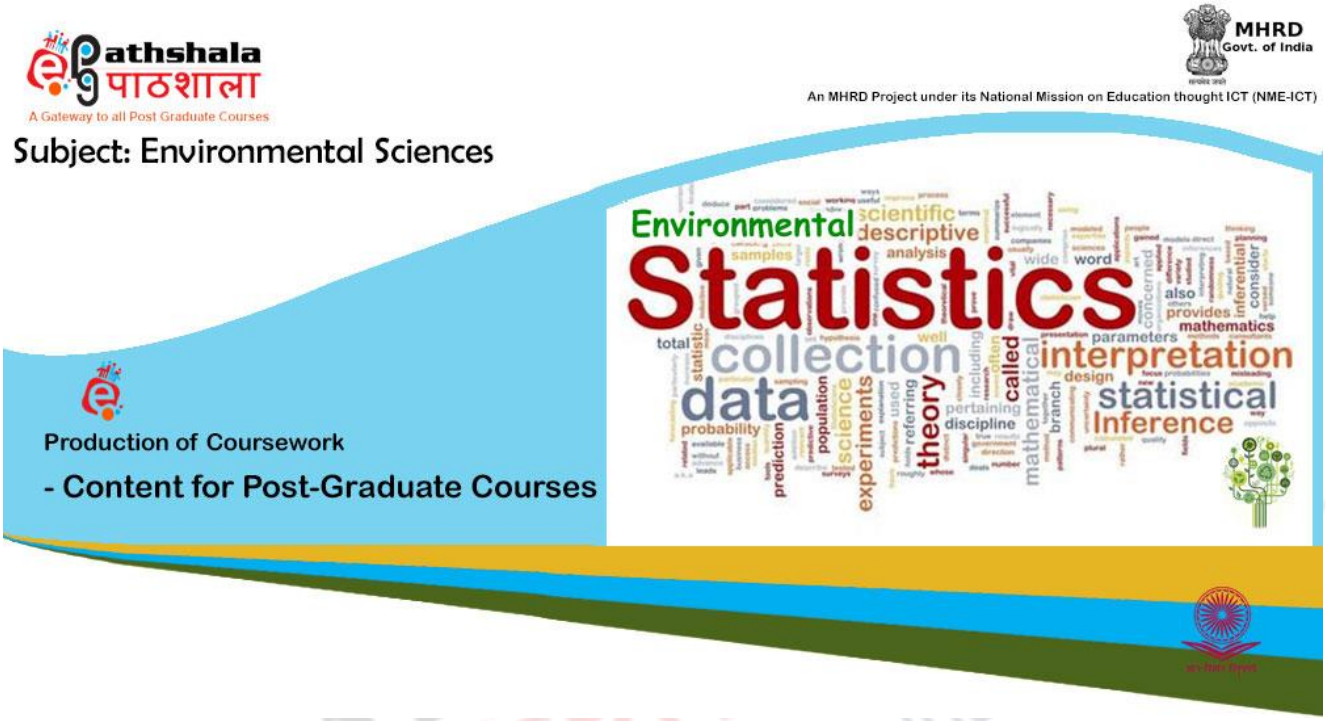

Paper No: 14 Statistical Applications in Environmental Sciences

Module: 4 Diagrammatic and Graphical Representation of Data I

Pathshala
पाठशाला
A Gateway to all Post Graduate Courses
MHRD
Govt. of India

Development Team

Principal Investigator & Co- Principal Investigator	**Prof. R.K. Kohli** **Prof. V.K. Garg &Prof.AshokDhawan** **Central University of Punjab, Bathinda**
Paper Coordinator	**Dr. Harmanpreet Singh Kapoor,** **Central University of Punjab, Bathinda**
Content Writer	**Dr. Harmanpreet Singh Kapoor** **Central University of Punjab, Bathinda**
Content Reviewer	**Prof. Kanchan Jain,** **Panjab University, Chandigarh**
Anchor Institute	**Central University of Punjab**

Description of Module	
Subject Name	**Environmental Sciences**
Paper Name	Statistical Applications in Environmental Sciences
Module Name/Title	**Diagrammatic and Graphical Representation of Data I**
Module Id	EVS/SAES-XIV/4
Pre-requisites	Basic Mathematics
Objectives	Introduction to diagrammatic and graphical representation of data
Keywords	Frequency, cumulative frequency, diagrams, graphs

Module 4: Diagrammatic and Graphical Representation of Data I

- **Learning Objectives**
- **Introduction.**
- **Frequency Distribution**
- **Use of Diagrammatic and Graphical Presentation of the Data**
- **Summary**
- **Suggested Readings**

1. Learning Objectives

In this module, a complete explanation about different types of diagrammatic representation of data will be discussed. This module helps one to learn different methods of diagrammatic presentation and their properties. Through this module, one can learn about which method of representation is appropriate under what type of conditions. Questions with answers are included to give an in-depth knowledge of the topic.

2. Introduction

Statistics is a science that is based on description of data either numerically or diagrammatic or other way. This science is used to extract information from the data based on the objective. Data can be in quantitative form as well in qualitative form. Data are collected from the resources of study and are available in raw form. After collecting and editing the data the next stage is to organize the data. Classification and tabulation of the data are among the most important tools for the precise, clear and comprehensible representation of the data. However, sometimes these forms of presentation are not appealing to the common person. Due to technicality involved in these forms, it may not be interesting for a common person to be able to understand the things in a simple manner. Another way to represent the data is through diagram and graphs that present the data into attractive manner that appeal more to the mind of the spectators. These forms are more attractive, fascinating and impressive than the other methods. The best part of diagrammatic representation method is that even a layman person can understand this without any previous knowledge of statistics. This is the reason that diagrams and graphs are used to give basic education to the kids.

Another important feature of the diagrammatic and graphical representation is that it saves lot of time as these are easy to build up and one can draw meaningful inference from them. These methods are able to present that information that might be lost amid the details of classification and tabulation of data. These methods also facilitate the person for comparing the value of two or more sets of data. Graphs and charts are used to clarify a complex problem and reveal the hidden facts that are not clear from the tabular form. Hence, graphs and diagrams are important not only for the representation of the data but for visual comparison of two or more datasets. Now in the next section, a brief introduction about frequency distribution and how to form it will be discussed.

3. Frequency Distribution

In English language, frequency means the rate of occurrence of something in a repetitive manner in a particular time interval or time frame. Statistics is concerned with the extraction of information from the numbers collected in a raw form from the study. We already discussed about definition of variables in the second module.

In practical life, variables are used to represent these numbers in case of quantitative data like in a study of sales of luxury cars in a particular region in a year. One can consider "the sale of a car in a day" as a variable and note down the sale of a car in a day by x_i variable where i=1,2,…..365. Here the values, $x_1, x_2, \ldots . x_{365}$ represent the sales of a car in the first day of a year or financial year and similarly second day of the year and so on. The values of $x_1, x_2, \ldots , x_{365}$ are known as the values of the variable.

Frequency of a value of the variable means the number of times a same value is repeated in the whole dataset. For example, if we assume that the no. of sales of a car are 2 units on 10^{th} day, 3 units on 20^{th} day, 2 units on the 45^{th} day, 4 units on the 70^{th} day of the year and so on. Here, 2 units appear two times, 3 and 4 units appear single time from the available information. So, from the available information the frequency of sale of 2 cars per day is two and other has only one. With an increase in the available information, one can construct frequency table that represent the repetition of the values in a dataset.

For example, let us consider the observation of sales of car in the month of February (28 values). These are 2, 3, 1, 2, 4, 5, 3, 2, 1, 3, 5, 6, 1, 2, 4, 2, 3, 5, 1, 3, 4, 2, 3, 5, 4, 1, 4, 2

Observations	Frequency
1	5
2	7
3	6
4	5
5	4
6	1
Total	28

Table No. 1

In the above table, we observe that for seven days two units of cars are sold, for six days three cars are sold and so on. Hence one can form the frequency table by counting the same observations in the data set this means *frequency* of a value of a variable is the number of times it occurs in a given series of observations. The table that represents the frequency of the value side by side is called *frequency table*. There are two forms of frequency distributions- Ungrouped frequency distribution and Grouped frequency distribution.

To understand the difference between two methods, we consider a data set of IQ score of 30 students and construct ungrouped and grouped frequency distribution.

IQ Scores of students				
35	65	58	79	41
45	47	87	85	47
51	81	88	76	83
38	62	94	79	84
61	74	86	64	85
54	78	73	71	77

Table No. 2

3.1 Ungrouped Frequency Distribution

Ungrouped frequency distribution table shows the frequency of the values in the dataset on individual basis. Table no. 3 is an example of ungrouped frequency distribution as the first column in the table represents the observations and second column shows corresponding frequency.

Ungrouped Frequency Table

IQ score	Frequency	IQ Score	Frequency
35	1	74	1
38	1	76	1
41	1	77	1
45	1	78	1
47	2	79	2
51	1	81	1
54	1	83	1
58	1	84	1
61	1	85	2
62	1	86	1
64	1	87	1

65	1	88	1
71	1	94	1
73	1		

Table No. 3

From Table No. 3, one can observe that 2nd column and 4th column show the frequency values of the dataset (from Table No. 2). Hence in ungrouped frequency distribution, the values of the dataset are shown on individual basis.

3.2 Group Frequency Distribution

In this method, the values of the variables are shown in the group or interval. In the following table, observations from Table no. 2 are used to present them in group frequency distribution. The smallest observation in the dataset is 35 and maximum value is 94. In Table no. 4, we consider width of the interval as 10 and lowest class value as 30 and so on till we cover the maximum value in the dataset. In the next section, we will discuss about how to choose, type of class, class interval, width of class interval etc.

From Table No. 4, we can see 70-80, 80-90 intervals have maximum frequency value that is 8. Hence we can conclude that IQ score of the most of the students lie between70 to 90.

On the comparison of Table No. 3 and Table No. 4, one can see that group frequency distribution visualize the important characteristics of the data in a simple and understandable manner about the tendency of IQ score of students over ungrouped frequency distribution.

Grouped Frequency Distribution

IQ Score	Frequency
30-40	2
40-50	4
50-60	3
60-70	4
70-80	8
80-90	8
90-100	1

Table No. 4

Steps for forming a group frequency distribution

Many techniques are used for the formation of the group frequency distribution. For group of observations, we divide the data into class intervals and difference between upper and lower interval is called the width of the class interval. There are few points that must be kept in mind while preparing a group frequency table.

(a) **Class type**: One should define class type in a very clear manner. It should be exhaustive and mutualy exclusive so that variable's value must be assigned to only one class in the table.

(b) **Class Intervals:** It means how many intervals should be formed for the available data. Number of intervals depend on the things like number of observations in the data, its magnitude value, level of precision and further analysis of the data.

The most common formula that is used for the determination of the interval in the group frequency distribution is Sturge's rule:

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

where k is the number of classes and N is the total number of observations in the data. This rule is used for correct determination of intervals in the data and it is further used for the determination of the width of the class interval.

(c) **Width of class interval:** Width of the interval means the difference between the lower limit and upper limit of the interval. The width of the interval is defined through the formula that is $h = \frac{Range}{Number\ of\ classes}$

where h denote the width of the class interval and range is defined as the difference between the highest and lowest value of the data set.

(d) Class limits methods: There are different methods that are used for the classification of the data set on the basis of class interval. The limit consists of two numbers that are used for the purpose of tallying observations into various classes. There are two different methods for the classification of the data on the basis of class intervals. These are:- (a) Inclusive method and (b) Exclusive Method.

(i) **Inclusive method:** In inclusive method, the upper limit of a class interval is considered in the interval itself and is not related with the next class. For example, in inclusive method, the class limits are 0-4, 5-9, 10-14, 15-19 and so on. Hence one can see that both the upper and lower limits are included in the class and thus it is termed as inclusive. The main drawback of this method for continuous data observations for example if data value is 4.5 then with this method one cannot tabulate or assign the value to any interval.

(ii) **Exclusive method:** In this method, the data are classified into class interval of such time that upper limit of one interval is the lower limit of next succeeding class interval. For example, in exclusive method the class limits should be of such type that is 0-5, 5-10,10-15 and so on. Hence all those values that are less than 5 are considered in first interval and

all those data values that are above than 5 but less than 10 are counted in second interval. Hence in exclusive method, the problem of inclusive method is taken care of.

(e) Mid value or Mid points: Mid value is calculated by taking the sum of upper and lower limit of the interval and dividing that sum by 2. This value is used as a representative value of the class interval and it is used for evaluation of mean, median, mode and higher moments of the data.

In the next section, we will discuss about cumulative frequency distribution and how to construct it.

3.3 Cumulative Frequency Distribution

We have already discussed about the frequency distribution in the previous section. Frequency distribution counts the occurrence of the same value in the data but sometime one is interested in the number of observations that are small or greater than a given value. In such type of situation, one has to calculate the accumulated frequency less than or greater than some specified value. This accumulated value is known as Cumulative frequency distribution.

The frequency of observations till a given value is considered as less than cumulative frequency and the frequency of observations that are greater than a value is called more than cumulative frequency.

Using the same observation as given in Table No.4. The cumulative frequency distribution for both more and less than are given in the following table.

IQ Score	Frequency	Less than Cumulative Frequency	More than Cumulative Frequency
30-40	2	2	30
40-50	4	6	28
50-60	3	9	24
60-70	4	13	21
70-80	8	21	17
80-90	8	29	9
90-100	1	30	1

Table No. 5

From the above table, one can see the less than and more than cumulative frequency values of the data. These values are further used for graphical representation of the data. For example, ogive curve of more than and less than type use these values for plotting on the axis.

4. Use of Diagrammatic and Graphical Presentation of the Data

Diagrammatic and Graphical presentation of the data are useful in practice due to the following reasons. These are:

(1) The information that we acquire from the graphical and diagrammatic representation of the data is easy to understand even for a layman person due to its simplicity.
(2) People are more interested in graphical presentation of facts then just numbers due to eye caching effect of diagram or pictures.
(3) Graphs and picture can simplify the complexity of the data that cannot be easily be understood with the figures.
(4) With the graphical presentation, one can easily compare the statistical data relating to different time and places to bring out the hidden facts and relationship among the statistical variables.

There are some limitations of diagrammatic and graphical representations like they do not show the details behind the numbers that can only be shown from the table in a better way. A single diagram or graph does not have a great importance rather than it is used for comparison purpose with other diagram or graph. In the next section, difference between graph and diagram will be discussed so that reader can understand the difference between them in a clear manner and use them at their proper place without any confusion.

4.1 Difference between diagrams and graphs

There are few rules based on them, one can differentiate between graphs and diagram but these rules are not standard for all so there is scope of changes in these rules among different persons. But we will discuss few rules that are considered common for all. These rules are:

(a) Diagram are plotted on the paper while graphs are plotted on a paper called graph paper graphs helps in the study of mathematical or numerical relationship between variables but in diagram precise relationship among variables are not discussed.
(b) In diagram, different tools like bars, rectangles, circles etc are used to present the information in the data. Whereas in graphs, different tools like lines, dots etc are used to present the data.
(c) Diagrams only give approximate information regarding the data as this information will not be used further for analysis purpose. On the other hand, graphs give more precise, accurate information about the data and they are used for further analysis purpose.
(d) Diagrams are used for the presentation of the categorical and geographical information in the data. On the other hand, graphs are used for the presentation of the time series and frequency distribution.
(e) Diagrams are more eye catching than graphs. Also diagrams are used for the understanding of the layman person but graphs are used by experts from the field for the further analysis of the information.

(f) Graphs are easier to build than comparative to the diagram.

There are few points based on which one keep in mind while constructing the diagrams. These are:-

- Diagram gives only a pictorial representation of the quantitative data for rough guesses;
- it can only be used for homogenous data;
- it is not reliable to make further inference about the data.

So, basically diagram are used for the graphical interpretation purpose only. One cannot use it to find out reasons or inference from the data.

While constructing diagrams, there are some general rules that should be followed. These are:-

- An appropriate diagram can only present the data in a better way. Thus, it is essential to choose the right diagram for the data that need expertise as well as knowledge. It may be possible that due to inappropriate selection of diagram the interpretation might be wrong that can lead to unbearable results.
- It is also important that a diagram should have an appropriate title corresponding to the nature of the data. With an appropriate title, a person can understand the main idea in the diagram.
- It should be constructed in such a way that it portray all the relevant information within an allotted space. So, it should be appropriate in terms of size and consistent in terms of dimensions.
- It should be neat, clean with footnotes and proper indexing that will attract the interest of the common man.

In the previous section, we discuss the characteristics of the appropriate diagram. Now in the following section, a brief note on different types of the diagrams will help to understand it importance.

4.2 Types of Diagrams

There are many types of diagram based on it dimensionality. These are

(i) One dimensional diagrams
(ii) Two dimensional diagrams
(iii) Pie Chart
(iv) Three dimensional diagrams

Each type of diagram is used for specific type of data i.e. for complex data, one need more dimensions to see the impact of one factor on the other. Hence the choice of the type of the data depends on the nature of the data. One dimensional diagram is discussed here to give elementary knowledge. One can read other dimensions diagram also from the references.

One dimensional diagrams

These types of diagram use only one dimension i.e. only length of bars and lines are taken into account. So, these diagrams are known as one-dimensional diagrams. Bars may be vertical or horizontal. Vertical bars are mostly used to represent growth or decline rate of the variable under study while horizontal bars are used to represent the data of attributes. There are few points that should be kept in mind while using bars.

(a) Bars should be constructed within an allotted space and of uniform shape and size.
(b) Scale should be chosen according to the magnitude of the observations.
(c) Bars must have the same base line for a given data.
(d) It is better to represent the value at the top of the bar for the convenience of the reader.
(e) Bars should be arranged from the left to right in order of magnitude for consistency.

The following example help you to understand the point given above.

The following data give the approximate average yield of rice in kg. per acre in different state of a country during a 2000-2001.

State	Punjab	H.P.	Gujarat	W.B.
Yield in Kg. per acre	640	520	320	1090

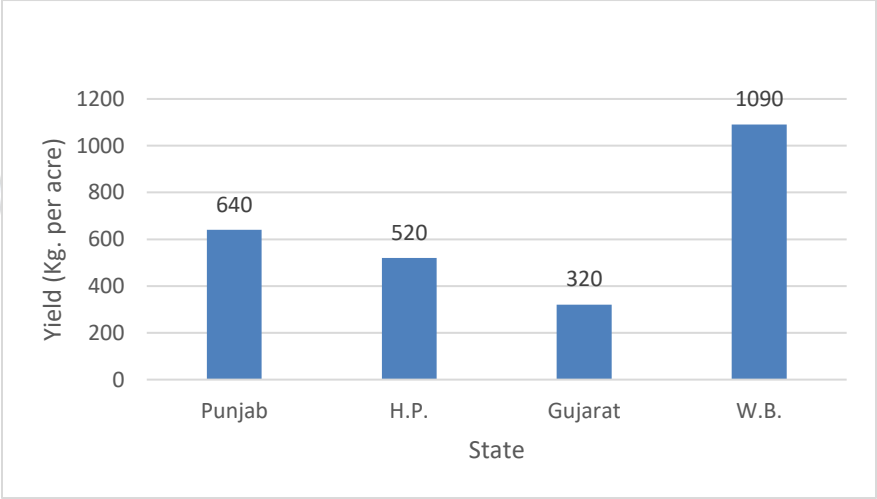

Figure 1

From the above figure, one can observe that the above diagram satisfies all the conditions. Hence one can easily understand the characteristics or facts of the data through diagrams in an easy manner.

Self- Checked Exercise

Question What are the benefits of using diagrammatic and graphical representation of the data?

Ans With the help of diagrammatic and graphical representation of the data even a layman person can understand the facts related with it. Although these techniques are not helpful to explain hidden factors influencing the variables or data.

Question How diagrammatic and graphical representation help in understanding the information contained in the data?

Ans As we are aware of the fact that one can understand the diagram or graphs in a better way than just numbers. Hence, one can easily understand the information contained in the data by using graphical and diagrammatical representations.

Question How graphical representation is different from diagrammatic presentation?

Ans In diagrammatic mode of presentation, one can use the devices like bars, rectangle etc whereas in graphical methods, one can use points, lines of different kind etc to present the information.

5. Summary

Data are collected from the resources of study and they are available in raw form. Thus after collecting and editing the data the next stage is to organize the data. Classification and tabulation of the data are among the most important tools for the precise, clear and comprehensible representation of the data. Frequency distribution and its various forms are discussed in the module. Graphical and diagrammatically forms and differences between them are discussed. Various forms of diagram one-diagram, two dimensional etc. are also used to represent the data in an understandable manner.

6. Suggested Readings

Agresti, A. and B. Finlay, Statistical Methods for the Social Science, 3rd Edition, Prentice Hall, 1997.

Daniel, W. W. and C. L. Cross, C. L., Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition, John Wiley & Sons, 2013.

Hogg, R. V., J. Mckean and A. Craig, Introduction to Mathematical Statistics, Macmillan Pub. Co. Inc., 1978.

Meyer, P. L., Introductory Probability and Statistical Applications, Oxford & IBH Pub, 1975.

Triola, M. F., Elementary Statistics, 13th Edition, Pearson, 2017.

Weiss, N. A., Introductory Statistics, 10th Edition, Pearson, 2017.