3. Organisation of Data

Q1. Which of the following alternative is true?

(i) The class midpoint is equal to:

a. The average of the upper class limit and the lower class limit.

b.The product of upper class limit and the lower class limit.

c. The ratio of the upper class limit and the lower class limit

d.None of the above.

Answer: a. The average of the upper class limit and the lower class limit.

Explanation: The class mid value lies half way between the lower and the upper class limits. It is calculated by dividing the sum of upper class limit and lower class limit by 2.

(ii) The frequency distribution of two variables is known as

a. Univariate Distribution Bivariate Distribution
b. Multivariate Distribution d. None of the above

Answer: b. Bivariate Distribution

Explanation: A frequency distribution where two variables are measured for a same set of items through cross classification is called bivariate or two way distribution.

(iii) Statistical calculations in classified data are based on

a. the actual value of observations
b.the upper class limits
c. the lower class limits
d.the class midpoints

Answer: d. the class midpoints

Explanation: Class mid points are the most important values, as they are representatives of the class in case of classified continuous series data and are take for the use in further statistical calculations.

(vi) Under Exclusive method,

a. the upper class limit of a class is excluded in the class interval

b.the upper class limit of a class is included in the class interval

c. the lower class limit of a class is excluded in the class interval

d.the lower class limit of a class is included in the class interval

Answer: a. the upper class limit of a class is excluded in the class interval

Explanation: Under exclusion method the upper limits of one class is the lower limits of the next class. In this way continuity of the data is maintained.

(v) Range is the

a. difference between the largest and the smallest observation

b.difference between the smallest and the largest distribution

c. average of the largest and the smallest observations

d.ratio of the largest to the smallest observation

Answer: a. difference between the largest and the smallest observation

Explanation: Range is equal to largest value minus the smallest value. i.e. R= L-S

Q2. Can there be any advantage in classifying things? Explain with an example from your daily life.

Answer: Classification is a technique of arranging collected data into different groups or classes that have some common characteristics to facilitate analysis and interpretation.

There are various advantages of classifying data which are as follows:

1. It presents the data in a simple form: Classification process eliminates unnecessary details and makes the mass of complex data simple, brief and logical so that a normal person can understand data at a glance. For Example: Data collected in population census is huge and fragmented. It is very difficult to draw any conclusion from these data. But when Data is classified according to some common characteristic e.g. age, occupation, etc. it can be easily understood.

2. Classification of data facilitates comparison: Classified data enable a person to do comparisons and draw inferences about fact For Example: You are given marks obtained by 60

students and you have been told to compare their intelligence level. It is a difficult task, but, when we classify students into 1st, 2nd and 3rd divisions on the basis of marks obtained, it makes comparison easy.

3. Classification of data highlights points of similarity and dissimilarity. For example population data can be classified as Employed and Unemployed, educated and uneducated, male and female, etc.

4. Classification of data prepares the basis for tabulation and statistical study of data.

5. Classification helps in finding out cause effect relationship in the data if there is any in the data. For example: Data of unemployed workers can help in finding out whether unemployed workers are more in educated or uneducated population.

Q3. What is a variable? Distinguish between a discrete and a continuous variable.

Answer: A variable refers to the characteristics of the sample or population under study that can be expressed in numbers and are generally represented by the letters X, Y, Z. Each value of a variable is an observation. For example: weight of students, income of workers, number of members in a family etc.

Continuous variable are those variable which can take all possible values in a given range. For example: heights and weights of individuals, prices of commodities. They can take any value. Integral values (1,2,3,4…) fractional values (1/2, 2/3, 4/5 ….) or values which are not rational numbers (√2, √3 , √7 √8 …)

Height does not change from 5ft 2” to 5ft 3”. But it will pass all possible values between these. Thus the variable height is capable of manifesting in any conceivable value.

Discrete variables on the other hand can take only some particular values. For example: members of a household. We can have 1member, 2 members, 3 members and so on. So it cannot have fractional values. The value will jump from 1 to 2, from 2 to 3 and so on.

But it does not mean that a discrete variable can take only whole numbers as values. We can have a variable ‘x’ which may take values like 1/2, 1/4, 1/8, 1/16 etc. provided there is a uniform difference from one variable to other variable. It means it jumps from one value to the other and cannot take values in between those two fractions.

Q4. Explain the ‘exclusive’ and ‘inclusive’ methods used in classification of data.

Answer: There are two methods of classifying data into class intervals.

1. Exclusive method: Under this method the upper limits of one class is the lower limits of the next class. In this way continuity of the data is maintained. E.g. class intervals are 0-5, 5-10, 10-

15 and so o Now 5 is coming twice so is 10 and 15. So the upper limit of the class is excluded , means if a student has obtained 5 marks he is not included in the first group i.e. 0-5, but in the second i.e.5-10.

2. Inclusive method: Under this method upper limits of the class interval are also included in that class. The class interval will be made like 0-4, 5-9, 10–14 and so o This does not exclude the upper class limit in a class interval. Both class limits upper and lower limits are parts of the class interval.

Q5. Use the data in Table 3.2 that relate to monthly household expenditure (in `) on food of 50 households and

(i) Obtain the range of monthly household expenditure on food.

(ii) Divide the range into appropriate number of class intervals and obtain the frequency distribution of expenditure.

(iii) Find the number of households whose monthly expenditure on food is

(a) less than Rs 2000 (b) more than Rs 3000

(c) between Rs 1500 and Rs 2500

Answer: Table 3.2

Monthly Household Expenditure (in Rupees) on food of 50 households

1904	1559	3473	1735	2760
2041	1612	1753	1855	4439
5090	1085	1823	2346	1523
1211	1360	1110	2152	1183
1218	1315	1105	2628	2712
4248	1812	1264	1183	1171

1007	1180	1953	1137	2048
2025	1583	1324	2621	3676
1397	1832	1962	2177	2575
1293	1365	1146	3222	1396

(i) The range of monthly household expenditure on food is obtained by subtracting the lowest expenditures from the highest expenditure i.e. Range = 5090 – 1007

Range = Rs 4083

(ii) Let to class interval be equal to Rs 500.

Dividing the range by class interval of we get, Number of classes = 4083/500 = 8.166 $\sim$ 9 classes.

Now, these 9 classes will include all the given values of expenditures of households of food. We will prepare a continuous frequency distribution by exclusion method.

The frequency distribution of expenditure is as follows:

Expenditure on food (Rs)	Tally Bars	Frequency
1000-1500	llll llll llll llll	20
1500-2000	llll llll lll	13
2000-2500	llll l	6
2500-3000	llll	5
3000-3500	ll	2

3500-4000	I	1
4000-4500	II	2
4500-5000	…	0
5000-5500	I	1
Total		50

(iii)

a. the number of households whose monthly expenditure on food is less than Rs 2000 is 20 + 13 = 33 households.
b. the number of households whose monthly expenditure on food is more than Rs 3000 =2 +1+2+0+1= 6 households.
c. the number of households whose monthly expenditure on food is between Rs 1500 and Rs 2500 is 13 + 6 = 19 households.

Q6. In a city 45 families were surveyed for the number of cell phones they used. Prepare a frequency array based on their replies as recorded below.

1	3	2	2	2	2	1	2	1	2	2	3	3	3	3
3	3	2	3	2	2	6	1	6	2	1	5	1	5	3
2	4	2	7	4	2	4	3	4	2	0	3	1	4	3

Answer: The frequency array of cell phones used by 45 families is as follows:

Number of cell phones	**Tally bars**	**Frequency**
0	I	1

1	IIII II	7
2	IIII IIII IIII	15
3	IIII IIII II	12
4	IIII	5
5	II	2
6	II	2
7	I	1
Total		45

Thus from the above frequency array we can conclude that, out of 45 families, one family is not using any cell phone, 7 families are using 1 cell phone, 15 families are using 2 cell phones, 12 families are using 3 cell phones, 5 families are using 4 cell phones, 2 families are using 5 cell phones, 2 are using 6 cell phones and only 1 family is using 7 cell phones.

Q7. What is 'loss of information' in classified data?

Answer: Once the data is grouped into classes, an individual observation has no significance in further statistical calculations. For example we have a class interval as 20-30. There are 6 observations in it i.e. 24, 25, 25, 27, 22, 28. When that data is grouped,the class 20-30 will show 6 frequencies. All values in this class will be assumed to be equal to the mid-value or the class mark. i.e. 25. Further calculations are based on 25 only.

Thus the use of class mark instead of actual values involves considerable loss of information. While classification summarises the data it does not show the details.

Q8. Do you agree that classified data is better than raw data? why?

Answer: Yes, classified data is better than raw data. After data is collected the next task is to arrange it in such a way that becomes easy to handle and convenient for further analyses.

Classification means arranging things in an appropriate order and putting them in some homogeneous groups or classes. For example in a library, books are kept in some order, according to the subjects or authors, or in alphabetical order.

Classified data is certainly better than unclassified data for many reasons.

1. Classification saves our time by making it easier to find an item. If we want to find the highest mark of a student from marks of 1000 students we will have to go through the whole unclassified data. It would be a very tedious task. But if we had it classified under different classes in ascending or descending order we could easily find it.
2. Classification compresses data into groups of similar types. So it is easier to look at it and make comparisons with the other groups.
3. Proper organisation is needed before any further systematic statistical analyses can be undertaken.

Q9. Distinguish between univariate and bivariate frequency distribution.

Answer: The frequency distribution of a single variable is called Univariate distribution. If we collect figures about marks in mathematics of 100 students, it is a univariate distribution.

A bivariate frequency distribution is a frequency distribution of two variables. If we have data of students' heights and weights in the same table that is bivariate distribution. The values of heights are classed in columns and values of weights are classed in rows. Each cell shows the frequency of the corresponding row and column. For example 3 children are 6 feet tall and weigh 55 kg.

Q10 Prepare a frequency distribution by inclusive method taking class interval of 7 from the following data.

28	**17**	**15**	**22**	**29**	**21**	**23**	**27**	**18**	**12**
7	**2**	**9**	**4**	**1**	**8**	**3**	**10**	**5**	**20**
16	**12**	**8**	**4**	**33**	**27**	**21**	**15**	**3**	**36**
27	**18**	**9**	**2**	**4**	**6**	**32**	**31**	**29**	**18**
14	**13**	**15**	**11**	**9**	**7**	**1**	**5**	**37**	**32**

28	**26**	**24**	**20**	**19**	**25**	**19**	**20**	**6**	**9**

Answer: Frequency distribution using (inclusive method and taking the class interval of 7 is as follows:

Class Intervals	Tally Bars	Frequency
0-7	IIII IIII IIII	15
8-15	IIII IIII IIII	15
16-23	IIII IIII IIII	14
24-31	IIII IIII I	11
32-39	IIII	05
Total		60