2·1. INTRODUCTION

Quantitative data in a mass exhibit certain general characteristics or they differ from each other in the following ways :

1. They show a tendency to concentrate at certain values, usually somewhere in the centre of the distribution. Measures of this tendency are called *measures of central tendency* or *averages*.
2. The data vary about a measure of central tendency and these measures of deviation are called *measures of variation or dispersion*.
3. The data in a frequency distribution may fall into symmetrical or asymmetrical patterns. The measures of the direction and degree of asymmetry are called *measures of skewness*.
4. Polygons of frequency distributions exhibit flatness or peakedness of the frequency curves. The measures of peakedness or flatness of the frequency curves are called *measures of kurtosis*.

2·2. FREQUENCY DISTRIBUTION

When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in Statistics obtained by 250 candidates selected at random from among those appearing in a certain examination.

TABLE 2·1 : MARKS IN STATISTICS OF 250 CANDIDATES

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46
41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38
31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	34	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as *array*. But this does not reduce the bulk of the data. A much better representation is given in *Table* 2·2 :

TABLE 2·2

Marks	*No. of Students —Tally Marks*	*Total Frequency*	*Marks*	*No. of Students —Tally Marks*	*Total Frequency*
15	ǀǀ	= 2	40	卌 卌 ǀ	= 11
17	ǀǀǀ	= 3	41	卌 卌	= 10
18	ǀǀ	= 2	42	卌 卌 ǀǀǀ	= 13
19	ǀǀ	= 2	43	卌 ǀǀǀ	= 8
21	ǀǀ	= 2	44	卌 卌 ǀǀ	= 12
22	ǀǀ	= 2	45	卌 ǀǀ	= 7
23	ǀǀǀ	= 3	46	卌 ǀǀ	= 7
24	ǀǀǀǀ	= 4	47	卌 ǀǀǀ	= 8
25	ǀ	= 1	48	卌 卌 ǀǀ	= 12
26	ǀǀǀ	= 3	49	ǀǀǀ	= 3
27	ǀ	= 1	50	卌 卌	= 10
28	ǀǀǀ	= 3	51	ǀǀǀǀ	= 4
29	ǀǀ	= 2	52	卌	= 5
30	卌	= 5	53	ǀǀǀǀ	= 4
31	卌 卌	= 10	54	ǀǀǀ	= 3
32	卌 卌	= 10	55	ǀǀ	= 2
33	卌 ǀǀǀ	= 8	56	ǀǀ	= 2
34	卌 卌 ǀ	= 11	57	ǀǀ	= 2
35	卌	= 5	58	ǀǀ	= 2
36	卌	= 5	60	ǀǀǀ	= 3
37	卌 卌 ǀǀ	= 12	61	ǀ	= 1
38	卌 卌 卌 ǀǀ	= 17	62	ǀ	= 1
39	卌 ǀ	= 6	68	ǀ	= 1

A bar (|) called *tally mark* is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (|) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

The representation of the data as above is known as *frequency distribution.* Marks are called the *variable* (*x*) and the '*number of students*' against the marks is known as the '*frequency*' (*f*) of the variable. The word '*frequency*' is derived from 'how frequently' a variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks. This representation, though better than an 'array', does not condense the data much and it is quite cumbersome to go thorough this huge mass of data.

If the identity of the individuals about whom a particular information is taken is not relevant, nor the order in which the observations arise, then the first real step of condensation is to divide the observed range of variable into a suitable number of *class-intervals* and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in *Table 2·3*.

TABLE 2·3 : FREQUENCY TABLE

Marks (x)	*No. of students* (f)
15—19	9
20—24	11
25—29	10
30—34	44
35—39	45
40—44	54
45—49	37
50—54	26
55—59	8
60—64	5
65—69	1
Total	250

Such a table showing the distribution of the frequencies in the different classes is called a *frequency table* and the manner in which the class frequencies are distributed over the class intervals is called the *grouped frequency distribution* of the variable.

Remark. The classes of the type 15—19, 20—24, 25—29 etc., in which both the upper and lower limits are included are called *'inclusive classes'*. For example, the class 20— 24, includes all the values from 20 to 24, both inclusive and the classification is termed as *inclusive type classification.*

In spite of great importance of classification in statistical analysis, no hard and fast rules can be laid down for it. The following points may be kept in mind for classification :

1. The classes should be clearly defined and should not lead to any ambiguity.

2. The classes should be exhaustive, *i.e.,* each of the given values should be included in one of the classes.

3. The classes should be mutually exclusive and non-overlapping.

4. The classes should be of equal width. The principle, however, cannot be rigidly followed. If the classes are of varying width, the different class frequencies will not be comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The ratios thus obtained are called, *'frequency densities'*.

5. Indeterminate classes, *e.g.,* the open-end classes like less than '*a*' or greater than '*b*' should be avoided as far as possible since they create difficulty in analysis and interpretation.

6. The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However, the number of classes may be more than 15 depending upon the total frequency and the details required, but it is desirable that it is not less than 5 since in that case the classification may not reveal the essential characteristics of the population. The following formula due to Struges may be used to determine an approximate number k of classes :

$$k = 1 + 3{\cdot}322 \log_{10} N, \text{ where } N \text{ is the total frequency.}$$

The Magnitude of the Class Interval. Having fixed the number of classes, divide the range (the difference between the greatest and the smallest observation) by it and the nearest integer to this value gives the magnitude of the class interval. Broad class intervals (*i.e.,* less number of classes) will yield only rough estimates while for high degree of accuracy small class intervals (*i.e.,* large number of classes) are desirable.

concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use :

(*i*) *Arithmetic Mean or Simple Mean ,* (*ii*) *Median,* (*iii*) *Mode,*
(*iv*) *Geometric Mean, and* (*v*) *Harmonic Mean.*

2·4·1. Requisites for an Ideal Measure of Central Tendency. According to Professor Yule, the following are the characteristics to be satisfied by an ideal measure of central tendency :

(*i*) It should be rigidly defined.

(*ii*) It should be readily comprehensible and easy to calculate.

(*iii*) It should be based on all the observations.

(*iv*) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series, we should be able to calculate the average of the composite series obtained on combining the given series.

(*v*) It should be affected as little as possible by fluctuations of sampling .

In addition to the above criteria, we may add the following (which is not due to Prof. Yule) :

(*vi*) It should not be affected much by extreme values.

2·5. ARITHMETIC MEAN

Arithmetic mean of a set of observations is their sum divided by the number of observations, *e.g.*, the arithmetic mean $\overline{x}$ of n observations $x_1, x_2, \ldots, x_n$ is given by :

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots (2\cdot1)$$

In case of the frequency distribution $x_i | f_i$, $i = 1, 2, \ldots n$, where f_i is the frequency of the variable x_i,

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N\right] \qquad \ldots (2\cdot1a)$$

In case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Remark. The symbol Σ is the letter capital sigma of the Greek alphabet and is used in mathematics to denote the sum of values.

Example 2·1. (*a*) *Find the arithmetic mean of the following frequency distribution :*

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6

(*b*) *Calculate the arithmetic mean of the marks from the following table :*

Marks :	0—10	10—20	20—30	30—40	40—50	50—60
No. of Students :	12	18	27	20	17	6