6

Classification and Tabulation

Introduction

The collected data in any statistical investigation are known as raw data. They are huge and confusive. As such they cannot be easily understood by persons, and are not fit for further analysis and interpretation. Prof. J.R. Hicks points out that "Classified and arranged facts speak themselves; unarranged, they are as dead as mutton." Hence after having collected and edited the data, the next important step is to organise it in a systematic manner.

The first step in the analysis and interpretation of data is classification and tabulation. Classification is the first step in tabulation, even though the phrase 'classification and tabulation' is used. Proper classification helps proper tabulation.

Meaning of Classification

Classification is the process of arranging the available facts into homogeneous groups or classes according to resemblances and similarities. The following are the definitions:

"Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities, and giving expression to the unity of attributes that may subsist amongst a diversity of individuals..." —*R.L. Connor*

"Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different but related parts..." —*Secrist*

"The process of grouping a large number of individual facts or observations on the basis of similarity among the items, is called classification." —*Stockton and Clark*

Chief Characteristics of Classification are:

1. All the facts are classified into homogeneous groups by the process of classification.
2. The basis of classification is unity in diversity.
3. Classification may be either real or imaginary.
4. The classification may be according to either similarities or dissimilarities.

5. It should be flexible to accommodate adjustments.

Objects of Classification

The chief objectives of classification are:

1. to condense the mass of data
2. to present the facts in a simple form
3. to bring out clearly the points of similarity and dissimilarity.
4. to facilitate comparison.
5. to bring out the relationship.
6. to prepare data for tabulation.
7. to facilitate the statistical treatment of the data.
8. to facilitate easy interpretation.
9. to eliminate unnecessary details.

Rules of Classification

It is important that classification should possess the following guiding principles.

(*a*) *Exactness*. The classes should be rigidly defined. They should not lead to any ambiguity or confusion.

(*b*) *Mutually exclusive*. Each item of data must find its place in one class. The classes must not overlap.

(*c*) *Stability*. Only one principle must be maintained (*i.e.*, the same pattern of classification) throughout the analysis. Then only it will facilitate meaningful comparison and become an ideal classification.

(*d*) *Flexibility*. The classification should be flexible and easy to adjust to new situations and circumstances.

(*e*) *Suitability*. The classification should be suitable for the object of the enquiry.

(*f*) *Homogeneity*. The items included in each class must be homogeneous; for example, a classification into employed and unemployed youth is not adequate to judge the effect of education; but further, each of them may then be classified into literate and illiterate.

(*g*) *Mathematical accuracy*. Items included in total and subtotals of each class and sub-class must be the same. Therefore, mathematical accuracy is very important in the classification of data.

TYPES OF CLASSIFICATION

The classification of data primarily depends on the purpose and objectives of the enquiry. There are four important types of classification. They are:

1. Geographical *i.e.*, areawise or regionwise or districtwise.
2. Chronological or historical *i.e.*, on the basis of time.
3. Qualitative by character or by attributes.
4. Quantitative or numerical or by magnitudes.

1. **Geographical classification (spatial).** In geographical or spatial classification, the basis of classification is the geographical or locational differences between various items in the statistical data like states, districts, cities, talukas, regions, zone, area, etc. Geographical classification is illustrated in the following table:

Sales data (of pressure cookers) for 2002 (T.N.)

Name of town	*Number of cookers*
Madras	15,000
Tiruchi	13,000
Madurai	11,000
Coimbatore	8,000
Kanyakumari	4,000

2. **Chronological classification.** This type of statistical data is classified according to the time of its occurrence, such as years, months, weeks, days, hours, etc. For example, census data are expressed in decades, national income is expressed every year, departmental sales are expressed every month or week.

Time series are also called chronological classification. They are further classified into the period of time and at the point of time. Statistical data regarding population, imports, exports, sales in a firm, etc., also come under this classification.

Chronological Classification is illustrated below:

Population of India from 1921 to 1981

Year	*Population (in million)*
1921	248
1931	276
1941	313
1951	357
1961	438
1971	536
1981	684

3. **Qualitative classification.** When the data are classified according to some quality or attributes, such as sex, honesty, intelligence, literacy, blindness, colour, deafness, religion, marital status, etc., the classification is termed as qualitative or descriptive

attributes. In this type we can only find out the presence or absence of the attributes, in the given units.

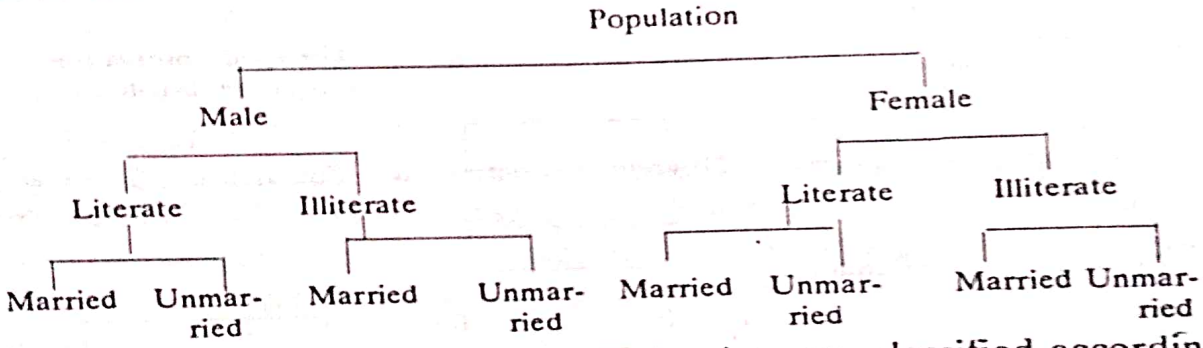
This again can be classified into two types:

(*a*) Simple classification.

(*b*) Manifold classification.

(*a*) **Simple classification.** If the data are classified into only two classes, such as literate and illiterate or honest and dishonest or skilled and unskilled, the classification is termed as simple classification. This classification is normally dichotomy or twofold; for example,

(*b*) **Manifold classification.** In manifold classification, the universe is classified on the basis of more than one attribute at a time; for example, we may first divide the population into males and females on the attribute of sex; then further divide them on the basis of literacy and so on:



4. **Quantitative classification.** If the data are classified according to some characteristic which is capable of quantitative measurement like age, income, height, weight, price, production, sales, profits, *etc.*, it is called quantitative classification or classification according to variables. Variable is the quantitative phenomenon under study.

Marks	*No. of students*
10—20	10
20—30	7
30—40	13
40—50	18
50—60	12

In the above classification, there are two elements, viz.

(1) the VARIABLE *i.e.*, the marks in the above example and

(2) the FREQUENCY i.e., the number of students in each class. There are ten students who scored marks between 10 and 20, seven students scored marks between 20 and 30 and so on.

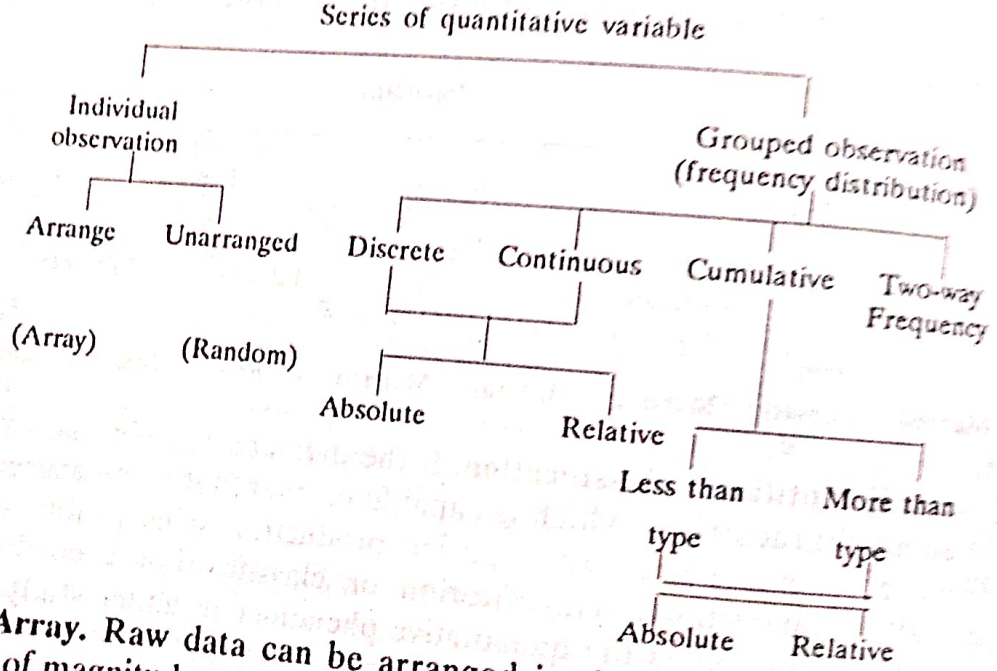
STATISTICAL SERIES

Statistical series are prepared to present the collected and classified data in a properly arranged way. In the words of Secrist, "A series, as used statistically, may be defined as things or attributes of things arranged according to some logical order." For example, if the data pertaining to the height of 10 students are put in a systematic way, it can be called statistical series.

According to L.R.Conner "If two variable quantities can be arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other the result is said to form a statistical series".

TYPES OF SERIES

The following chart brings out clearly the various types of series based on numerical or quantitative values.



Array. Raw data can be arranged in descending or ascending order of magnitude, and it is called an array. But it does not reduce the volume of data.

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 the data are grouped into classes and the number of cases which fall in each class is recorded.

Basically, frequency distribution can be of two kinds :

(1) Univariate Frequency Distribution

(2) Bivariate Frequency Distribution (Two-Way Frequency Distribution)

Again, Univariate Frequency Distribution is of 3 types :

(1) Series of Individual Observations

(2) Discrete Frequency Distribution

(3) Continuous Frequency Distribution.

(A) Individual Observation

Series of Individual observations is a series where items are listed singly after observation, as distinguished from listing them in groups. For some statistical calculations, the series of individual observation are to be arranged in either ascending or descending order. This is called an array.

Roll Nos.	*Marks*	*Roll Nos.*	*Marks*
1	40	6	48
2	33	7	44
3	27	8	51
4	38	9	39
5	41	10	55

The data in the above form is called raw or disorganised data. The above presentation of data in its raw form does not give us any useful information and is rather confusing to the mind. A better presentation of the above raw data would be to arrange them in an ascending or descending order of magnitude which is called the "arraying" of the data. However, this presentation (arraying), though better than the raw data does not reduce the volume of the data. The above illustration is arranged in ascending and descending order in the table.

(B) Discrete (Ungrouped) Frequency Distribution

In a discrete series the data are presented in a way that exact measurements of units are clearly indicated. There is definite difference between the variables of different groups of items. Each class is distinct and separate from the other classes. Non-continuity from one class to another exists.

We have to count the number of times each value of the variable is repeated in the data and it is called the frequency of that class. Boddington says, "Discrete variable is one where the variates differ from each other by definite amounts". For example,

Arrayine

Observed Values	Ascending Order	Descending Order
40	27 .. Lowest	55.. Highest
33	33	51
27	38	48
38	39	44
41	40	41
48	41	40
44	44	39
51	48	38
39	51	33
55	55.. Highest	27.. Lowest

Number of children	Number of families
0	12
1	84
2	110
3	65
4	29
Total	300

Making a Frequency Table

Data may be given in the form of individual observation. They are to be converted into discrete frequency distribution.

Steps

We should form a table with three headings, *viz.*, variable, tally-marks and frequency.

In the first column we place all possible values of the variable. In the second column vertical bar (I) called tally-mark is put against the number. After a particular value has occurred four times, for the fifth occurrence we put a cross tally-mark (IIII) cutting the first four tally-marks; and this gives as a block of 5. For the sixth item we put another tally-mark leaving some space. By putting cross tally-marks, and allowing little space after a block of five, easy and correct counting is facilitated. Finally we count the number of bars corresponding to each value of the variable and place it in the column entitled frequency.

Illustration : 1

Consider the marks scored by 30 students:

9	7	5	3	4	8	6	0	6	5
9	1	7	2	3	8	6	8	7	4
9	4	5	10	6	5	9	6	9	5

We are unable to understand the significance of marks scored by the 30 students, as it is given in raw form. We have to form discrete series out of the above data. First, we note down the lowest and highest values. In the first column we place all possible values of the variables. In the second column a vertical bar (I), called tally mark is put against the number (variable) whenever it occurs. After a particular value occurred four times, for the fifth occurrence, we put a cross tally marks (IIII) or (IIII) , cutting the first four tally marks and this gives us a block of five.

Marks	*Tally-sheet*	*Number of students (Frequency)*
0	I	1
1	I	1
2	I	1
3	II	2
4	III	3
5	IIII	5
6	IIII	5
7	III	3
8	III	3
9	IIII	5
10	I	1
	Total	30

Marks	*Frequency*
0	1
1	1
2	1
3	2
4	3
5	5
6	5
7	3
8	3
9	5
10	1
Total	30

(C) Continuous or Grouped Frequency Distribution

Continuous series is one where measurements are only approximations and are expressed in class intervals, *i.e.*, whithin certain limits. According to Boddington, "the variable which can take any intermediate value between the smallest and longest value in the distribution." In a continuous frequency distribution the class intervals theoretically continue from the beginning of the frequency distribution to the end without break. The continuous frequency distribution can always be distinguished from the discrete frequency distribution in that it will contain two limits–upper limits and lower limits of each class interval – while the discrete frequency distribution will possess only one list of classification of values.

A collection of items, which cannot be exactly measured, but placed within certain limits, is called continuous series. The following technical terms are important, when a continuous frequency distribution is formed or data are classified according to class-intervals.

Class-limits. The class-limits are the smallest or the lowest and the largest or the highest values in the class. For example, take the class 10–20. The lowest value is 10 and the highest value is 20. The two boundaries of the class are known as the lower limit and upper limit of the class. Class limit is also known as class boundaries.

Class-intervals. The difference between the lower limit and the upper limit of the class is known as the class-interval; for example in the class 10–20, the class-interval is 10 (*i.e.*, 20–10). The formula to find the class-interval of a given problem is:

$$i = \frac{L-S}{k}$$

L = largest item

S = smallest item

k = the number or classes

For example, if the marks of 50 students are varied between 10 and 80 and if we want to form 7 classes, then the class-interval would be:

$$i = \frac{L-S}{k} \quad L = 80, S = 10, k = 7$$

$$\frac{80-10}{7} = \frac{70}{7} = 10$$

Therefore, the class-interval would be 10–20, 20–30, 30–40, 40–50, 50–60, 60–70, and 70–80.

There are two methods of forming class-intervals (1) exclusive method and (2) inclusive method.

1. Exclusive method. (Overlapping).

cell.

TABULATION OF DATA

Meaning

By tabulation we mean, a systematic presentation of numerical data in columns and rows in accordance with some salient features or characteristics. Columns are vertical arrangement and rows are horizontal arrangement. Croxton and Cowden state that "Either for one's own use or for the use of others, the data must be presented in a suitable form." It facilitates comparison. It also facilitates analysis.

Definition

According to Prof. H. Secrist, "Tables are a means of recording in permanent form the analysis that is made through classification and of placing in juxtaposition things that are similar and should be compared." In the words of Prof. Neiswanger, "A statistical table is a systematic organisation of data in columns and rows." Tabulation, is the process of presenting data in tables.

Objects

Tabulation helps in understanding complex numerical data and makes them in a simple and clear way that their similar and dissimilar facts are separated. According to D.W. Peden and E.F. Lindquist, "The purpose of a table is to summarise a mass of numerical information and to present it in the simplest possible form, consistent with the purpose for which it is to be used." In the words of R. Gregory and H. Ward, "Tabulation is the process of condensing classified data in the form of a table, so that it may be more easily understood and any comparisons involved may be more readily made." The tabulation is a medium of communication of great economy and effectiveness for which ordinary prose is inadequate. In addition its function in simple presentation, the statistical table is a useful tool of analysis.

"A good statistical table is not a mere careless grouping of columns and rows of figures; it is a triumph of ingenuity and technique, a master-piece of economy of space combined with a maximum of clearly presented information. To prepare a first class table, one must have a clear idea of the facts to be presented, the contrasts to be stressed, the points upon which emphasis is to be placed and lastly a familiarity with the technique of preparation." Harry Jerome

The main objectives of tabulation are :

1. to clarify the object investigation.
2. to simplify complex data.
3. to clarify the characteristics of data.
4. to present facts in the minimum of space.
5. to facilitate comparison.
6. to detect errors and omission in the data.
7. to depict trend and tendencies of the problem under consideration.
8. to facilitate statistical processing.
9. to help reference.

Difference between Classification and Tabulation

Classification and tabulation are important processes in statistical investigation. Through these processes, the collected data are summarised and put in a systematic order.

1. Both classification and tabulation are important for statistical investigation. First the data are classified; then they are presented in tables; classification is the basis for tabulation.

2. Tabulation is a mechanical function of classification, because in tabulation classified data are placed in columns and rows.

3. Classification is a process of statistical analysis; tabulation is a process of presenting data in suitable structure.

Parts of Tabulation

A good statistical table is an art. The following parts must be present in all tables:

1. Table number
2. Title
3. Head note
4. Caption
5. Stubs
6. Body of the table
7. Foot-note
8. Source-note.

1. **Table number.** A table should always be numbered for identification and reference in the future. Each column should also be numbered as shown in the illustration.

2. **Title of the table.** Each table should be given a suitable title. It must be written on the top of table. It must describe the contents of the table. It must explain (1) what the data are (2) where the data are (3) time or period of data (4) how the data are classified, etc.

3. **Head note.** It is a statement, given below the title and enclosed in brackets; for example, the unit of measurement is written as a head-note, such as 'in millions' or 'in crores'.

4. **Captions.** These are headings for the vertical columns. They must be brief and self-explanatory. They have main heading and sub-headings and must be written in small letters.

5. **Stubs.** These are the headings or designation for the horizontal rows. Stubs are wider than columns.

6. **Body of the table.** It contains the numerical information. It is the most important part of the table. The arrangement in the body is generally from left to right in rows and from top to bottom in columns.

7. **Foot-note.** If any explanation or elaboration regarding any item is necessary, foot notes should be given.

8. **Source-note.** It refers to the source from where information has been taken. It is useful to the reader to check the figures and gather additional information.

STRUCTURE OF A TABLE

Number

Title

(Head-note if any)

Stub Heading.	Caption			Total
	Col. Heading	Col. Heading	Col. Heading	
Stub entries	B O D Y			
Total				

Foot-note: *Source*

Rules for Tabulation

The construction of a good statistical table is a specialised art and requires great skill, experience and common sense on the part of the tabulator. There is no hard and fast rule regarding it. According to Bowley, "In collection and tabulation, common sense is the chief requisite, and experience, the chief teacher." These rules may be divided into two groups : (a) Rules rèlating to table structure, which is explained earlier and (b) General rules.

General Rules

1. The table should be simple and compact. It should not be overloaded with details.

2. The captions and stubs in the tables should be arranged in a systematic manner. It must be easy to read the important items. There are many types. They are alphabetical, chronological, geographical, conventional, etc.

3. It should suit the purpose of the investigation.

4. The unit of measurements should be clearly defined and given in the tables; for example, height in metres, weight in kilograms, etc.

5. Figures may be rounded off to avoid unnecessary details in the table. But a foot-note must be given to this effect.

6. Suitable approximation may be adopted.

7. A miscellaneous column should be added to include unimportant items.
8. A table should be complete and self-explanatory.
9. A table should be attractive to draw the attention of readers.
10. As it forms a basis for statistical analysis, it should be accurate and free from all sorts of errors.
11. Abbreviations should be avoided.
12. Do not use ditto marks that may be mistaken.
13. Proper lettering will help to adjust the size of the table.
14. If it is a big table, it will lose its simplicity and understandability; and in such a case break it into two or three tables.

TYPES OF TABLES

Statistical tables can be classified into a number of ways. There are many categories depending upon : (1) The basis of coverage which can be further classified into simple table and complex table. A complex table can be classified into twofold, threefold, or manifold table. (2) The basis of objective or purpose. This can be further classified into general purpose table or reference table and special purpose table or summary table. (3) The basis of nature of enquiry, which can further be classified into original or primary table and derived or derivative table.

1. On the basis of coverage

Simple and Complex

In a simple table the data are classified according to only one characteristic. It is termed as one way or single table and it takes form of frequency table. In a complex table two or more characteristics are shown. It is more popular, because it helps appropriate consideration of all related facts.

Simple table :

Distribution of Marks

Class marks	*No. of students*
20—30	10
30—40	18
40—50	22
Total	50

Two-way table

If the caption or stub is classified into two characteristics and if it gives information of two interrelated questions, then such a table is called two-way table; for example,

Distribution of Marks (Girls & Boys)

Class marks	*Number of students*		
	Boys	*Girls*	*Total*
20–30	6	4	10
30–40	8	10	18
40–50	10	12	22
Total	24	26	50

Three-way table

In this type of table three characteristics are shown. It gives information regarding three interrelated characteristics of a phenomenon; for example.

Distribution of Population by Age, Sex and Literacy

	Males			*Females*			*Total*		
Age group (years)	*Liter-ate*	*Illiter-ate*	*Total*	*Literate*	*Illiterate*	*Total*	*Literate*	*Illiterate*	*Total*
0–18									
18–25									
25–35									
35–45									

A large number of interrelated problems or characteristics are represented in the same table; for example, the distribution of students in a college according to faculty, class, sex and residence.

Manifold or Higher order table

Number of students in M.K. University
(according to faculty, age, sex and residence)

Faculty age group (year)	Students						Total			Total
	Boys			Girls						
	Hosteller	Day Scholar	Total	Hosteller	Day Scholar	Total	Hosteller	Day Scholar	Total	
Commerce										
20—25										
25—30										
above 30										
Arts										
20—25										
25—30										
above 30										
Science										
20—25										
25—30										
above 30										
Law										
20—25										
25—30										
above 30										

2. On the Basis of Objective (Purpose)

General purpose table

It is also known as informative table and provides information for general use; and usually in chronological order. The detailed table in the census reports are of this kind. Govt. agencies prepare this type of tables. These are used by research workers and statisticians. These are placed in the appendix of a report for reference.

Special purpose table

It is also called a summary table or text table or analytical table or derivative table or derived table. It presents the data relating to a particular or a special purpose. Ratios, percentages, etc., are used to facilitate comparison.

3. On the Basis of Originality

The statistical table may be classified into (1) primary table (2) derived table.

In primary table (original), the statistical forms are expressed in original. It contains actual and absolute figures. In a derived table, figures and results are derived from the primary data. It presents totals, percentages, ratios, averages, dispersion, coefficient of correlation, etc. Both primary and derived tables are generally used in practice: