive’ and the ‘more than ogive intersect at (15.5, 20). The median of the data is:

(a) 4.5 (b) 20 (c) 50 (d) 15.5

Answer: (d) 15.5

[12] The abscissa of the point of intersection of the less than type and of the more thantype cumulative frequency curves of a grouped data gives its:

(a) mean (b) median

(c) mode (d) none of these

Answer: (b) median

[13] The measures of central tendency which can’t be found graphically is

(a) mean (b) median

(c) mode (d) none of these

Answer: (a) mean

[14] The 30 students in a class did a survey of their favourite movie series and recorded the results as follows:

Movies	Frequency
Twilight	10
Harry Potter	6
Narnia	2
High school Musical	9
Pirates of the Caribbean	3
Total	30

What was the relative frequency for High School Musical?

(a) 0.09 (b) 0.3 (c) 0.33 (d) 0.43

Answer: (b) 0.3

Explanation: - 9 students out of 30 said their favourite was High School Musical. Therefore the relative frequency was 9/30 = 3/10 = 0.3

[15] Nationality of a person is

(a) a variable (b) an attributes

(c) a discrete variable (d) none of these

Answer: (b) an attributes

[16] Diagrams and Graphs are tool of_____

(a) collection of data (b) analysis

(c) presentation of data (d) all of these

Answer: (c) presentation of data

[17] The mid -point of a class is obtained by

(a) adding upper and lower limits

(b) dividing the difference of upper and lower limits by two

(c) adding upper and lower limits and dividing it by two

(d) by deducting the upper limit from the lower limit.

Answer: (c) adding upper and lower limits and dividing it by two

[18] Mode is found graphically by

(a) frequency polygon (b) ogive curve

(c) simple bar diagram (d) histogram

Answer: (d) histogram

[19] Median is found graphically by

(a) frequency polygon (b) ogive curve

(c) simple bar diagram (d) histogram

Answer: (b) ogive curve

[20] If the class interval 10-20 then mid value or mid- point of the class is

(a) 10 (b) 15

(c) 20 (d) 30

Answer: (b) 15

[21] The series of observation which contains two modes is called

(a) unimodal (b) bimodal

(c) trimodal (d) none of these

Answer: (b) bimodal

EXERCISE:

[1] Following is the distribution of marks of students in a class:

Marks	0-9	10-19	20-29	30-39	40-49
Frequency	12	15	24	12	8

Answer the following questions:

(i) Find the class-mark of the third class.

(ii) Find the class –width of second class.

(iii) State the class boundaries of fourth class.

(iv) Find the No. of students whose marks are less than 29

[2] Following is the frequency distribution of marks of students in a class:

Marks	0-10	10-20	20-30	30-40	40-50
Frequency	17	19	24	10	4

Answer the following questions:

(i) Obtain the class-mark of the third class.

(ii) Obtain the class –width of second class.

(iii) State the class boundaries of fourth class.

(iv) State the type of classification

[3] Following is the frequency distribution of number of students according to marks scored in a certain examination:

Class	Below 20	20-40	40-60	60-80	80-100
Frequency	18	36	34	22	15

Answer the following questions:

(i) State the type of classification

(ii) Obtain the class-mark of the second class

(iii) Obtain the class –width of third class

(iv) State open end class

(v) How many students getting marks less than 80?

[4] Distinguish between 'inclusive method of classification' and 'exclusive method d classification'.

[5] Explain the construction of the following graphs along with the rough sketches: (i) histogram (ii) frequency polygon (iii) frequency curve (iv) ogives.

[6] Draw Ogive curves for the following grouped frequency distribution

Class	0-20	20-40	40-60	60-80	80-100
Frequency	18	36	34	22	15

[7] Following is the distribution of marks of students in a class:

Marks	0-9	10-19	20-29	30-39	40-49
Frequency	12	15	24	12	8

Draw Histogram and find mode graphically.

[8] From the following information find mode graphically:

Age	20-30	30-40	40-50	50-60	60-70
No. of person	3	7	14	16	8

[9] From the following information find mode graphically.

Income Rs.	30-40	40-50	50-60	60-70	70-80
Frequency	7	8	10	8	5