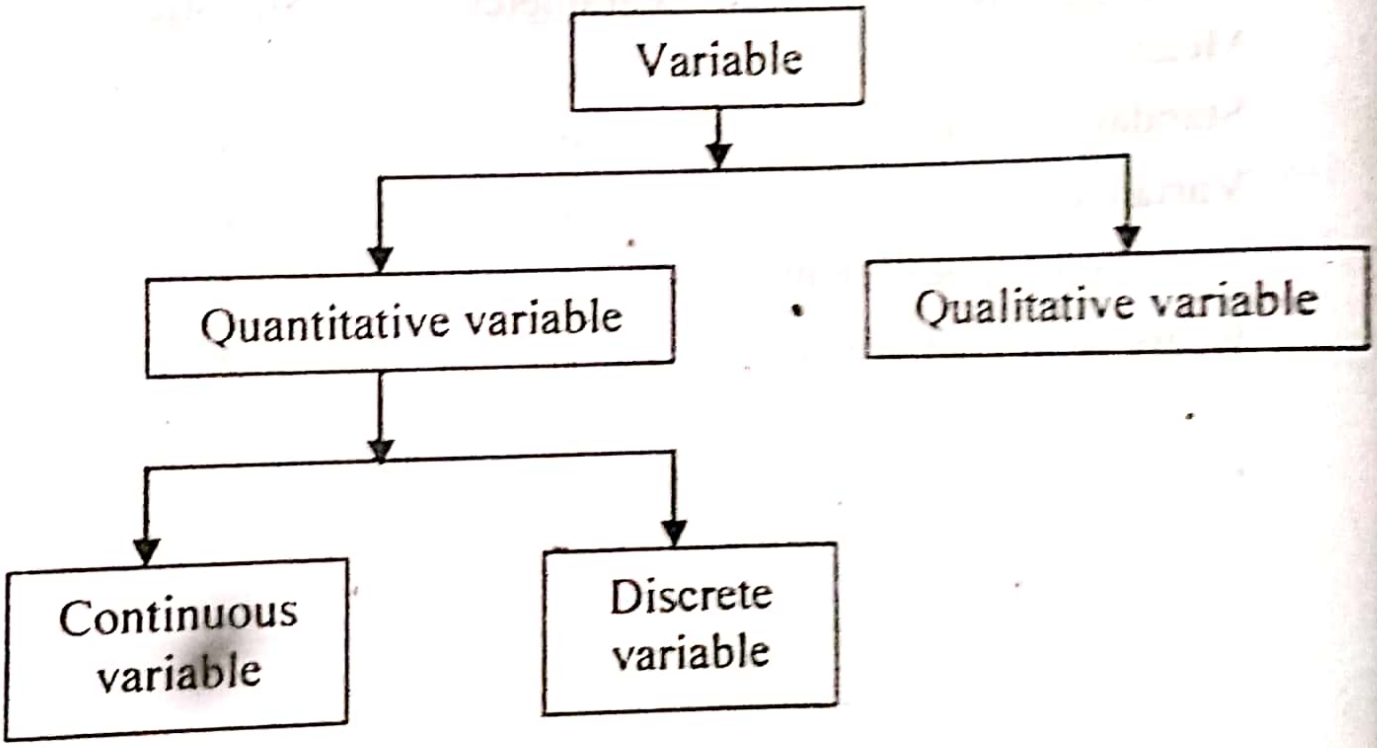
2.1. Variable :

Measurable characteristics of a population that may vary from element to element either in magnitude or in quality are called *variables*. Variables are of two types - *quantitative variable* and *qualitative variable*.

Variables and its classification can be demonstrated as shown below :



Quantitative Variable :

Variable characteristics, whose values are expressed numerically, are known as quantitative variables: Height or weight of students, length or breadth of fishes, weight of tomato, number of grapes per bunch, number of grains per panicle, etc. are some examples of quantitative variables.

Qualitative Variable :

Some variables, which express the quality of population elements, cannot be numerical measured but can be classified or

categorised, these are called qualitative variables. For example, merit of students, educational attainment, type of farmers (big, medium, small), type of fishes (sea fish, river fish) etc. cannot be numerically measured but can be grouped into classes or categories. Qualitative variables are also known as attributes.

Quantitative variables are of two types - continuous and discrete.

A variable which can assume any value, integral or fractional, within specified limits, is called a *continuous variable*. For example, height of students, weight of tomato, length of fish, height of trees, weight of animal etc. are continuous variables which can take both integral or fractional values.

On the other hand some variables can take only integral values e.g., number of grains per panicle, number of students per class, number of fishes caught per unit time etc. These are called *discrete variables*.

2.2. Frequency and Frequency Distribution :

In many situations several population elements assume same values; i.e., numerical values of population characteristics may often repeated again and again. For example, several panicles may contain the same number of grains, a number of fishes may have the same length or weight, several students in a class may have the same height. Such repeatation of the value of variable is called frequency. That is, the number of times a particular observation occurs in the data set is the frequency of that particular observation. Frequency is usually denoted by 'f'.

Frequency Distribution :

Information collected in any process are usually classified or grouped according to specific characteristics. Arrangement of

observational data according to frequencies of the observations is called *frequency distribution*.

Frequency distribution should be such that the arrangement according to the observations becomes easily understandable. Frequency distributions are constructed mainly to present the data in condensed form and for easy understanding. Frequency distribution is very important in statistical studies.

Construction of Frequency Distributions :

Steps in constructing a frequency distribution are discussed below :

1. Finding the Range : In constructing frequency distribution the highest and the lowest value in the data set are first identified and their difference is obtained. This difference between the highest value and the lowest value is called the *range* usually denoted by R.

$$\text{Range} = \text{Highest value} - \text{Lowest value}.$$

2. Decision About the Number of Classes : After finding the range, it is necessary to decide the number of classes in which the entire data set should be divided. Choice of the number of classes should be realistic; this number should not be very small and at the same time it should not be very large so that the aim of construction of frequency distribution (condensation) is not achieved. It is generally expected to limit the number of classes between 7 and 15. There is no hard and fast rule for choosing the number of classes. However, M.A. Sturge's formula gives a guideline for desired number of classes. The formula is -

$$k = 1 + 3.322 \log_{10} N$$

where N is the total number of observations in the data set and k is the desired number of classes.

3. Choosing the Class Interval : The next step of constructing frequency distribution is the calculation of the class interval. Each class will have two limits, the lower limit (the lower value) and the upper limit (the higher value). The difference of the upper limit and the lower limit of a class is known as *class interval*, usually denoted by c or h. If the range is divided by the number of classes, we get the class interval.

$$\text{Class Interval (C)} = \frac{\text{Range}}{\text{No. of classes}}$$

The value of c is taken as the next integral value of the ratio R/k. For choosing the class interval, there is no rigid rule as to use the exact end values of the data set, rather convenient values near the highest and lowest observations of the data set may be used. However, class interval should be such that classes are distinct and separate from each other. Depending on the nature of the variable, two different methods are used in choosing the class limits. If the variable is discrete, closed intervals like $a \leq x \leq b$ are used, both the lower and upper limits are included (e.g., 0-4, 5-9, 10-14, ..., etc.). On the other hand, if the variable is continuous, open interval system ($a < x \leq b$ or $a \leq x < b$) is used; one of the class limits is included and the other is excluded (usually the lower limit is included). In this case the classes will be 0-5, 5-10, 10-15, ..., etc.

4. Counting of Frequencies : For convenience of counting the number of observations falling within each class tally marks are used; frequency of each class is determined by counting the tally marks.

Sometimes it may be necessary to know the observations greater or smaller than a particular value or class of values. For this, cumulative frequencies for observation or class are obtained.

Example 2.1 :

Suppose the marks obtained by 50 students in an examination in Economics are as follows :

32	27	19	40	31	17	15	18	21	27	38	15	33	34	29
26	16	25	33	36	24	22	26	19	36	18	25	20	25	25
31	24	16	28	30	24	29	42	29	28	26	27	47	43	22
25	28	22	24	23										

Here the variable is the marks obtained by the students. The data as shown above are called raw or ungrouped data.

If it is needed to describe the performance of the students, it may be done in a number of ways.

We may enumerate the grade of each student either in ascending or descending order; data such arranged are said to be arranged in array. Counting the number of times each value of the variable occurs, we get a table of the following type :

Table 2.1 Frequency Distribution of Marks

Marks	Frequency (No. of students)	Cumulative frequency	Marks	Frequency (No. of students)	Cumulative frequency
15	2	2	28	3	33
16	2	4	29	3	36
17	1	5	30	1	27
18	2	7	31	2	39
19	2	9	32	1	40
20	1	10	33	2	42
21	1	11	34	1	43
22	3	14	36	2	45
23	2	16	38	1	46
24	4	20	40	1	47
25	5	25	42	1	48
26	3	28	43	1	49
27	2	30	47	1	50

Such a table is known as frequency table or frequency distribution. The above arrangement is an improvement over the raw data, but to get a still better idea of the performance of the students we reclassify the data into grouped frequency distribution as shown below :

Table 2.2: Grouped Frequency Distribution of Marks

Class interval	Tally mark	Frequency (No. of students)	Cumulative frequency	
			Ascending	Descending
15-20	𝍸 ||||	9	9	50
20-25	|||| 𝍸 |	11	20	41
25-30	𝍸 𝍸 𝍸 |	16	36	30
30-35	𝍸 ||	7	43	14
35-40	|||	3	46	7
40-45	|||	3	49	4
45-50	|	1	50	1

This type of classification of raw data is called grouped frequency distribution or simply frequency distribution.

In the above example the highest value is 47, the lowest value is 15 and the range is, R = 47-15 = 32.

According to Sturge's formula,

$$k = 1 + 3.322 \log_{10} 50 = 6.47$$

That is, 6 to 7 groups are appropriate in this case.

Again, $C = \frac{R}{k} = \frac{32}{7} = 4.57$; accordingly 5 is taken as the class interval.