Methods of Data Collection, Organization and Presentation

Before any statistical work can be done data must be collected. Depending on the type of variable and the objective of the study different data collection methods can be employed.

Data Collection Methods

Data collection techniques allow us to systematically collect data about our objects of study (people, objects, and phenomena) and about the setting in which they occur. In the collection of data we have to be systematic. If data are collected haphazardly, it will be difficult to answer our research questions in a conclusive way.

Various data collection techniques can be used such as:

1- Observation
2- Face-to-face and self-administered interviews
3- Postal or mail method and telephone interviews
4- Using available information
5- Focus group discussions (FGD)
6- Other data collection techniques – Rapid appraisal techniques, life histories, case studies, etc.

The statistical data may be classified under two categories, depending upon the sources. (1) Primary data (2) Secondary data.

Primary Data: are those data, which are collected by the investigator himself for the purpose of a specific inquiry or study. Such data are original in character and are mostly generated by surveys conducted by individuals or research institutions.

Secondary Data: When an investigator uses data, which have already been collected by others, such data are called "Secondary Data". Such data are primary data for the agency that collected them, and become secondary for someone else who uses these data for his own purposes.

The selection of the method of data collection is also based on practical considerations, such as:

1) The need for personnel, skills, equipment, etc. in relation to what is available and the urgency with which results are needed.

2) The acceptability of the procedures to the subjects - the absence of inconvenience, unpleasantness, or untoward consequences.

3) The probability that the method will provide a good coverage, i.e. will supply the required information about all or almost all members of the population or sample.

Methods of data organization and presentation

The data collected in a survey is called ***raw data***. In most cases, useful information is not immediately evident from the mass of unsorted data. Collected data need to be organized in such a way as to condense the information they contain in a way that will show patterns of variation clearly. Precise methods of analysis can be decided up on only when the characteristics of the data are understood. For the primary objective of this different techniques of data organization and presentation like **order array**, **tables** and **diagrams** are used.

Frequency Distributions

For data to be more easily appreciated and to draw quick comparisons, it is often useful to arrange the data in the form of a table, or in one of a number of different graphical forms.

When analyzing voluminous data collected from say, a health center's records, it is quite useful to put them into compact tables. Quite often, the presentation of data in a meaningful way is done by preparing a frequency distribution.

Array (ordered array) is a serial arrangement of numerical data in an ascending or descending order. This will enable us to know the range over which the items are spread and will also get an idea of their general distribution. Ordered array is an appropriate way of presentation **when the data are small in size (usually less than 20).**

A study in which 400 persons were asked how many full-length movies they had seen on television during the preceding week. The following gives the distribution of the data collected.

Number of movies	Number of persons	Relative frequency %
0	72	18.0
1	106	26.5
2	153	38.3
3	40	10.0
4	18	4.5
5	7	1.8
6	3	0.8
7	0	0.0
8	1	0.3
Total	400	100.0

In the above distribution Number of movies represents the variable under consideration, Number of persons represents the frequency, and the whole distribution is called frequency distribution particularly simple frequency distribution.

Frequency distributions present data in a relatively compact form, gives a good overall picture, and contain information that is adequate for many purposes, but there are usually some things which can be determined only from the original data. For instance, the above grouped frequency distribution cannot tell how many of the arrested persons are 19 years old, or how many are over 62.

The construction of grouped frequency distribution consists essentially of four steps:

(1) Choosing the classes, (2) sorting (or tallying) of the data into these classes, (3) counting the number of items in each class, and (4) displaying the results in the form of a chart or table.

Choosing suitable classification involves choosing the number of classes and the range of values each class should cover, namely, from where to where each class should go. Both of these choices are arbitrary to some extent, but they depend on the nature of the data and its accuracy, and on the purpose the distribution is to serve. The following are some rules that are generally observed:

1) We seldom use fewer than 6 or more than 20 classes; and 15 generally is a good number, the exact number we use in a given situation depends mainly on the number of measurements or observations we have to group A guide on the determination of the number of classes (k) can be the Sturge's Formula, given by:

K = 1 + 3.322 × log(n), where **n** is the number of observations

And the length or width of the class interval (**w**) can be calculated by:

W = (Maximum value – Minimum value)/K = Range/K

2) We always make sure that each item (measurement or observation) goes into one and only one class, i.e. classes should be mutually exclusive. To this end we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into possible gaps between successive classes, and that the classes do not overlap, namely, that successive classes have no values in common.

Note that the Sturges rule should not be regarded as final, but should be considered as a guide only. The number of classes specified by the rule should be increased or decreased for convenient or clear presentation.

3) **Determination of class limits:** (i) Class limits should be definite and clearly stated. In other words, open-end classes should be avoided since they make it difficult, or even impossible, to calculate certain further descriptions that may be of interest. These are classes like less than 10, greater than 65, and so on. (ii) The starting point, i.e., the lower limit of the first class be determined in such a manner that frequency of each class get concentrated near the middle of the class interval. This is necessary because in the interpretation of a frequency table and in subsequent calculation based up on it, the mid-point of each class is taken to represent the value of all items included in the frequency of that class.

It is important to watch whether they are given to the nearest inch or to the nearest tenth of an inch, whether they are given to the nearest ounce or to the nearest hundredth of an ounce, and so forth. For instance,

to group the weights of certain animals, we could use the first of the following three classifications if the weights are given to the nearest kilogram, the second if the weights are given to the nearest tenth of a kilogram, and the third if the weights are given to the nearest hundredth of a kilogram:

Weight (Kg)	Weight (Kg)	Weight (Kg)
10 – 14	10.0 – 14.9	10.00 – 14.99
15 – 19	15.0 – 19.9	15.00 – 19 .99
20 – 24	20.0 – 24.9	20.00 – 24.99
25 – 29	25.0 – 29.9	25.00 – 29.99
30 – 34	30.0 – 34.9	30.00 – 34.99

Example: Construct a grouped frequency distribution of the following data on the amount of time (in hours) that 80 college students devoted to leisure activities during a typical school week:

23	24	18	14	20	24	24	26	23	21
16	15	19	20	22	14	13	20	19	27
29	22	38	28	34	32	23	19	21	31
16	28	19	18	12	27	15	21	25	16
30	17	22	29	29	18	25	20	16	11
17	12	15	24	25	21	22	17	18	15
21	20	23	18	17	15	16	26	23	22
11	16	18	20	23	19	17	15	20	10

Using the above formula, $K = 1 + 3.322 \times \log(80) = 7.32 \approx 7$ classes

Maximum value = 38 and Minimum value = 10 ; Range = 38 – 10 =28 and W = 28/7 = 4

We can construct grouped frequency distribution for the above data as:

Time Spent (hours)	Frequency	Cumulative Frequency
10 – 14	8	8
15 – 19	28	36
20 – 24	27	63
25 – 29	12	75
30 – 34	4	79
35 - 39	1	80

Cumulative and Relative Frequencies:

When frequencies of two or more classes are added up, such total frequencies are called **Cumulative Frequencies**. This frequencies help as to find the total number of items whose values are less than or greater than some value. On the other hand, relative frequencies express the frequency of each value or class as a percentage to the total frequency.

Mid-Point of a class interval

Mid-point or class mark (Xc) of an interval is the value of the interval which lies mid-way between the lower true limit (LTL) and the upper true limit (UTL) of a class. It is calculated as:

$$Xc = \frac{Upper\ Class\ Limit + Lower\ Class\ Limit}{2}$$

True limits (or class boundaries) are those limits, which are determined mathematically to make an interval of a continuous variable continuous in both directions, and no gap exists between classes. The true

limits are what the tabulated limits would correspond with if one could measure exactly.

Example: Frequency distribution of weights (in Ounces) of Malignant Tumors Removed from the Abdomen of 57 subjects

Weight	Class boundaries	Xc	Freq.	Cum. Freq.	Relative Freq. %
10 – 19	9.5 – 19.5	14.5	5	5	0.0877
20 – 29	19.5 – 29.5	24.5	19	24	0.3333
30 – 39	29.5 – 39.5	34.5	10	34	0.1754
40 – 49	39.5 – 49.5	44.5	13	47	0.2281
50 – 59	49.5 – 59.5	54.5	4	51	0.0702
60 – 69	59.5 – 69.5	64.5	4	55	0.0702
70 – 79	69.5 – 79.5	74.5	2	57	0.0352
Total			**57**		**1.0000**

Note: The width of a class is found from the true class limit by subtracting the true lower limit from the upper true limit of any particular class.

For example, the width of the above distribution is (let's take the fourth class) w = 49.5 - 39.5 = 10.