Module 1

INTRODUCTION

The term statistics seems to have been derived from the Latin word *'status'* or Italian word *'statista'* or the German word *'statistic,* each of which means *political state.*

The word 'Statistics' is usually interpreted in two ways. The first sense in which the word is used is a plural noun just refer to a collection of numerical facts. The second is as a singular noun to denote the methods generally adopted in the collection and analysis of numerical facts. In the singular sense the term 'Statistics' is better described as statistical methods.

Different authors have defined statistics in different ways. According to Croxton and Cowden statistics may be defined as *''collection, organisation presentation, analysis and interpretation of numerical data''*

Population and sample

Population

An aggregate of individual items relating to a phenomenon under investigation is technically termed as 'population'. In other words a collection of objects pertaining to a phenomenon of statistical enquiry is referred to as population or universe. Suppose we want to collect data regarding the income of college teachers under University of Calicut,, then, the totality of these teachers is our population.

In a given population, the individual items are referred to as elementary units, elements or members of the population. The population has the statistical characteristic of being finite or infinite. When the number of units under investigation are determinable, it is called finite population. For example, the number of college teachers under Calicut University is a finite population. When the number of units in a phenomenon is indeterminable, eg, the number of stars in the sky, it is called an infinite population.

Sample

When few items are selected for statistical enquiry, from a given population it is called a 'sample'. A sample is the small part or subset of the population. Say, for instance, there may be 3000 workers in a factory. One wants to study their consumption pattern. By selecting only 300 workers from the group of 3000, sample for the study has been taken. This sample is not studied just for its own sake. The motive is to know the true state of the population. From the sample study statistical inference about the population can be done.

Census and sample Method

In any statistical investigation, one is interested in studying the population characteristics. This can be done either by studying the entire items in the population or on a part drawn from it. If we are studying each and every element of the population, the process is called *census method* and if we are studying only a sample, the process is called sample survey, *sample method* or *sampling*. For example, the Indian population census or a socio economic survey of a whole village by a college planning forum are examples of census studies. The national sample survey enquiries are examples of sample studies.

Advantages of Sampling

The sample method is comparatively more economical.

The sample method ensures completeness and a high degree of accuracy due to the small area of operation

It is possible to obtain more detailed information, in a sample survey than complete enumeration.

Sampling is also advocated where census is neither necessary nor desirable.

In some cases sampling is the only feasible method. For example, we have to test the sharpness of blades-if we test each blade, perhaps the whole of the product will be wasted; in such circumstances the census method will not be suitable. Under these circumstances sampling techniques will be more useful.

A sample survey is much more scientific than census because in it the extent of the reliability of the results can be known where as this is not always possible in census.

Frequency Distribution

Variables and Attributes

A quantity which varies from one person to another or one time to another or one place to another is called a variable. It is actually a numerical value possessed by an item. For example, price of a given commodity, wages of workers, production and weights of students etc.

Attribute means a qualitative characteristic possessed by each individual in a group. It can't assume numerical values. For example, sex, honesty, colour etc.

This means that a variable will always be a quantitative characteristic. Data concerned with a quantitative variable is called *quantitative data* and the data corresponding to a qualitative variable is called *qualitative data*.

We can divide quantitative variables into two (i) discrete (ii) continuous. Those variables which can assume only distinct or particular values are called *discrete* or *discontinuous* variables. For example, the number of children per family, number rooms in a house etc. Those variables which can take any numerical value with in a certain range are known as *continuous* variables. Height of a boy is a continuous variable, for it changes continuously in a given range of heights of the boys. Similar is the case of weight,: production, price, demand, income, marks etc.

Types of Frequency Distribution

Erricker states "frequency distribution is a classification according to the number possessing the same values of the variables''. It is simply a table in which data are grouped into classes and the number of cases which fall in each class is recorded. Here the numbers are usually termed as 'frequencies'. There are discrete frequency distributions and continuous frequency distributions.

1. Discrete Frequency Distribution

If we have a large number of items in the data it is better to prepare a frequency array and condense the data further. Frequency array is prepared by listing once and consecutively all the values occurring in the series and noting the number of times each such value occurs. This is called discrete frequency distribution or ungrouped frequency distribution.

Illustration: The following data give the number of children per family in each of 25 families 1, 4, 3, 2, 1, 2, 0, 2, 1, 2, 3, 2, 1, 0, 2, 3, 0, 3, 2, 1, 2, 2, 1, 4, 2. Construct a frequency distribution.

No of children	Tally marks	No of families
0	|||	3
1	𝍸 |	6
2	𝍸 𝍸	10
3	||||	4
4	||	2
Total		25

2. Continuous Frequency Distribution

An important method of condensing and presenting data is that of the construction of a continuous frequency distribution or grouped frequency distribution. Here the data are classified according to class intervals.

The following are the rules generally adopted in forming a frequency table for a set of observations.

Note the difference between the largest and smallest value in the given set of observations

Determine the number classes into which the difference can be divided.

The classes should be mutually exclusive. That means they do not overlap.

Arrange a paper with 3 columns, classes, tally marks and frequency.

Write down the classes in the first column.

Go though the observations and put tally marks in the respective classes.

Write the sum of the tally marks of each class in the frequency column.

Note that the sum of the frequencies of all classes should be equal to the total number of observations.

Concepts of a Frequency Table

i. Class limits: The observations which constitute a class are called class limits. The left hand side observations are called lower limits and the right hand side observations are called upper limits.

Working classes: The classes of the form 0-9, 10-19, 20-29,... are called working classes or nominal classes. They are obtained by the inclusive method of classification where both the limits of a class are included in the same class itself.

Actual classes: If we are leaving either the upper limit or the lower limit from each class, it is called exclusive method of classification. The classes so obtained are called 'actual classes' or 'true classes'.

The classes 0.5 - 9.5, 9.5 - 19.5, 19.5 - 29.5,... are the actual classes of the above working classes. The classes of the type 0-10, 10

20, 20 - 30,... are also treated as actual classes. There will be no break in the actual classes. We can convert working classes to the corresponding actual classes using the following steps.

Note the difference between one upper limit and the next lower limit.

Divide the difference by 2.

Subtract that value from the lower limits and add the same to the upper limits.

For example

Working Classes	Frequency	Actual Classes
1-2.9	2	0.95-2.95
3-4.9	8	2.95-4.95
5-6.9	10	4.95-6.95
7-8.9	5	6.95-8.95

Class boundaries: The class limits of the actual classes are called actual class limits or class boundaries.

Class mark: The class marks or mid value of classes is the average

of the upper limit and lower limit of that class. The mid value of working classes and the corresponding actual classes are the same. For example, the class mark of the classes 0 - 9, 10 - 19, 20 - 29,... are respectively 4.5, 14.5, 24.5,...

Class interval: The class interval or width of a class is the difference between upper limit and lower limit of an actual class. It is better to note that the difference between the class limits of a working class is not the class interval. The class interval is usually denoted by 'c' or i or 'h'.

Example

Construct a frequency distribution for the following data

70	45	33	64	50	25	65	74	30	20
55	60	65	58	52	36	45	42	35	40
51	47	39	61	53	59	49	41	20	55
46	48	52	64	48	45	65	78	53	42

Solution

Classes	Tally marks	Frequency
20-29	𝍸	3
30-39	𝍸	5
40-49	𝍸 𝍸 ||	12
50-59	𝍸 𝍸	10
60-69	𝍸 ||	7
70-79	|||	3
Total		40

Cumulative Frequency Distribution

An ordinary frequency distribution show the number of observations falling in each class. But there are instances where we want to know how many observations are lying below or above a particular value or in between two specified values. Such type of information is found in cumulative frequency distributions.

Cumulative frequencies are determined on either a less than basis or more than basis. Thus we get less than cumulative frequencies (<CF) and greater than or more than cumulative frequencies (>CF). Less than CF give the number of observations falling below the upper limit of a class and greater than CF give the number of observations lying above the lower limit of the class. Less than CF are obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulation is started from the lowest size of the class to the highest size, (usually from top to bottom). They are based on the upper limit of actual classes.

More than CF distribution is obtained by finding the cumulation or total of frequencies starting from the highest size of the class to the lowest class, (ie., from bottom to top) More than CF are based on the lower limit of the actual classes.

Classes	f	UL	<CF		LL	>CF	
0-10	2	10	2	2	0	3+7+10+8+5+1	235
10-20	5	20	2+5	7	10	3+7+10+8+5	33
20-30	8	30	2+5+8	15	20	3+7+10+8	28
30-40	10	40	2+5+8+10	25	30	3+7+10	20
40-50	7	50	2+5+8+10+7	32	40	3+10	13
50-60	3	60	2+5+8+10+7+3	35	50	3	3

EXERCISES

Multiple Choice Questions

A qualitative characterisic is also known as

a. attribute b. variable

c . variate d. frequency

A variable which assumes only integral values is called

a. continuous b. discrete

c . random d. None of these

An example of an attribute is

a. Height b. weight

c . age d. sex

Number of students having smoking habit is a variable which is

a. Continuous b. discrete

c . neither disrete nor continuous

None of these

A series showing the sets of all district values individually with their frequencies is known as

grouped frequency distribution

simple frequency distribution

cumulative frequency distribution

none of the above

A series showing the sets of all values in classes with their corersponding frequencies is knowsn as

grouped frequency distribution

simple frequency distribution

cumulative frequency distribution

none of the above

If the lower and upper limits of a class are 10 and 40 respectively, the mid points of the class is

a. 25.0 b. 12.5 c. 15.0 d. 30.0

13. In a grouped data, the number of classes preferred are

a. minimum possible b. adequate

c. maximum possible d. any arbitrarily chosen number

Class interval is measured as:

the sum of the upper and lower limit

half of the sum of upper and lower limit

half of the difference between upper and lower limit

the difference between upper and lower limit

A group frequency distribution with uncertain first or last classes is known as:

exclusive class distribution

inclusive class distribution

open end distribution

discrete frequency distribution

Very Short Answer Questions

Define the term ‘statistics’.

Define the term population.

What is sampling

What is a frequency distribution?

21 Distinguish between discrete and continuous variables.

Short Essay Questions

Explain the different steps in the construction of a frequency table for a given set of observations.

Explain the terms (i) class interval (ii) class mark (iii) class frequency.

Distinguish between census and sampling

What are the advantages of sampling over census?

23. State the various stages of statistical investigation.

Long Essay Questions

Present the following data of marks secured in Statistics (out of 100) of 60 students in the form of a frequency table with 10 classes of equal width, the lowest class being 0-9

41	17	83	60	54	91	60	58	70	07
67	82	33	45	57	48	34	73	54	62
36	52	32	72	60	33	07	77	28	30
42	93	43	80	03	34	56	66	23	63
63	11	35	85	62	24	00	42	62	33
72	53	92	87	10	55	60	35	40	57

Solomon Joseph