CHAPTER 1

Classification and Tabulation of Data

1.1. INTRODUCTION

The word '**statistics**' has been derived from the Latin word '**status**'. In the plural sense it means a set of numerical figures called '**data**' obtained by counting, or, measurement. In the singular sense it means collection, classification, presentation, analysis, comparison and meaningful interpretation of '**raw data**'.

It has been defined in different ways by different authors. Croxton and Cowdon defined it as 'the science which deals with the collection, analysis and interpretation of numerical data'.

Statistical data help us to understand the economic problems, e.g., balance of trade, disparities of income and wealth, national income accounts, supply and demand curves, living and whole sale price index numbers, production, consumption, etc., formulate economic theories and test old hypothesis. It also helps in planning and forecasting.

The success of modern business firms depends on the proper analysis of statistical data. Before expansion and diversification of the existing business or setting up a new venture, the top executives must analyse all facts like raw material prices, consumer-preferences, sales records, demand of products, labour conditions, taxes, etc., statistically. It helps to determine the location and size of business, introduce new products or drop an existing product and in fixing product price and administration. It has also wide application in Operations Research.

1.2. LIMITATIONS OF STATISTICS AND ITS CHARACTERISTICS

(*i*) Statistics studies a group but not an individual.

(*ii*) Statistics cannot be applied to study the qualitative phenomenon.

(*iii*) Statistical decisions are true on an average only. For better results a large number of observations are required.

(*iv*) Statistical data are not mathematically accurate.

(*v*) Statistical data must be analysed by statistical experts otherwise the results may be misleading.

(*vi*) The laws of statistics are not exact like the laws of sciences. The first law states "a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group". The second law states "other things being equal, as the sample size increases, the result tends to be more reliable and accurate".

(*vii*) Statistics collected for a given purpose must be used for that purpose only.

(*viii*) Statistical relations do not always establish the 'cause and effect" relationship between phenomena.

(*ix*) A statistical enquiry has four phases, viz.,

(*a*) Collection of data;

(*b*) Classification and tabulation of data;

(*c*) Analysis of data;

(*d*) Interpretation of data.

1.3. FUNCTIONS OF STATISTICS

(*i*) It simplifies complex data and helps us to study the trends and relationships of different phenomena and compare them.

(*ii*) It helps us to classify numerical data, measure uncertainty, test the hypothesis, formulate policies and take valid inferences.

1.3.1 Collection of Data

Data means information. Data collected expressly for a specific purpose are called **'Primary data'** e.g., data collected by a particular person or organisation from the primary source for his own use, collection of data about the population by censuses and surveys, etc. Data collected and published by one organisation and subsequently used by other organisations are called **'Secondary data'**. The various sources of collection for secondary data are: newspapers and periodicals; publications of trade associations; research papers published by university departments, U.G.C. or research bureaus; official publications of central, state and the local and foreign governments, etc.

The collection expenses of primary data are more than secondary data. Secondary data should be used with care. The various methods of collection of primary data are: (*i*) Direct personal investigation (interview/observation); (*ii*) Indirect oral investigation; (*iii*) Data from local agents and correspondents; (*iv*) Mailed questionnaires; (*v*) Questionnaires to be filled in by enumerators; (*vi*) Results of experiments, etc. Data collected in this manner are called **'raw data'**. These are generally voluminous and have to be arranged properly before use.

1.4. CLASSIFICATION OF DATA

[C.A. Foun. Nov. 1997)

Connor defined classification as: "the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals".

The raw data, collected in real situations and arranged haphazardly, do not give a clear picture. Thus to locate similarities and reduce mental strain we resort to classification. Classification condenses the data by dropping out unnecessary details. It facilitates comparison between different sets of data clearly showing the different points of agreement and disagreement. It enables us to study the relationship between several characteristics and make further statistical treatment like tabulation, etc.

During population census, people in the country are classified according to sex (males/females), marital status (married/unmarried), place of residence (rural/urban), Age (0–5 years, 6–10 years, 11–15 years, etc.), profession (agriculture, production, commerce, transport, doctor, others), residence in states (West Bengal, Bihar, Mumbai, Delhi, etc.), etc.

1.4.1. Primary Rules of Classification

[C.A. Foun. Nov. 1997]

In quantitative classification, we classify data by assigning arbitrary limits called **class-limits**. The group between any two class-limits is termed as **class or class-interval**. The primary rules of classification are given below:

(*i*) There should not be any ambiguity in the definition of classes. It will eliminate all doubts while including a particular item in a class.

(*ii*) All the classes should preferably have equal width or length. Only in some special cases, we use classes of unequal width.

(*iii*) The class-limits (integral or fractional) should be selected in such a way that no value of the item in the raw data coincides with the value of the limit.

(*iv*) The number of classes should preferably be between 10 and 20, i.e., neither too large nor too small.

(*v*) The classes should be exhaustive, i.e., each value of the raw data should be included in them.

(*vi*) The classes should be mutually exclusive and non-overlapping, i.e., each item of the raw data should fit only in one class.

(*vii*) The classification must be suitable for the object of inquiry.

(*viii*) The classification should be flexible and items included in each class must be homogeneous.

(*ix*) Width of class-interval is determined by first fixing the no. of class-intervals and then dividing the total range by that number.

1.4.2. Modes of Classification

There are four types of classification, viz., (*i*) qualitative; (*ii*) quantitative; (*iii*) temporal and (*iv*) spatial.

(*i*) **Qualitative classification:** It is done according to attributes or non-measurable characteristics; like social status, sex, nationality, occupation, etc. For example, the population of the whole country can be classified into four categories as married, unmarried, widowed and divorced. When only one attribute, e.g., sex, is used for classification, it is called **simple classification**. When more than one attributes, e.g., deafness, sex and religion, are used for classification, it is called **manifold classification**.

(*ii*) **Quantitative classification:** It is done according to numerical size like weights in kg or heights in cm. Here we classify the data by assigning arbitrary limits known as class-limits. The quantitative phenomenon under study is called a variable. For example, the population of the whole country may be classified according to different variables like age, income, wage, price, etc. Hence this classification is often called '**classification by variables**'.

(*a*) **Variable:** A variable in statistics means any measurable characteristic or quantity which can assume a range of numerical values within certain limits, e.g., income, height, age, weight, wage, price, etc. A variable can be classified as either discrete or continuous.

(1) **Discrete variable:** A variable which can take up only exact values and not any fractional values, is called a 'discrete' variable. Number of workmen in a factory, members of a family, students in a class, number of births in a certain year, number of telephone calls in a month, etc., are examples of discrete-variable.

(2) **Continuous variable:** A variable which can take up any numerical value (integral/fractional) within a certain range is called a 'continuous' variable. Height, weight, rainfall, time, temperature, etc., are examples of continuous variables. Age of students in a school is a continuous variable as it can be measured to the nearest fraction of time, i.e., years, months, days, etc.

(*iii*) **Temporal classification:** It is done according to time, e.g., index numbers arranged over a period of time, population of a country for several decades, exports and imports of India for different five year plans, etc.

(*iv*) **Spatial classification:** It is done with respect to space or places, e.g., production of cereals in quintals in various states, population of a country according to states, etc.

1.5. STATISTICAL SERIES (DISCRETE/CONTINUOUS)

Corner defined statistical series as: "If two variable quantities can be arranged side by side so that measurable difference in the one corresponds with measurable difference in the other, the result is said to form a statistical series". Statistical series may be either discrete or continuous (Table 1.1).

ILLUSTRATION 1.1

Table 1.1

Discrete series		Continuous series	
Marks	Number of students	Height [inch]	No. of students
40	12	54–60	15
50	15	60–66	14
60	16	66–72	9
70	7	72–78	12

A discrete series may be formed from items which are exactly measurable. For example, the number of students getting exactly 40, 50, 60, 70 marks can be easily counted. But height or weight cannot be measured with absolute accuracy. Hence the number of students with height exactly 5′–6″ cannot be counted. Exact height will be either (5′–6″+0.01″), or, (5′–6″– 0.01″). Here, we are to count the number of students whose heights fall between 5′–0″ to 5′–6″. Such series are known as 'continuous' series.

Discrete or ungrouped frequency distribution does not condense the data much and is quite cumbersome to grasp and comprehend. It becomes handy if the values of the variable are largely repeated. A discrete variable sometimes may be presented in the form of a continuous frequency distribution when the discrete distribution is too large and unwieldy to handle.

1.6. FREQUENCY DISTRIBUTION

If the value of a variable, e.g., height, weight, etc. (continuous), number of students in a class, readings of a taxi-meter (discrete) etc., occurs twice or more in a given series of observations, then the number of occurrence of the value is termed as the "**frequency**" of that value.

The way of tabulating a pool of data of a variable and their respective frequencies side by side is called a '**frequency distribution**' of those data. Croxton and Cowden defined frequency distribution as "a statistical table which shows the sets of all distinct values of the variable arranged in order of magnitude, either individually or in groups, with their corresponding frequencies side by side".

Let us consider the marks obtained by 100 students of a class in Economics.

Table 1.2: Marks of 100 Students of a Class in Economics

72	61	63	65	62	68	69	64	65	67
69	(56)	60	66	62	57	72	67	65	70
64	66	71	(73)	67	65	64	63	61	58
64	62	69	66	65	63	63	59	61	64
65	57	66	71	68	70	67	66	60	62
65	58	63	68	64	61	62	65	66	59
62	65	65	60	64	61	64	69	62	64
62	63	68	67	65	62	65	68	61	63
62	72	62	66	66	65	63	67	66	63
63	66	65	63	62	62	66	64	62	62

If the raw-data of Table 1.2 are arranged in either ascending, or, descending order of magnitude, we get a better way of presentation, usually called an "array" (Table 1.3).

Table 1.3: Array of Marks Shown in Table 1.2

56	61	62	62	63	64	65	66	67	69
57	61	62	63	63	64	65	66	67	69
57	61	62	63	64	65	65	66	67	70
58	61	62	63	64	65	65	66	68	70
58	61	62	63	64	65	65	66	68	71
59	61	62	63	64	65	65	66	68	71
59	62	62	63	64	65	66	66	68	72
60	62	62	63	64	65	66	67	68	72
60	62	62	63	64	65	66	67	69	72
60	62	62	63	64	65	66	67	69	73

Now let us present the above data in the form of a simple (or, ungrouped) frequency distribution using the tally marks. A tally mark is an upward slanted stroke (/) which is put against a value each time it occurs in the raw data. The fifth occurrence of the value is represented by a cross tally mark (\) as shown across the first four tally marks.

Finally, the tally marks are counted and the total of the tally marks against each value is its frequency.

Let us now represent the data in Table 1.3 as simple (or, ungrouped) frequency distribution.

Table 1.4: Simple Frequency Distribution of Marks of 100 Students

Marks	Tally marks	Frequency	Marks	Tally marks	Frequency
56	/	1	65	~~////~~ ~~////~~ ////	14
57	//	2	66	~~////~~ ~~////~~ /	11
58	//	2	67	~~////~~ /	6
59	//	2	68	~~////~~	5
60	///	3	69	////	4
61	~~////~~ /	6	70	//	2
62	~~////~~ ~~////~~ ~~////~~	15	71	//	2
63	~~////~~ ~~////~~ /	11	72	///	3
64	~~////~~ ~~////~~	10	73	/	1
Total	–	52	Total	–	48
				Total Frequency = 52 + 48 = 100	

1.6.1. Grouped Frequency Distribution

The data in Table 1.3 can be further condensed by putting them into smaller groups, or, classes called "**class-Intervals**". The number of items which fall in a class-interval is called its "**class-frequency**".

The tabulation of raw data by dividing the whole range of observations into a number of classes and indicating the corresponding class-frequencies against the class-intervals, is called "**grouped frequency distribution**".

Let us now represent the data in Table 1.3 as grouped frequency distribution. We find that the lowest value is 56 and the highest value is 73. Thus for approximately 10 classes the difference of values between two consecutive classes will be $\frac{73-56}{10}=\frac{17}{10}=1.7\approx 2$ and the nine class-intervals will be 56–57, 58–59, ..., etc. (Table 1.5).

Table 1.5: Marks Obtained by 100 Students of a Class in Economics

Class–intervals (Marks)	Tally marks	Frequency (No. of students)
56–57	///	3
58–59	////	4
60–61	~~////~~ ////	9
62–63	~~////~~ ~~////~~ ~~////~~ ~~////~~ ~~////~~ /	26
64–65	~~////~~ ~~////~~ ~~////~~ ~~////~~ ////	24
66–67	~~////~~ ~~////~~ ~~////~~ //	17
68–69	~~////~~ ////	9
70–71	////	4
72–73	////	4
	Total	100

Thus the steps in preparing the grouped frequency distribution are:

1. Determining the class intervals.
2. Recording the data using tally marks.
3. Finding frequency of each class by counting the tally marks.

1.6.2. Several Important Terms

(*a*) **Class-limits:** The maximum and minimum values of a class-interval are called upper class-limit and lower class-limit respectively. In Table 1.5 the lower class-limits of nine classes are 56, 58, 60, 62, 64, 66, 68, 70, 72 and the upper class-limits are 57, 59, 61, 63, 65, 67, 69, 71, 73.

(*b*) **Class-mark, or, Mid-value:** The class-mark, or, mid-value of the class-interval lies exactly at the middle of the class-interval and is given by:

(*i*) Class-mark, or, Mid-value = $\dfrac{(\text{lower class limit} + \text{upper class limit})}{2}$,

or, $\dfrac{(\text{lower class boundary} + \text{upper class boundary})}{2}$,

or, Lower class-limit $+\dfrac{1}{2}$(upper class-limit − lower class-limit)

(*c*) **Class boundaries:** Class boundaries are the true-limits of a class interval. It is associated with grouped frequency distribution, where there is a gap between the upper class-limit and the lower class-limit of the next class. This can be determined by using the formula:

$$\text{Lower class boundary} = \text{lower class-limit} - \frac{1}{2}d$$

$$\text{Upper class boundary} = \text{upper class-limit} + \frac{1}{2}d,$$

where d = common difference between the upper class-limit of a class-interval and the lower class-limit of the next higher class interval.

The class-boundaries of the class-intervals of Table 1.5 will be 55.5 – 57.5; 57.5 – 59.5; 59.5 – 61.5; etc., since d = 58 – 57 = 60 – 59 = ...= 1. The class-boundaries convert a grouped frequency distribution (inclusive type) into a continuous frequency distribution.

(*d*) **Type of class-interval:** Different type of class-intervals with their class-limits are given below:

(A) Exclusive type

Upper limit excluded	Lower limit excluded	Upper limit excluded	Upper limit excluded
(*i*) 10–15	(*ii*) Above 10 but not more than 15	(*iii*) 50–100	(*iv*) 30–
15–20	" 15 " " 20	100–150	40–
20–25	" 20 " " 25	150–250	50–
25–30	" 25 " " 30	250–400	60–70

(B) Inclusive type	(C) Open-end type (*i*)	(C) Open-end type (*ii*)	(D) Unequal class-interval
60–69	0–50	Below 50	0–20
70–79	50–100	50–60	20–50
80–89	100–150	60–70	50–100
90–99	150–200	70–80	100–200
	200–250	80 and above	200–450
	250–over		

Class-intervals (A) like 10–15, 15–20, 50–100, 100–150; 30–, 40–; are upper limit exclusive type, i.e., an item exactly equal to 15, 100 and 40 are put in the intervals 15–20, 100–150 and 40–, respectively and not in intervals 0–15, 50–100 and 30–, respectively. Similarly, 15 is included and 10 excluded (lower limit) in "above 10 but not more than 15" class-interval. In the exclusive type the class-limits are continuous, i.e., the upper-limit of one class-interval is the lower limit of the next class-interval and class limits of a class-interval coincide with the class boundaries of that class-interval. It is suitable for continuous variable data and facilitates mathematical computations.

Again class-intervals (B) like 60–69, 70–79, 80–89, etc., are inclusive type. Here both the upper and lower class-limits are included in the class-intervals, e.g., 60 and 69 both are included in the class-interval 60–69. This is suitable for discrete variable data. There is no ambiguity to which an item belongs but the idea of continuity is lost. To make it continuous, it can be written as 59.5–69.5, 69.5–79.5, 79.5–89.5, etc.

In 'open-end' class-interval (C) either the lower limit of the first class- interval, or, upper limit of the last class-interval, or, both are missing. It is difficult to determine the mid-values of the first and last class-intervals without an assumption. If the other closed class-intervals have equal width, then we can assume that the open-end class-intervals also have the same common width of the closed class-intervals.

Grouped frequency distributions are kept open ended when there are limited number of items scattered over a long interval. Unequal class-intervals (D) are preferred only when there is a great fluctuation in the data.

(*e*) **Width or Length (or size) of a Class-interval:** Width of a class-interval = Upper class-boundary − Lower class-boundary

Common width of a class-interval = difference between two successive upper Class-limits (or, two successive lower class-limits) (when the class-intervals have equal widths)

= difference between two successive upper class-boundaries (or, two successive lower class-boundaries)

= difference between two successive class marks, or, mid values

$$\text{Width of the class-interval } = \frac{\text{largest value } - \text{ smallest value}}{\text{no. of class-intervals}}$$

$$= \frac{\text{range}}{\text{no. of class-intervals}} = \frac{\text{range}}{1 + 3.322 \log_{10} \text{N}}$$

where, N = total no. of observations in the data (Formula suggested by M.A. Sturges)

(*f*) **Relative frequency:** Relative frequency $= \dfrac{\text{class frequency}}{\text{total frequency}}$

e.g., the relative frequency of class-interval 62–63 in Table 1.5 is 26/100 = 0.26.

(*g*) **Percentage frequency:**

Percentage frequency of a class-interval $= \dfrac{\text{class frequency}}{\text{total frequency}} \times 100$, e.g., the percentage frequency of above class-interval $= \dfrac{26}{100} \times 100 = 26.$

(*h*) **Frequency density:** Frequency density of a class-interval $= \dfrac{\text{class frequency}}{\text{width of the class}}$

e.g., the frequency density of class-interval 62–63 in Table 1.5 is 26/2 = 13. If the class-intervals of a frequency distribution are of unequal width, the frequency densities of the class-intervals can be used to compare the concentration of frequencies in the class-intervals and construct the histogram.

(*i*) We have defined other related terms under summary.

1.6.3. Two-way, or, Bivariate Frequency Distribution

Often we find data composed of measurements made on two, or, more variates for each item, e.g., weights and heights of a group of boys, ages of wives and husbands for a group of couples, etc. Such data require classification w.r. to characteristics, or, criteria simultaneously. The frequency distribution obtained by this cross classification is called the bivariate frequency distribution.

Example 1.1. (*a*) 25 values of two variables X and Y are given below. Form a two-way frequency table showing the relationship between the two. Take class-intervals of X as 10 to 20; 20 to 30; etc., that of Y as 100 to 200; 200 to 300, etc.

X	12	34	33	22	44	37	26	36	55	48	27
Y	140	266	360	470	470	380	480	315	420	390	440
X	37	21	51	27	42	43	52	57	44	48	48
Y	390	590	250	550	360	570	290	416	380	492	370
X	52	41	69								
Y	312	330	590								

[CA (Inter), May 1980]

(*b*) Also construct the marginal frequency table of X and Y.

(*c*) The conditional distribution of X when Y lies between 300 and 400.

Answer: **(a)** By using the method of tally marks, we can get the cell frequencies, as shown in the bivariate frequency Table 1.6 given below:

Table 1.6: Bivariate Frequency Table

Y↓ \ X→	10–20	20–30	30–40	40–50	50–60	60–70	f_y
100–200	/ (1)	–	–	–	–	–	1
200–300	–	–	/ (1)	\ –	// (2)	–	3
300–400	–	–	//// (4)	//// (5)	/ (1)	–	10
400–500	–	/// (3)	–	// (2)	// (2)	–	7
500–600		// (2)		/ (1)		/ (1)	4
600–700	–	–	–	–	–	–	–
f_x	1	5	5	8	5	1	25

(b) Table 1.7: The Marginal Distributions of X and Y

Marginal distribution (X)		Marginal distribution (Y)	
X	$f = f_x$	Y	$f = f_y$
10–20	1	100–200	1
20–30	5	200–300	3
30–40	*5	300–400	10
40–50	*8	400–500	7
50–60	5	500–600	4
60–70	1	600–700	–
Total	$\Sigma f_x = 25$		$\Sigma f_y = 25 = \Sigma f_x$

Note: f_x = marginal frequencies of X

f_y = " " of Y

* f(X, Y) = the frequencies of the pairs (X, Y).

X and Y = the two variates.

$\Sigma f_x = \Sigma f_y$ = grand total of all the frequencies.

(c) Table 1.8: Conditional Distribution X when Y Lies Between 300 and 400

X	Frequency
10–20	0
20–30	0
30–40	4
40–50	5
50–60	1
60–70	0
Total	10

1.6.4. Steps of Construction of a Bivariate Table

(*i*) Find the class-interval of each of the variables.

(*ii*) Write one of the variables on the left-hand corner of the table and the other on top of the table.

(*iii*) The first item X = 12 (Table 1.6) falls in the class-interval. 10–20 and its Y value 140 falls in the class-interval 100–200. Insert a tally mark in the 1st cell where the column corresponding to X = 12 intersects the row corresponding to Y = 140.

(*iv*) Similarly, insert tally marks for each pair of values (X, Y) for all the 25 sets. The total frequency for each cell is written within brackets ( ) immediately after the tally marks.

(*v*) Next count the frequencies of each row and put the final figure at the extreme right column. Again count the frequencies of each column and put the final figure at the bottom row.

1.7. CUMULATIVE FREQUENCY DISTRIBUTION

A frequency distribution becomes cumulative when the frequency of each class-interval is cumulative. Cumulative frequency of a class-interval can be obtained by adding the frequency of that class-interval to the sum of the frequencies of the preceding class-intervals.

Often we want to know the number of cases which fall below, or, above a certain value. Hence, there are two types of cumulative frequencies, i.e., (1) less than (or, from below) cumulative frequency, and (2) more than (or, from above) cumulative frequencies. In the less than type the cumulative frequency of each class-interval is obtained by adding the frequencies of the given class and all the preceding classes, when the classes are arranged in the ascending order of the value of the variable. In the more than type the cumulative frequency of each class-interval is obtained by adding the frequencies of the given class and the succeeding classes. For grouped frequency distribution, the cumulative frequencies are shown against the class-boundary points.

Example 1.2. Construct the cumulative frequency distribution (both "less than" and "more than" types) from the following data:

Weight (in kg):	10–19	20–29	30–39	40–49	50–59	Total
No. of students:	6	10	25	15	12	68

Answer: Here, d = Gap between two consecutive classes = 1, $\therefore \frac{1}{2}$ d = 0.5

$\therefore$ Lower class-boundary points are 9.5, 19.5, 29.5, etc. and the last upper class-boundary point is 59.5. Hence, the class boundary points are 9.5, 19.5, ..., 59.5.

Table 1.9: Cumulative Frequency Distribution of Weights of 68 Students

Class–boundary points (Weight in kg)	Cumulative Frequency	
	"Less than"	"More than"
9.5	0	68
19.5	6	62
29.5	10 + 6 = 16	52
39.5	6 + 10 + 25 = 41	27
49.5	41 + 15 = 56	12
59.5	56 + 12 = 68	0

Note: No. of students of weights less than 9.5 kg is nil; hence 0 is written against 9.5; no. of students of weights less than 19.5 is 6, hence 6 is written against 19.5 kg and so on. Similarly, the number of students of weights 59.5 kg and above is nil, hence 0 is written against 59.5 kg; number of students of weights 49.5 kg or more is 12, hence 12 is written against 49.5 kg and so on.

Median, Quartiles, Deciles and percentiles can be obtained from cumulative frequency distributions.

1.8. PRESENTATION OF STATISTICAL DATA

Statistical data can be presented in three different ways: (*a*) Textual presentation, (*b*) Tabular presentation, and (*c*) Graphical presentation.

(*a*) **Textual presentation:** This is a descriptive form. The following is an example of such a presentation of data about deaths from industrial diseases in Great Britain in 1935–39 and 1940–44.

Example 1.3. Numerical data with regard to industrial diseases and deaths thereform in Great Britain during the years 1935–39 and 1940–44 are given in a descriptive form:

"During the quinquennium 1935–39, there were in Great Britain 1, 775 cases of industrial diseases made up of 677 cases of lead poisoning, 111 of other poisoning, 144 of anthrax, and 843 of gassing. The number of deaths reported was 20 p.c. of the cases for all the four diseases taken together, that for lead poisoning was 135, for other poisoning 25 and that for anthrax was 30.

During the next quinquennium, 1940–44, the total number of cases reported was 2, 807. But lead poisoning cases reported fell by 351 and anthrax cases by 35. Other poisoning cases increased by 784 between the two periods. The number of deaths reported decreased by 45 for lead poisoning, but decreased only by 2 for anthrax from the pre-war to the post-war quinquennium. In the later period, 52 deaths were reported for poisoning other than lead poisoning. The total number of deaths reported in 1940–44 including those from gassing was 64 greater than in 1935–39".

[C.A. (Final), May 1958]

The **disadvantages** of textual presentation are: (*i*) it is too lengthy; (*ii*) there is repetition of words; (*iii*) comparisons cannot be made easily; (*iv*) it is difficult to get an idea and take appropriate action.

(*b*) **Tabular presentation, or, Tabulation**

[C.A. Foun. Nov. 1996, May 1999]

Tabulation may be defined as the systematic presentation of numerical data in rows or/and columns according to certain characteristics. It expresses the data in concise and attractive form which can be easily understood and used to compare numerical figures. Before drafting a table, you should be sure what you want to show and who will be the reader.

The descriptive form of Example 1.3 has been condensed below in the form of a Table.

Table 1.10: Deaths from Industrial Diseases in Great Britain

Date:..........

Sl. No.	Diseases	1935–39			1940–44		
		Number of cases	Number of deaths	% of deaths	Number of cases	Number of deaths	% of deaths
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Lead poisoning	677	135	19.94	326	90	27.60
2	Anthrax	144	30	20.83	109	28	25.69
3	Gassing	843	165	19.57	1, 477	249	16.86
4	Other poisoning	111	25	22.52	895	52	5.81
	Total	1, 775	355	20.00	2, 807	419	14.93

[*Source:* C.A. (Final), May, 1958]

The **advantages** of a tabular presentation over the textual presentation are: (*i*) it is concise; (*ii*) there is no repetition of explanatory matter; (*iii*) comparisons can be made easily; (*iv*) the important features can be highlighted; and (*v*) errors in the data can be detected.

An ideal statistical table should contain the following items:

(*i*) **Table number:** A number must be allotted to the table for identification, particularly when there are many tables in a study.

(*ii*) **Title:** The title should explain what is contained in the table. It should be clear, brief and set in bold type on top of the table. It should also indicate the time and place to which the data refer.

(*iii*) **Date:** The date of preparation of the table should be given.

(*iv*) **Stubs, or, Row designations:** Each row of the table should be given a brief heading. Such designations of rows are called “stubs”, or, “stub items” and the entire column is called “stub column”.

(*v*) **Column headings, or, Captions:** Column designation is given on top of each column to explain to what the figures in the column refer. It should be clear and precise. This is called a “caption”, or, “heading”. Columns should be numbered if there are four, or, more columns.

(*vi*) **Body of the table:** The data should be arranged in such a way that any figure can be located easily. Various types of numerical variables should be arranged in an ascending order, i.e., from left to right in rows and from top to bottom in columns. Column and row totals should be given.

(*vii*) **Unit of measurement:** If the unit of measurement is uniform throughout the table, it is stated at the top right-hand corner of the table along with the title. If different rows and columns contain figures in different units, the units may be stated along with "stubs", or, "captions". Very large figures may be rounded up but the method of rounding should be explained.

(*viii*) **Source:** At the bottom of the table a note should be added indicating the primary and secondary sources from which data have been collected.

(*ix*) **Footnotes and references:** If any item has not been explained properly, a separate explanatory note should be added at the bottom of the table.

A table should be logical, well-balanced in length and breadth and the comparable columns should be placed side by side. Light/heavy/thick or double rulings may be used to distinguish sub-columns, main columns and totals. For large data more than one table may be used.

(*c*) **Graphical presentation:** Quantitative data may also be presented graphically by using bar charts, pie diagrams, pictographs, line diagrams, etc.

Example 1.4. Draw up a blank table to show the number of employees in a large commercial firm, classified according to (*i*) Sex: Male and Female; (*ii*) Three age-groups: below 30, 30 and above but below 45, 45 and above; and (*iii*) Four income-groups: below Rs. 400, Rs. 400–750, Rs. 750–1, 000, above Rs. 1, 000. **[C.A. May 1963; C.U. B.Com (Hons) 1980]**

Date:.......

Answer:

Table 1.11: Number of Employees in a Large Commercial Firm Classified by Sex, Three Age-Groups and Four Income-Groups

Age–groups	Below 30 (Nos)	30–45 (Nos)	45 and above (Nos)	Total (Nos)
Sex / Income groups	Male Female Total	Male Female Total	Male Female Total	Male Female Total
1. Below Rs. 400				
2. Rs. 400 to Rs. 750				
3. Rs. 750 to Rs. 1000				
4. Above Rs. 1000				
Grand total				

[*Source:* C.A. May 1963]

Example 1.5. (*a*) Industrial finance in India showed great variation in respect of sources during the first, second and third plans. There were two main sources, viz., internal and external. The former had two sources—depreciation and free reserves and surplus. The latter had three sources—capital issues, borrowing and 'other sources'.

During the first plan internal and external sources accounted for 62% and 38% of the total, and in this depreciation, fresh capital and 'other sources' formed 29%, 7% and 10.6% respectively.

During the second plan internal sources decreased by 17.3% and external sources increased by 17.3% as compared to the first plan, and depreciation was 24.5%. The external finance during the same period consisted of fresh capital 10.9% and borrowings 28.9%.

Compared to the second plan, during the third plan external finance decreased by 4.4% and borrowings and other sources were 29.4% and 14.9%. During the third plan, internal finance increased by 4.4% and free reserves and surplus formed 18.6%.

Tabulate the above information with the given details as clearly as possible observing the rules of tabulation.

[C.A. (Inter) Adapted]

(*b*) A survey of 370 students from commerce faculty and 130 students from science faculty revealed that 180 students were studying for only C.A. Examinations; 140 for only Costing Examinations and 80 for both C.A. and Costing Examinations. The rest had offered part-time management courses. Of those studying for Costing only 13 were girls and 90 boys belonged to commerce faculty. Out of 80 studying both C.A. and Costing; 72 were from commerce faculty amongst which 70 were boys. Amongst those who offered part-time management courses; 50 boys were from science faculty and 30 boys and 10 girls from commerce faculty. In all there were 110 boys in science faculty. Out of 180 studying C.A., only, 150 boys and 8 girls were from commerce faculty and 6 girls from science faculty.

Present the above information in a tabular form. Find the number of students from science faculty studying for part-time management courses.

[C.A. (Inter) Adapted]

Date:.......

Answer:

(a) Table 1.12: Industrial Finance Pattern in India (in percentage)

Plans	Sources						
	Internal			External			
	Depre–ciation	Free reserves and surplus	Total	Capital issues	Borrow–ings	Other sources	Total
First	29	33	62	7	20.4	10.6	38
Second	24.5	20.2	44.7	10.9	28.9	15.5	55.3
Third	30.5	18.6	49.1	6.6	29.4	14.9	50.9

[*Source:* C.A. Inter–Adapted]

Date:............

(b) Table 1.13: Number of Students Studying in Different Faculties and Courses

Faculty / Courses	Commerce (Nos)			Science (Nos)			Total (Nos)		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
Part–time Management	30	10	40	50	10	60	80	20	100
C.A.	150	8	158	16	6	22	166	14	180
Costing	90	10	100	37	3	40	127	13	140
C.A. and Costing	70	2	72	7	1	8	77	3	80
Total	340	30	370	110	20	130	450	50	500

[*Source:* C.A. Inter Adapted]

Note: One (simple), two, or, three (complex)—way tables classify the data w.r. to one, two, or, three characteristics respectively. Table (1.11) is a three-way table where the characteristics are age, sex and income.

1.8.1. Objectives of Tabulation

The main objectives of tabulation are stated below:

(*i*) to carry out investigation; (*ii*) to do comparison; (*iii*) to locate omissions and errors in the data; (*iv*) to use space economically; (*v*) to study the trend; (*vi*) to simplify data; (*vii*) to use it as future reference.

1.8.2. Sorting

Sorting of data is the last process of tabulation. It is a time-consuming process when the data is too large. After classification the data may be sorted using either of the following methods:

(*i*) **Manual method:** Here the sorting is done by hand by giving tally marks for the number of times each event has occurred. Next the total tally marks are counted. The method is simple and suitable for limited data.

(*ii*) **Mechanical and electrical method:** To reduce the sorting time mechanical devices may be used. This is described as mechanical tabulation.
For electrical tabulation data should be codified first and then punched on card. For each data a separate card is used. The punched cards are checked by a machine called '**verifier**'. Next the cards are sorted out into different groups as desired by a machine called '**sorter**'. Finally, the tabulation is done by using a tabulator. The same card may be sorted out more than once for completing tables under different titles.

(*iii*) **Tabulation using electronic computer:** It is convenient to use electronic computer for sorting when (*a*) data are very large; (*b*) data have to be sorted for future use and (*c*) the requirements of the table are changing. Such a tabulation is less time-consuming and more accurate than the manual method.

1.9. TYPICAL EXAMPLES

Example 1.6. Represent the following information in a suitable tabular form with proper rulings and headings:

The annual report of the Ishapore Public Library reveals the following points regarding the reading habits of its members.

Out of the total 3, 713 books issued to the members in the month of June 1983, 2, 100 were fictions. There were 467 members of the library during the period and they were classified into five classes A, B, C, D, and E. The number of members belonging to the first four classes were respectively 15, 176, 98 and 129; and the number of fiction issued to them were 103, 1, 187, 647, and 58 respectively. Number of books, other than text books and fictions, issued to these four classes of members were respectively 4, 390, 217 and 341. The textbooks were issued only to members belonging to the classes C, D and E and the number of textbooks issued to them were respectively 3, 317 and 160.

During the same period, 1, 246 periodicals were issued. These included 396 technical journals of which 36 were issued to members of class B, 45 to class D and 315 to class E.

To members of the classes B, C, D and E the number of other journals issued were 419, 26, 231 and 99 respectively.

The report, however, showed an increase by 3.9% in the number of books issued over last month, though there was a corresponding decrease by 6.1% in the number of periodicals and journals issued to members.

[ICWAI June 1977, June 1983]

Answer:

Table 1.14: Statistics Showing Reading Habits of Members of Ishapore Public Library, June 1983

Classes	No. of books issued					Periodicals and journals (Nos)		
	No. of members	Fictions	Text books	Other books	Total	Technical	Other journals	Total
A	15	103	–	4	107	–	75	75
B	176	1, 187	–	390	1, 577	36	419	455
C	98	647	3	217	867	–	26	26
D	129	58	317	341	716	45	231	276
E	49	105	160	181	446	315	99	414
Total	467	2, 100	480	1, 133	3, 713	396	850	1, 246
Last* Month (May)	–	2, 021	462	1, 091	3, 574	422	905	1, 327

Note: *Figures for last month (May 1983)

(*a*) **Books:** (*i*) No. of books issued $= \dfrac{3{,}713 \times 100}{103.9}$ = 3, 574 nos (approx)

(*ii*) No. of fictions issued $= \dfrac{2{,}100 \times 100}{103.9}$ = 2, 021 nos (approx)

(*iii*) No. of textbooks issued $= \dfrac{480 \times 100}{103.9}$ = 462 nos (approx).

$\therefore$ Total no. of other books issued = 3, 574 – (462 + 2, 021) = 1, 091 nos

(*b*) **Periodicals:** (*i*) No. of periodicals issued $= \dfrac{1{,}246 \times 100}{93.9}$ = 1, 327 Nos (Approx)

(*ii*) No. of Technical Journals issued $= \dfrac{396 \times 100}{93.9}$ = 422 nos (approx)

(*iii*) No. of other Journals issued = (1, 327 – 422) = 905 nos (approx).

Example 1.7. The total number of accidents in Southern Railway in 1960 was 3, 500, and it decreased by 300 in 1961 and by 700 in 1962. The total number of accidents in metre gauge section showed a progressive increase from 1960 to 1962. It was 245 in 1960, 346 in 1961; and 428 in 1962. In the metre gauge section, "not compensated" cases were 49 in 1960, 77 in 1961, and 108 in 1962. "Compensated" cases in the broad gauge section were 2, 867, 2, 587 and 2, 152 in these three years respectively.

From the above report, you are required to prepare a neat table as per the rules of tabulation.

[C.A. Nov. 1971]

Answer:

Table 1.15: Number of Accidents in Southern Railway from 1960 to 1962

Section	1960 (Nos)			1961 (Nos)			1962 (Nos)		
	N	C	Total	N	C	Total	N	C	Total
Metre gauge	49	196	245	77	269	346	108	320	428
Broad gauge	388	2, 867	3, 255	267	2, 587	2, 854	220	2, 152	2, 372
Total	437	3, 063	3, 500	344	2, 856	3, 200	328	2, 472	2, 800

Example 1.8. You are given below the wages paid to some workers in a small factory. Form a frequency distribution with class-interval of 10 paise;

Wages in Rs.

1.10	1.13	1.44	1.44	1.27	1.17	1.98	1.36	1.30
1.27	1.24	1.73	1.51	1.12	1.42	1.03	1.58	1.46
1.40	1.21	1.62	1.31	1.55	1.33	1.04	1.48	1.20
1.60	1.70	1.09	1.49	1.86	1.95	1.51	1.82	1.43
1.29	1.54	1.38	1.87	1.41	1.77	1.15	1.57	1.07
1.65	1.36	1.67	1.41	1.55	1.22	1.69	1.67	1.34
1.45	1.39	1.25	1.26	1.75	1.57	1.53	1.37	1.59
1.19	1.52	1.56	1.32	1.81	1.40	1.47	1.38	1.62
1.76	1.28	1.92	1.46	1.46	1.35	1.16	1.42	1.78
1.68	1.47	1.37	1.35	1.47	1.43	1.66	1.56	1.48

[C.A. May 1967]

Answer: From the inspection of the above observations, we find that largest and smallest values are 1.98 and 1.03 respectively.

$\therefore$ Number of the class–intervals $= \dfrac{\text{largest value} - \text{smallest value}}{\text{width of the class - interval}}$

$$= \frac{1.98 - 1.03}{0.10} = 9.5 \approx 10$$

$\therefore$ The class limits may be taken as 1.01–1.10; 1.11–1.20; ...... ; (1.91–2.00).

Table 1.16: Frequency Distribution of Wages

Class limits (Wages Rs.)	Tally marks	Frequency
1.01–1.10	~~||||~~	5
1.11–1.20	~~||||~~ //	7
1.21–1.30	~~||||~~ ~~||||~~	10
1.31–1.40	~~||||~~ ~~||||~~ ~~||||~~	15
1.41–1.50	~~||||~~ ~~||||~~ ~~||||~~ ///	18
1.51–1.60	~~||||~~ ~~||||~~ ////	14
1.61–1.70	~~||||~~ ////	9
1.71–1.80	~~||||~~	5
1.81–1.90	////	4
1.91–2.00	///	3
Total		90

Example 1.9. Calculate: (*a*) the number of cases between 112 and 134; (*b*) number less than 112; (*c*) number greater than 134, from the following:

Class limit :	90–100	100–110	110–120	120–130	130–140	140–150	150–160
Frequency :	16	22	45	60	50	24	10

[C.A. May 1969]

Answer:

Table 1.17: Cumulative Frequency Distribution

Class–boundary	Cumulative frequency (less than)
90	0
100	16
110	38
112 →	← x
120	83
130	143
134 →	← y
140	193
150	217
160	227

Now, $\dfrac{x - 38}{83 - 38} = \dfrac{112 - 110}{120 - 110}$ or, $x = \dfrac{2}{10} \times 45 + 38 = 47$ (Simple interpolation)

Again, $\frac{y-143}{193-143} = \frac{134-130}{140-130}$ or, $y = \frac{4}{10} \times 50 + 143 = 163$

∴ (*a*) No. of cases between 112 and 134 is 163 − 47 = 116,

(*b*) No. less than 112 is 47,

(*c*) No. greater than 134 is 227 − 163 = 64.

Example 1.10. Using Sturges' rule, $K = 1 + 3.322 \log_{10} N$,
where K = no. of class-intervals, N = total no. of observations;
classify, in equal intervals, the following hours worked by 20 workmen in a factory for one month.

155	120	50	110	116
95	125	42	175	130
160	90	68	71	135
147	115	108	140	98

Find the percentage frequency in each class-interval.

Answer: Now, $K = 1 + 3.322 \log_{10} 20$

$= 1 + 3.322 \times 1.3010 = 5.3219 \approx 6$ = No. of class-intervals

∴ Width of class-interval $= \frac{\text{largest value} - \text{smallest value}}{\text{No. of the class - interval}}$

$$= \frac{175-42}{6} = \frac{133}{6} = 22.17 \approx 23.$$

Hence the classes by exclusive type are:

40–63, 63–86, 86–109, 109–132, 132–155, 155–178

The frequency distribution is shown below:

Table 1.18: Frequency Distribution of Hours of 20 Workmen of a Factory

Class-intervals	Tally marks	Frequency (f)	Percentage frequency
40–63	//	2	$\frac{2}{20} \times 100 = 10$
63–86	//	2	10
86–109	////	4	20
109–132	~~////~~ /	6	30
132–155	///	3	15
155–178	///	3	15
Total		N = 20	100