Useful for Researchers, Students of Geography & Statistics

STATISTICAL ANALYSIS OF GEOGRAPHICAL DATA

Revised and Expanded Second Edition

Dr. Shrikant Karlekar

Dr. Mohan Kale

Diamond Publications

STATISTICAL ANALYSIS OF GEOGRAPHICAL DATA

Dr. Shrikant Karlekar
Dr. Mohan Kale

First Edition : September, 2006
Revised Second Edition : December, 2013
Reprint : 2022

ISBN : 978-81-8483-550-2

© Diamond Publications, Pune

Cover Page Design
Sham Bhalekar

Printed by
Shri Chintamani Printers, Pune

Published by
Diamond Publications
264/3 Shaniwar Peth, 302 Anugrah Apartment
Near Omkareshwar Temple, Pune - 411 030
☎ 020-24452387, 24466642

info@dpbooks.in
www.dpbooks.in

All rights reserved. No part of this book may be reproduced or utilised in any form and by any means, electronic or mechanical, including photocopying, recording or by any information storage or retrieval system, without permission in writing from the publisher.

Preface

This book is written for the graduate and post-graduate students of Geography. However, it may be viewed as a good text for any course on Statistical methods. It is mainly for non-statistitians, who are unfamiliar with statistical concepts. The aim therefore is to describe statistical concepts and methods in simple language. Attempt has been made to keep mathematical jargon as minimum as possible.

Each Statistical technique and method is demonstrated by an example from Geographical situation. Much of the material in the text is like an introductory course in bivariate and multivariate statistics. We hope that the book will be useful to the students of Geography in particular, having statistical methods as one of the papers at their graduate and post graduate studies.

The book will also be useful to the M.Phil. and Ph.D. research students in the 'Course Work' component as well as in the application of various statistical techniques in their research problems.

We are grateful to Mr. Dattatrya Pashte of Diamond Publications for timely publication of second edition of this book.

Dr. Shrikant Karlekar
Dr. Mohan Kale

About the Authors



Dr. Shrikant Karlekar (M.Sc., Ph.D)

- Presently working as a Honourary Prof. and Head, Dept. of Earth Sciences, Tilak Maharashtra Vidyapeeth, Pune.
- Former Prof. and Head, Dept. of Geography at Sir Parushurambhau College, Pune.
- Awarded Ph.D. in Geomorphology by University of Pune in 1981 on ìA Geomorphic study of South Konkanî.
- Worked as a professor in Geography, at University of Pune in 2004-2005.
- 17 Students received their Ph.D. degree and 7 M.Phil, under his supervision.
- About 60 research papers published in various National and International Research Journals.
- 25 academic books (text and reference) published till year 2013.
- Worked as a principal investigator in many U.G.C. and ICSSR funded projects.
- Received training in ëPhotogrammetryíat Indian Remote Sensing Institute, Dehradun, in 1978.



Dr. Mohan Kale (M.Sc., Ph.D)

- Presently working as a associate Prof. at Dept. of Statistics, University of Pune.
- Awarded Research Fellowship of CSIR, New Delhi for the period 1991-1994.
- Awarded Research Associationship of CSIR for the period 1994-1996.
- Taught Statistics, to PG classes at North Maharashtra University and Shivaji University for one year each.
- Rank holder in CSIR-UGC Joint Examination for Junior Research Fellowship and Lecturership.
- Winner of Young Scientist Award of Indian Science Congress Association in 1993.
- Published Nine Research papers in international and national level Journals of Statistics.
- Published Three articles explaining need of Statistical Techniques in Ayurvedic Research.
- Offered consultancy to around dozen M.D (Ayurved) students for Statistical Analysis of their research.
- Taught at various Departments and Schools of University of Pune as Visiting Faculty (These include Department of statistics and Centre for Advance Study in Statistics, Bioinformatics Centre, Inter Disiciplinary School of Scientific Computing etc.)
- Given around Twenty (20) Invited / Guest Lectures at various Institutes and Presented Papers in about 15 conferences.
- Worked as a Lecturer in Statistics, in Department of Computer Science, S. P. College Pune from 1996.
- Area of Interest - Inference in Branching process, Statistical Applications in Ayurved, Statistical Mehtods in Computer Science and Information Technology.

Table of Contents

	Preface	iii
	About the Authors	iv
1.	Geography & Statistical Methods	1
2.	Frequency Distribution & Data Representation	8
3.	Measures of Central Tendencies	15
4.	Measures of Dispersion	22
5.	Probability & Probability Distributions	31
6.	Bivariate Data and Correlation	41
7.	Regression Analysis	49
8.	Vectors and Matrices	59
9.	Bivariate non-linear (Curvilinear) Relationships	68
10.	Time Series Analysis	75
11.	Sampling Techniques	84
12.	Hypothesis testing	91
13.	The Analysis of Multivariate Data	125
14.	Trend Surfaces and Trend Surface Analysis	136
15.	Other Multivariate Techniques	144
	Statistical Tables	166
	Self Study Questions	205
	References	207

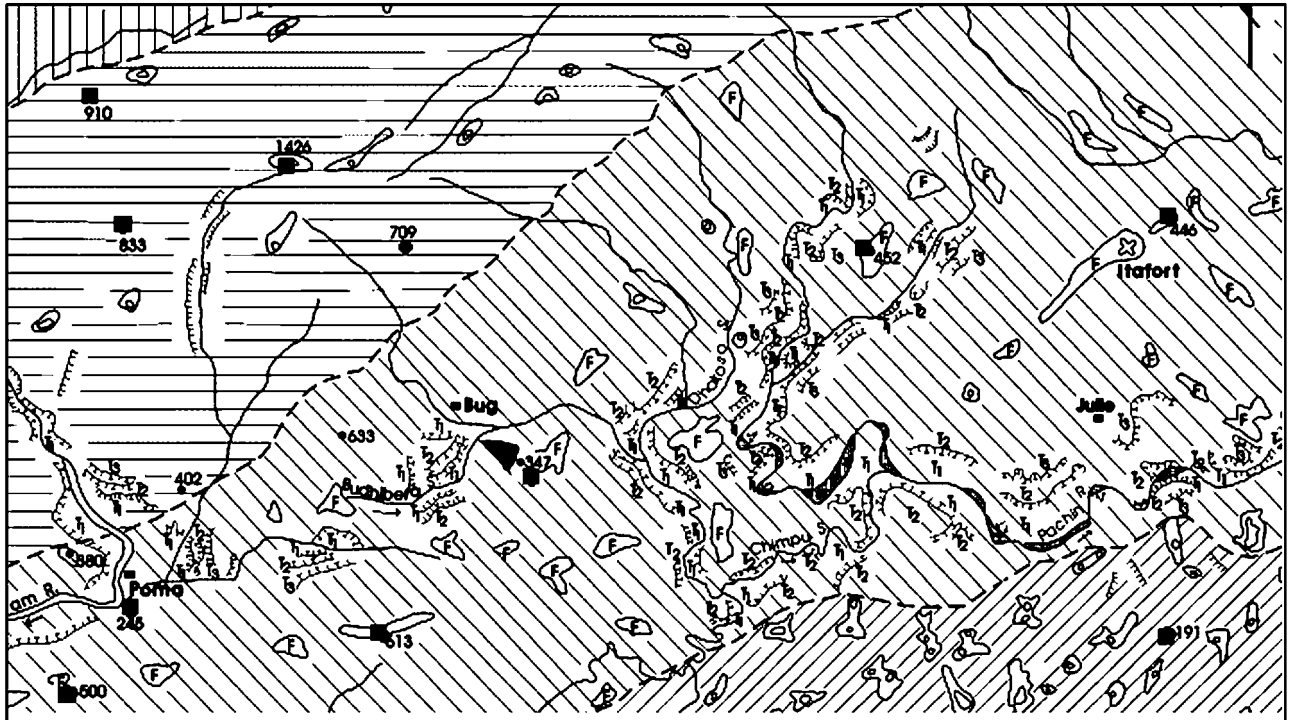
CHAPTER 1

GEOGRAPHY AND STATISTICAL METHODS

The major objective of all sciences, the physical, the biological and the social is to **explain** events. A mere statement of the facts is not useful if no explanation is involved. In geographical situations it is always necessary to provide rational explanations of spatial attributes and interrelationships of phenomena. This is achieved by statistical analysis of geographical data using different types of statistical methods and techniques.

Two basic types of statistics can be identified namely **descriptive** statistics and **inferential** statistics. The descriptive type is designed to describe and explain the geographical situation by obtaining quantitative information (e.g. an evapotranspiration index, location quotient, price index, vulnerability index etc.). The inferential statistics however helps to make probabilistic statements regarding hypothesis, about relationships between variables, about population from which the samples are drawn. Obtaining the point estimates and testing the hypotheses are examples of this type of statistics.

Following example will illustrate the difference between these two approaches in the analysis of a geographical event. The distribution of landslide affected hill sectors in Himalayan zone is shown in figure 1.



Map Showwing distribution of landslide sites. **Fig. 1**

The data regarding altitude and the average incidence of landslides per year in the area is shown in table 1 below.

Location	Altitude (m)	Landslide Frequency (per year)
1	446	5
2	245	4
3	500	5
4	500	6
5	833	7
6	910	9
7	191	2
8	347	3
9	613	6
10	1426	8

From this one can compute several descriptive parameters such as mean altitude (X), Median, standard deviation etc.

Descriptive parameters	Inferential Parameter
Altimate Landslides 601 X = 5.5 500 Median = 5.5 500 Mode = 5.0 370 Std. dev. = 2.2	correlation coefficient $r_{xy} = 0.85$ $y = 2.49 + .005x$

The information obtained from these parameters is no doubt useful but does not convey the relationship and association between altitude and frequency occurrence of landslides. This can be achieved by finding correlation coefficient, analysing reqression and testing hypothesis.

This is called an inferential statistic. Both these types are useful and required in the analysis of geographical events and situations.

The application of quantitative methods is an integral part of any analytical research work. In geography cause and effect relationships can be tested effectively by collecting information in numerical form and analyzing it statistically (e.g. High density of population in good quality soil area).

An event or observation can be described by different ways(e.g. Hill slopes in Western Ghat region are very steep/less steep than those in Himalayan region or the forests in Western Ghats are degraded or not very dense etc.). Such descriptions are necessarily qualitative and subjective. The description like this can be improved only by quantification of the event. As pointed out earlier, the relationships and associations among the data then can be analysed and tested(e.g. rate of soil erosion depends on amount of slope, rainfall, soil type and vegetation cover etc.).

The basis of all statistical analysis in geography is the observation value, from which by induction method a scientific truth can be developed (e.g. The height of a place above mean sea level is 50m and the atmospheric pressure recorded is 980 mb. At a place 100m ASL it is 960mb and at 120 m ASL it is 920 mb. Thus the atmospheric pressure decreases as height of a place above sea level increases).

A quantitative analysis in geography involves 4 major stages.

1) Data collection
2) Graphical representation of the data
3) Data analysis
4) Data interpretation

It is always necessary to know the nature of geographical data before any descriptive or inferential statistical analysis is undertaken.

NATURE OF GEOGRAPHICAL DATA

The geographical data used in most of the analytical exercises in geography has following salient features:

a) Data are obtained and collected from various geographical areas, regions, subregions and zones such as states. districts and villages. Due to this, there is a large spatial variability in the data

b) Numerical data collected have a reference of geolocation in the form of latitudes and longitudes.

c) Secondary data are primarily collected from reports, census handbooks, socioeconomic surveys etc. Primary data are obtained in the field by direct measurements.

d) Data are obtained in a discrete format (i.e. In the counted form such as number of houses, trees, people, gravels, slopes etc.) as well as in a continuous format (i.e. in the measurement form e.g. Distance between houses, height of trees, size of gravels, amount of hill slopes etc.) In a discrete data observation values are integers and no value will

occur between1 & 2 or 2 & 3 etc.The values in discrete data are obtained by counting. (Table 2)

Table 2

Number of pebbles in various potholes in a river channel

70	72	60	80	52
71	51	81	72	58
90	76	91	63	51
51	90	71	91	67
70	71	63	90	67
90	72	65	55	52
67	62	76	69	64
79	89	79	66	88
81	90	80	79	70
58	71	76	71	80
76	88	54	75	91

The data in the continuous format can be obtained directly (measurement of height, distance, area) or indirectly (measurement of temperature from the mercury column of thermometer)

Table 3

Size of pebbles (in cm) in a pot hole

4.6	4.9	12.3	7.8	8.7
4.8	14.5	29.2	13.3	18.6
5.5	15	5.5	9.3	12
27.2	6.6	8.4	17.9	13
21.2	16.8	20	24.3	23.4
8.3	19.7	16.2	17.4	13.7
15.1	22.2	14.8	17.5	9.8
10.2	13.4	14.9	14.3	11
16.3	21.6	28.3	11.4	12
12.3	1.1	3.2	5.4	13
14.3	11.2	6.9	11	30

e) In geography data are also extracted from maps like Survey of India toposheets.

f) The oral data collected through a questionnaire is also used in Geography

g) Data in geography can be assorted, compiled or piecemeal in nature

h) Data are written and obtained in 'space-time' language(e.g. Population of a particular place in a specific year).

i) The numerical value in the data can be zero dimensional (point), unidimensional (line), two dimensional(area), three dimensional (volume) or multidimensional(space-time)

j) Data in geography necessarily have a hierarchy and a scale.

k) If there is no natural hierarchy in the observations it has to be created.

The attributes in the data format have a specific unit(e.g. meters, milibars, hectapascals etc). Comparision is possible only in similar attributes (e.g. Agricultural density of village A can be compared with that of village B and village C etc).

l) Different variables in a data set in geography can not be measured by same method and using same level of measurement. Hence level of measurement must be decided before the data collection is undertaken.

LEVELS OF MEASUREMENT:

Method of quantitative analysis is governed partly by the level at which data are measured. Four ascending levels of measurement are frequently used. They are,

i) Nominal or classificatory scale : This is used when observations can be only placed in mutually exclusive categories(e.g. Streams classified by their order, people classified by their occupation etc.) This is the weakest level of measurement. It is necessary that members of the subclass are equivalent.

ii) Ordinal or ranking scale : This is a stronger level of measurement because the subclasses can be ranked above or below one another(e.g. Largest town has rank one followed by rank two of second largest town etc.) Sometimes two observations have same magnitude. In such cases tide rankings are given. Except in cases of

tide rankings the difference between sequentially ordered observations is always one.

iii) Interval scale : This is a still stronger level of measurement. Here not only observations are ordered but the magnitude of the difference separating any two observations along the measurement scale is variable (e.g. Cities ranked, give not only the ranks but also the number of people in each group)

iv) The ratio scale : It is the highest level of measurement. Here measurements are made relative to some absolute and non arbitrary base(e.g. The ratio of distance between factory 1 and market and factory 2 and market is 5/10 kms. It will be identical if distances are measured in miles i.e. 8/16).

■■

CHAPTER 2

FREQUENCY DISTRIBUTION AND DATA REPRESENTATION

A careful study of the observations shown in table 1 and 2 shows that some values occur frequently in the data set. This is natural or obvious tendency seen in the observations in Geography. (Sometimes a particular event occurs frequently in the area. e.g. flooding of Brahmaputra river near Gauhati or cyclones on the east coast of India).

Frequency is an incidence and therefore it is counted and represented in a discrete format. (Sometimes a relative frequency of an event is calculated which is expressed in fraction). The frequency distribution is the classification of different observations in different groups or classes. The frequency occurrence of the observations in each class is recorded. The data is then grouped in various classes of suitable sizes. Normally the total number of classes should be between 5 and 10. The tabulation of the data in different classes is carried out by using tally marks. The table thus obtained is called frequency table.

The data in table 2 is grouped in frequency table as follows

Table 4 : Frequency Distribution

Frequency distribution of pot holes shown in table 2
Table 4

Class	Talley Marks	Frequency **(f)**	Cumulative frequency **(cmf)** ↓	Cumulative frequency **(cmf)** ↑
51 - 60	𝍸 IIII	9	9	55
61 - 70	𝍸 𝍸 III	13	22	46
71 - 80	𝍸 𝍸 𝍸 𝍸	20	42	33
81 - 90	𝍸 𝍸	10	52	13
91 - 100	III	3	55	3

Thus the distribution of pebbles in various classes is obtained by assigning each observation a pre-decided class. The data in table 2 shows that the minimum number of pebbles found in the pot holes of the river channel is 51 and maximum is 91. Thus the total range in the observations is 40 (91-51). While grouping the data in various classes, the number of classes is fixed usually on the basis of this range. In this example the data is grouped in 5 classes as shown in table 4.

The lower and upper limit of the class is decided considering whether the data is discrete or continuous. While preparing classes the class limits/ boundaries are selected in such a way that the smallest observation is to be classified in first class and maximum observation is to be classified in the last class. If the upper limit of the current class is same as lower limit of the next class for all the classes then such a frequency distribution is called as frequency distribution with continuous classes. In such case, class limits are called as class boundaries.

In case of discrete data upper limit of the class is not equal to the lower limit of the next class. However in the case of continuous data, lower limit of the class is same as the upper limit of the previous class. Moreover in case of continuous data, it is also necessary to fix the decimal places of lower and upper limits. These places are usually one decimal higher than the decimals in actual data. For the data shown in table 3, the frequency distribution will be as shown in table 5 below.

Frequency distribution of size of pebbles shown in table 3
Table 5

Class	Talley Marks	Frequency **(f)**	Cumulative frequency **(cmf)**	**(cmf)**
0.01 - 5	𝍸	5	5	55
5.1 - 10	𝍸 𝍸 |	11	16	50
10.1 - 15	𝍸 𝍸 𝍸 𝍸	20	36	39
15.1 - 20	𝍸 𝍸	10	46	19
20.1 - 25	𝍸	5	51	9
25.1 - 30	||||	4	55	4

The frequency distribution shown in table 5 can be shown in graphical form. The representation of a frequency distribution in graphical format is called **frequency histograms.** It enhances the distributional properties of the data. A frequency histogram is constructed by showing frequency on Y-axis and classes on X-axis. A suitable scale is used to show the frequencies on Y-axis. The vertical rectangles drawn to scale have uniform width and all the rectangles are not drawn separately. The frequency histogram for the data in table 4 and 5 is shown in **figure 2 a, b**.

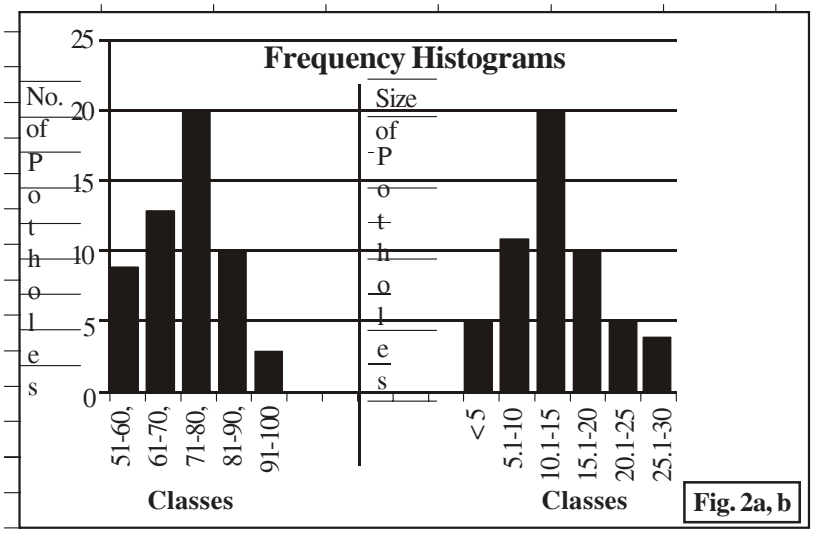


Fig. 2a, b

The height of the rectangle is equal to the frequency of the class.

The frequency distribution can also be shown as **frequency polygon**. In this, the frequency distribution is shown by a line graph. The class midmarks are shown on X-axis.The frequency in the class is shown on Y-axis. The points are joined in sequence by line segments.

While preparing a polygon the two ends of the polygon are joined to X-axis considering zero frequency in the class previous to first class and next to last class. The polygon thus created shows the distribution of frequencies and the total area under the curve is determined by total frequencies. Figure 3 a, b shows frequency polygons for the data in table 4 and 5.

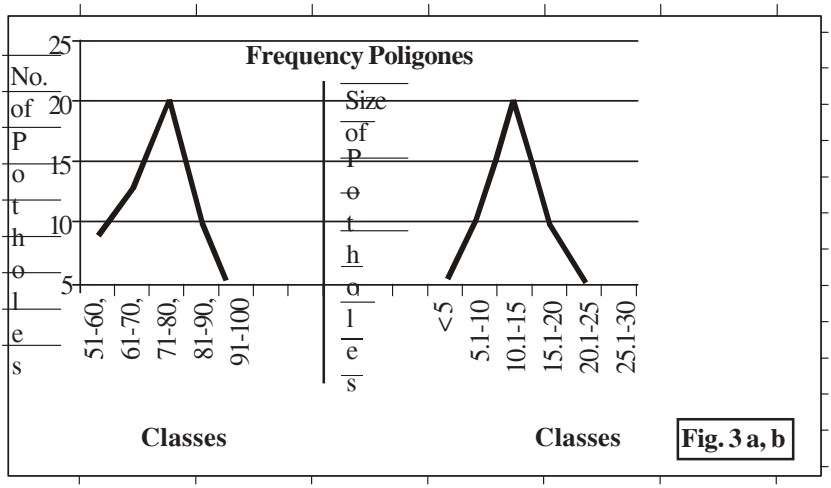


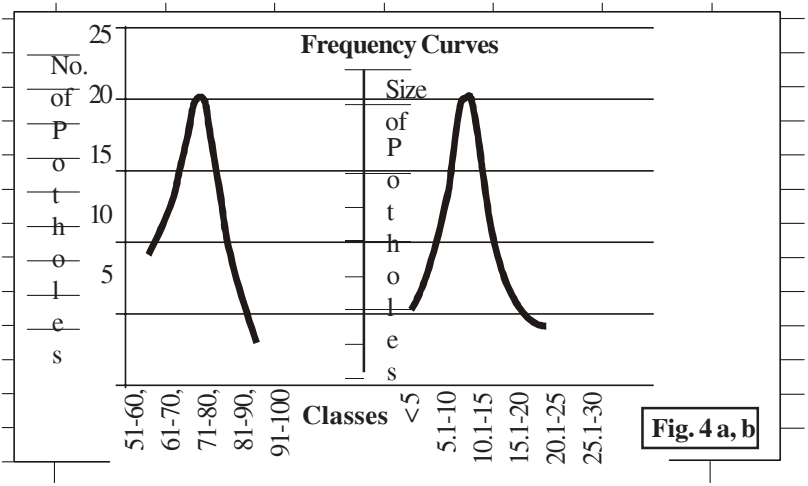
Fig. 3 a, b

When points are joined by smooth free hand curve and the two ends of the polygon do not meet the X-axis, the figure is called as **frequency curve (Fig 4 a, b).**

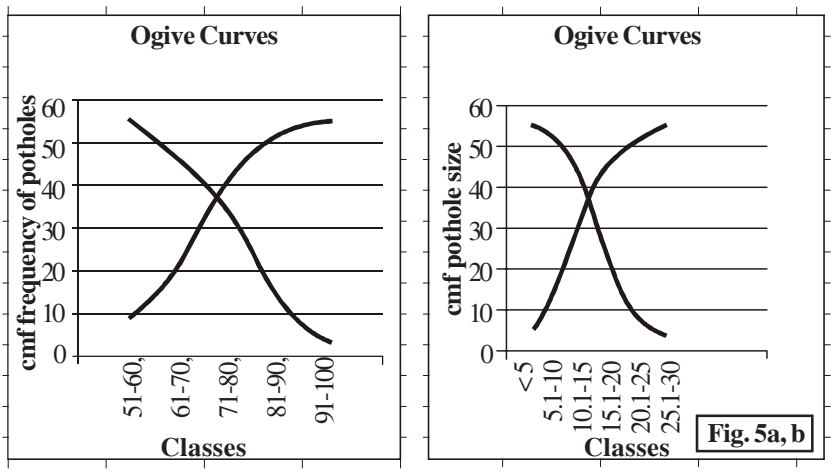
The other type of curve that is used to show the cumulative frequency is called **ogive curve**, as it resembles S shaped ogee curve. Cumulative frequency curve is of two types (i) Cumulative frequency less than type which means it is the number of observations less than equal to the upper class boundary of the class. Cumulative frequency less than type can be obtained by adding frequency downwards i.e. from lowest class to highest class. The distribution so prepared is called as cumulative frequency less than type distribution.

(ii) Cumulative frequency greater than type which means it is the number of observations greater than equal to the lower class boundary of the class. Cumulative frequency greater than type can be obtained by adding frequency upwards i.e. from highest class to lowest class. The distribution so prepared is called as cumulative frequency greater than type distribution.

Thus cumulation of the frequencies can be performed from lowest class to highest class or vice-versa. (Table 4 and 5).



Cumulative frequencies obtained in upward and downward direction when plotted on Y-axis with respective class boundaries give cumulative representation of frequencies in ogive curve. The major advantage of the ogive curve is that one can locate median, quartile and decile values on the curve. Ogive curves for the table 4 and 5 are shown in figure 5.



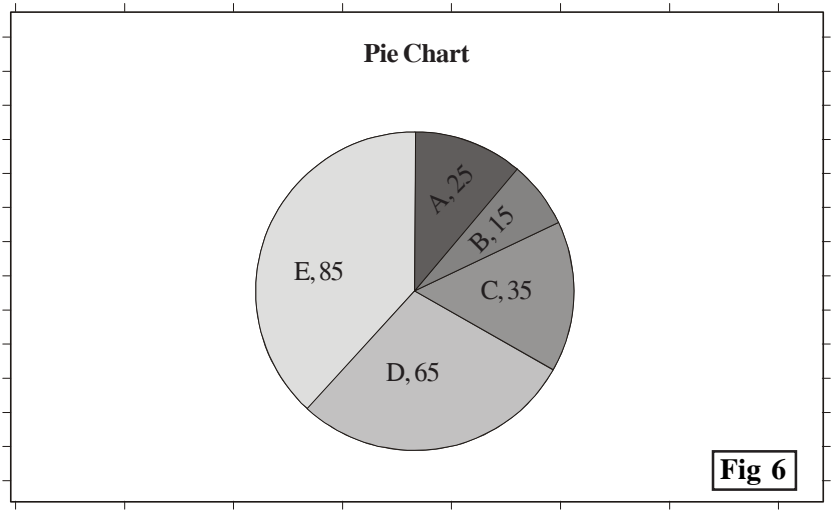
Pie Chart: This graphical representation is used when data are measured on Nominal or Classifactory scale. The brief procedure is as follows.

The angle corresponding to proportion of degrees out of 360 degrees is marked on circle and corresponding area of circular arc is the share of that category.

Table 6

Category	Number of objects	Proportion	Degrees
A	25	0.111111	40
B	15	0.066667	24
C	35	0.155556	56
D	65	0.288889	104
E	85	0.377778	136
	225	**1**	**360**

The data in above example can be shown by a pie chart as below. (Fig. 6)



Figs : Pie Chart

The data in table 4 will be represented in a pie chart as shown below.

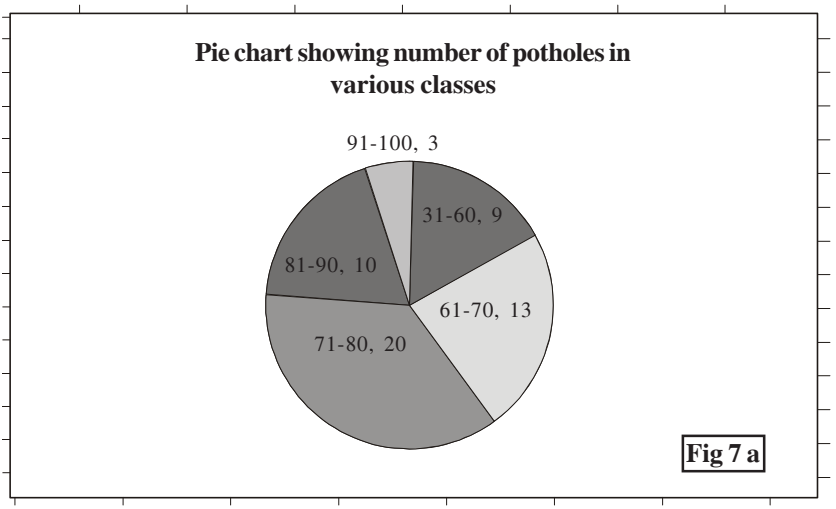


Fig 7 a

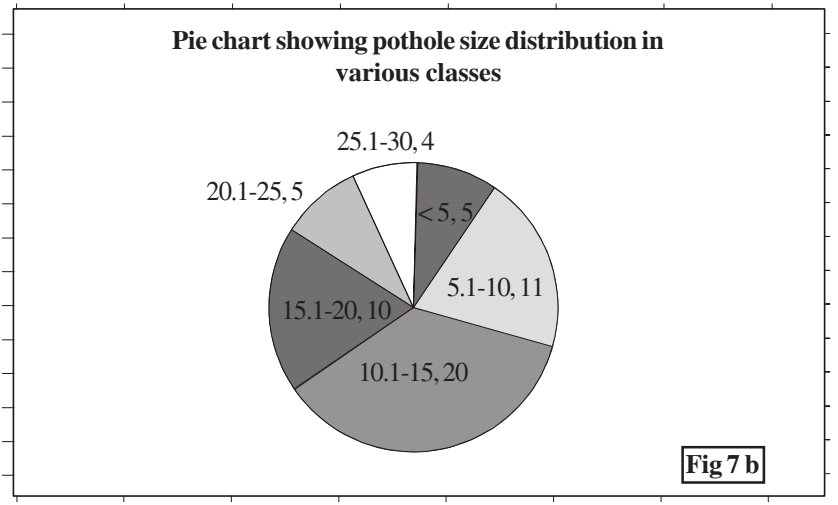


Fig 7 b

■■

CHAPTER 3

MEASURES OF CENTRAL TENDENCIES

There are various parameters which describe a distribution. They include measures of central tendencies, dispersion and other higher order moments. These are descriptive statistics and are relevant in number of non-parametric and parametric tests discussed in later chapters. They also play very crucial role in modeling and pattern recognition.

These parameters are used to describe the data by summarization. Average is one such parameter which is synonymously called the central tendency. Arithmetic mean, mode, median as well as geometric mean and harmonic mean are five ways of measuring the average of the set of the data.

Central tendency is an important and useful property of the variables which show a spatio-temporal change. (e.g. Precipitation, temperature, elevation, production etc.) These variables show a distribution (which is not the case for constants like gravity and π).The distribution may show a wide range in the observations. The average or the central tendency of the distribution is the best representation of the total distribution. The centre of the distribution range does not represent the distribution of data.

Average Or Mean: It is basically the arithmetic mean of the data. For the **ungrouped** data, it is calculated as,

$$\bar{X} = \sum x / n$$

Where $\bar{X}$ is the mean, $\sum x$ is the summation of all observations and n is the total number of observations in the data. The mean of the observations given in table number 2, is therefore,